Atten Percept Psychophys (2014) 76:2158–2170
DOI 10.3758/s13414-014-0649-8

Toward ecologically realistic theories in visual short-term memory research

A. Emin Orhan · Robert A. Jacobs

Published online: 22 March 2014

Abstract Recent evidence from neuroimaging and psychophysics suggests common neural and representational substrates for visual perception and visual short-term memory (VSTM). Visual perception is adapted to a rich set of statistical regularities present in the natural visual environment. Common neural and representational substrates for visual perception and VSTM suggest that VSTM is adapted to these same statistical regularities too. This article discusses how the study of VSTM can be extended to stimuli that are ecologically more realistic than those commonly used in standard VSTM experiments and what the implications of such an extension could be for our current view of VSTM. We advocate for the development of unified models of visual perception and VSTM—probabilistic and hierarchical in nature—incorporating prior knowledge of natural scene statistics.

Keywords Visual short-term memory · Ecological validity · Probabilistic model

Introduction

Research on visual short-term memory (VSTM) generally uses artificial visual displays consisting of simple objects with easily parameterizable features and with no statistical structure within or between objects. These displays are used to address questions such as the nature of capacity limits or the units of storage in VSTM (Brady, Konkle, & Alvarez, 2011). This choice of stimuli confers many advantages to the design of experiments and to the interpretation of results obtained from such experiments. Among the advantages of using artificial and simple stimuli are the ease with which such stimuli can be generated and manipulated, the fact that they make the formulation and testing of hypotheses straightforward, their relative unfamiliarity for subjects, and, lastly, their "bare bones" character, stripped of features that are irrelevant to the hypothesis under question. The last two properties help to minimize the effects of irrelevant prior knowledge or assumptions subjects might bring to a VSTM task.

Despite these advantages, psychologists have also been aware of the potential problems that are associated with the use of artificial stimuli and tasks in experimental studies (Brunswik, 1943, 1955; Neisser, 1976). The main concern here is the danger that these stimuli and tasks might be too artificial to give an accurate reflection of the problems faced by an observer in the natural world. Given that many aspects of perception and cognition can be profitably thought of as rational solutions or adaptations to problems that observers (and actors) encounter in their natural environments (Anderson, 1990; Geisler, Perry, & Ringach, 2009), unnatural stimuli and tasks might lead to misleading characterizations of perceptual and cognitive processes.

In this article, our goal is to take a critical look at the use in VSTM studies of impoverished, unnatural scenes lacking the rich statistical structure displayed by stimuli that are more representative of the natural environment. Although we acknowledge that experiments using artificial displays with simple statistical structure are often useful first steps in elucidating fundamental perceptual and cognitive processes, we argue that whether and how results obtained from such experiments would generalize to stimuli and tasks that are more representative of the natural environment should always be considered carefully. If there are any doubts about the

A. E. Orhan (✉)
Center for Neural Science, New York University,
New York, NY 10003, USA
e-mail: eorhan@cns.nyu.edu

R. A. Jacobs
Department of Brain & Cognitive Sciences,
University of Rochester, Rochester, NY 14627, USA
e-mail: robbie@bcs.rochester.edu

generalizability of the results, experimental stimuli and procedures will need to be refined accordingly. In the following sections, we discuss why we think the question of ecological validity in VSTM research should be taken more seriously and how the study of VSTM can be extended to ecologically more realistic stimuli, as well as possible implications of such an extension.

Why is it important to take the question of ecological validity in VSTM research seriously?

As was mentioned above, although the use of unnatural stimuli and tasks might afford greater experimental control, psychologists have been aware of the problems with generalizing the results of such studies (Brunswik, 1943, 1955; Neisser, 1976). This generalization problem may be more acute in some domains than in others, but it is one that every cognitive psychologist should take seriously if he or she does not want to study a problem that exists in a laboratory setting but that does not exist or is either ill-defined or much less important for real observers or actors living in natural environments. Below, we discuss several specific examples from the VSTM literature to illustrate the need to take the question of ecological validity seriously.

Nature of capacity limits

One of the most popular questions in the VSTM literature addresses the nature of capacity limits in VSTM. For example, from a single fixation, how much information can we encode and then maintain in VSTM across a brief interval, and what is the nature of this capacity limit? The notions of information and capacity are derived from information theory and, to be meaningful, require well-defined stimulus spaces. This, together with reasons mentioned in the Introduction, may explain why studies on the nature of capacity limitations in VSTM often use stimulus displays containing simple items with simple features. One consequence of this choice, however, is that theories or models attempting to explain the nature of VSTM capacity limitations are specifically tailored to these types of stimulus displays. For example, these theories put forward rivaling explanations of how a fixed amount of resources in VSTM gets distributed over some simple "items" in a display. Some theories (so-called high-threshold models) have claimed that only a few items can be selected from the display and encoded with near-perfect resolution, while the remaining items are not encoded at all (Luck & Vogel, 1997; Rouder et al., 2008). Other theories posit either a discrete (Zhang & Luck, 2008) or a continuous (Bays & Husain, 2008) resource that can be distributed among the items either evenly (Wilken & Ma, 2004) or unevenly (Van den Berg, Shin, Chou, George, & Ma, 2012).

However, even leaving aside the difficulties surrounding the issue of what constitutes an "item" in more natural scenes (what constitutes an item in the rightmost image in Fig. 1a? Are the arms or heads of the pedestrians separate "items," or are only whole bodies items? How many "items" are there on the traffic lights?), it seems evident that these accounts can give, at best, an incomplete picture of the nature of capacity limitations in VSTM for the simple reason that we do not just see and remember "items" in the natural world.

Figure 1 shows examples of different kinds of visual information present in natural environments that can be encoded and maintained in VSTM. These examples suggest that we can remember visual information about textures (Fig. 1b), material properties of objects (Fig. 1c), complex real-world scenes that are presumably encoded at multiple levels of abstraction (Fig. 1a), and even actions (Giese & Poggio, 2003) such as walking, running, etc.[1] (Fig. 1d).

Consider, for instance, the leftmost image in Fig. 1b. If we are briefly shown the texture in this figure, we will undoubtedly remember something about it. Subjectively, it seems sensible to suppose that we will not be able to remember every detail of the texture and that our memory will be more impoverished than our perception of the same texture.[2] Nonetheless, we will be able to remember *something* about it. Why is our memory more impoverished than our perception? Why can we not remember every visual detail in this texture, and what exactly can we remember about it? Similarly, for real-world scenes such as the one shown in the rightmost image in Fig. 1a, what exactly can we remember about this image from a brief presentation? Existing theories on the nature of capacity limitations in VSTM have little to say about these issues, because they are specifically tailored to the types of artificial stimuli typically used in VSTM studies.

The question of the distribution of resources among items in simple displays does not arise for stimuli such as those in Fig. 1b, c, because they do not contain any items, and is at best poorly defined for stimuli shown in Fig. 1a, d, because these stimuli are encoded not merely in terms of separate, individual items, but in a more complex, hierarchical way. Even the possible "items" or objects in these figures are not encoded as monoliths, but as structured objects composed of many parts put together in specific configurations. For example, the pedestrians in the rightmost image in Fig. 1a can be decomposed into heads, arms, torsos, and so forth. Their

[1] It may be claimed that the last example does not constitute visual information. However, we can perceive and remember rich visual details about, for example, somebody's manner of walking that cannot be not easily classified as anything other than "visual." It is doubtful whether there is a justifiable sense of "visual" according to which the shape of an object counts as visual information but the manner of movement does not.

[2] See, for example, Huang and Sekuler (2010) for a direct comparison of the precision of visual perception with the precision of VSTM in a simple estimation task.

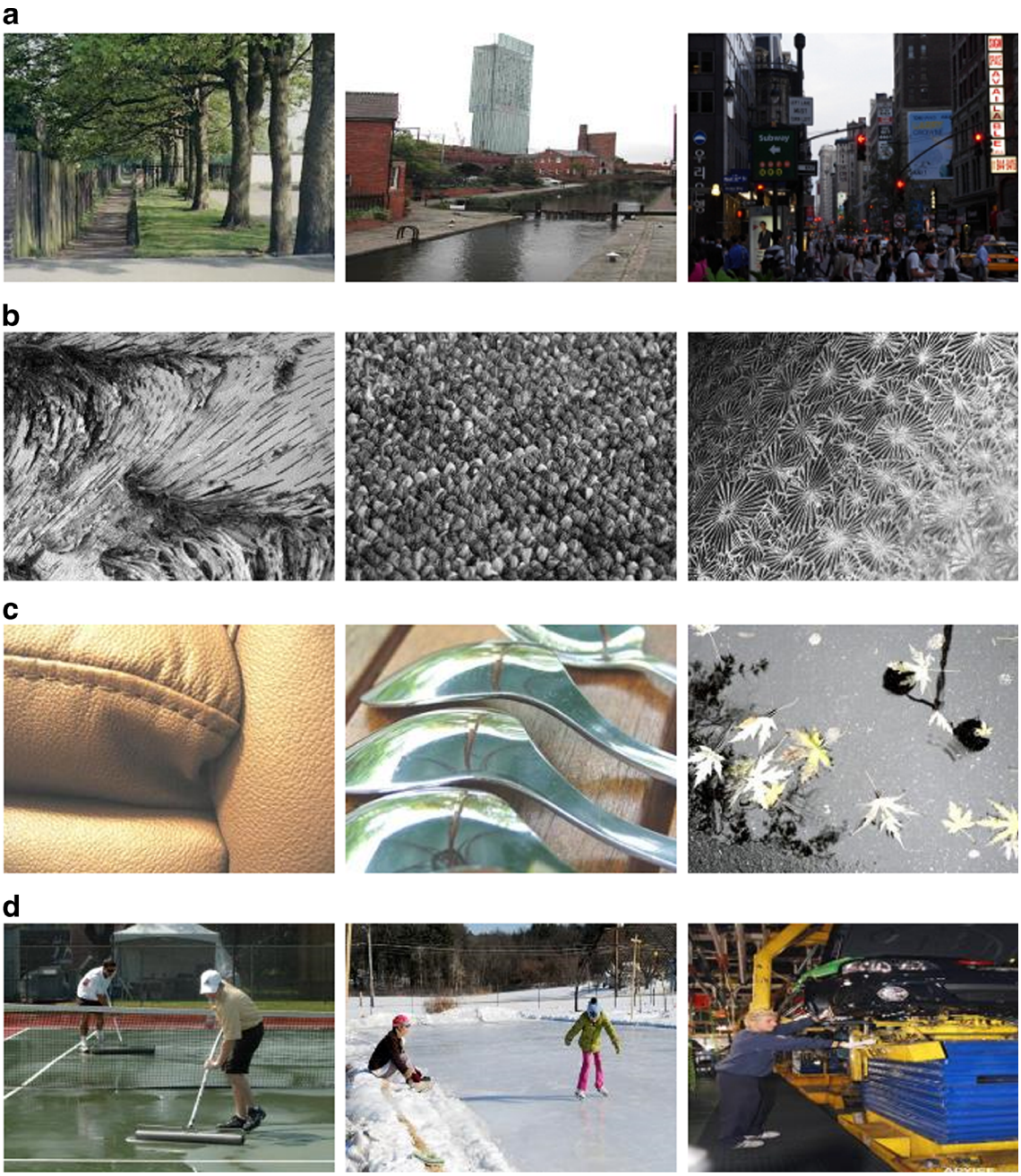

Fig. 1 Example natural images highlighting a variety of different types of visual information that we can perceive and remember in the real world. **a** Complex real-world scenes. The first two images are from the SUN database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010); the rightmost image is from http://www.flickr.com/photos/13774569@N07/5923460837/. **b** Textures. Images are from the UIUC texture database (Lazebnik, Schmid, & Ponce, 2005). **c** Material properties such as glossiness, rigidity, roughness, being made of leather, and so forth. Images are from the Flickr Material Database (Sharan, Rosenholtz, & Adelson, 2009). **d** Actions. Images are from the SUN database

heads can be further decomposed into various facial features, and so on. In the section titled How to Extend Models of VSTM to Natural Stimuli, we will discuss how the encoding of such richly structured scenes in VSTM can be modeled with hierarchical probabilistic models.

These examples demonstrate that one of the most popular questions currently debated in the VSTM literature, the nature of capacity limits, and the answers given to it are specific to the impoverished, artificial stimuli used in most VSTM studies and they are either not quite well-defined or simply do not arise within the context of ecologically more valid stimuli.

Units of storage

Another example is the question of the units of storage in VSTM—that is, the question of whether objects are represented

as bound units in VSTM or as independent features (Bays, Wu, & Husain, 2011; Fougnie & Alvarez, 2011; Luck & Vogel, 1997). Again, this question is not meaningful for stimuli that do not involve objects. A more meaningful question for such stimuli is what exactly people do represent about them in VSTM. For stimuli that involve objects, the answer to the question of whether features of an object are represented independently is likely to depend on factors such as familiarity with the object, the particular object category or features in question, and so forth. For example, encoding of different features might be expected to be much less independent for highly familiar objects than for less familiar objects. Similarly, for object categories with a higher degree of consistency between their features, the encoding of features might be expected to be less independent: Think of the encoding of the color and the shape of a banana versus the encoding of the color and the shape of a pen. To the best of our knowledge, these questions are simply not addressed in the current literature on VSTM because of a lack of interest in extending the research questions formulated within the context of simple, artificial stimuli (in this case, the question of whether a simple object is represented as a bound unit in VSTM) to ecologically more realistic conditions.

Effects of eye movements, eccentricity, viewing conditions

In many VSTM studies, stimuli are presented solely in the periphery, and eye movements are prevented either by monitoring them or by adopting short presentation times (shorter than about 150 ms). In addition, in most VSTM studies, subjects are not given a task other than to remember the stimuli. In natural vision, in contrast, we sample the world by making frequent eye movements under tight control of task demands (Hayhoe, 2000; Land, Mennie, & Rusted, 1999). Importantly, fixated objects and targets of upcoming saccades are encoded better than unfixated objects (Bays & Husain, 2008; Hollingworth & Henderson, 2002). Indeed, it has been argued that visual information from successively fixated objects can be integrated into a detailed visual representation of the scene (Hollingworth, 2004; Hollingworth & Henderson, 2002). In addition, perceptual grouping mechanisms appear to be more effective near fixation than in the visual periphery (Velisavljevic & Elder, 2008). These results suggest that, relative to natural (unconstrained) viewing conditions, the prevention of eye movements in experiments might engage different encoding mechanisms and lead to qualitatively different patterns of results. Therefore, researchers using constrained viewing conditions in VSTM experiments should carefully consider whether their results would generalize to more natural viewing conditions.

Another obvious difference between viewing conditions in natural vision and in typical experimental settings is the full-field, 3-D visual stimulation under natural conditions versus limited-field, 2-D projections used in most experiments. It remains to be seen what the consequences of these differences, if any, might be for VSTM studies.

Other issues

In our daily lives, most of our visual perception involves objects and environments that we are extremely familiar with. How does this familiarity affect the characteristics of visual short-term memories for such objects and environments? Could the results of experiments using unfamiliar stimuli be reliably generalized to extremely familiar stimuli? If not, how should they be modified (see, e.g., Hemmer & Steyvers, 2009)? For example, it has been suggested that object features are encoded approximately independently in VSTM (Bays, Wu, & Husain, 2011; Fougnie & Alvarez, 2011). Does this result generalize to objects that we are extremely familiar with? Similarly, do the set size effects (i.e., decline in memory precision with the number of objects presented) typically observed in VSTM studies generalize to stimuli we are extremely familiar with? If not, how should these results be modified? Context effects like the word superiority effect (Reicher, 1969) demonstrating that, in some cases, it may be easier to recognize objects or features within the context of other objects or features provide a reason for suspecting that the decline in memory precision with the number of objects may not easily generalize to stimuli we are extremely familiar with (see Bar, 2004, for a review of context effects in object recognition).

Another example where questions about the generalizability of experimental results to more natural conditions might arise is the recent finding that encoding precision is variable across items and across trials in typical VSTM tasks (Fougnie, Suchow, & Alvarez, 2012; Van den Berg et al., 2012). It remains unclear whether this result would generalize to the encoding of objects in more realistic scenes. In addition, these studies used peripheral presentation of stimuli, and thus, variability here refers to variability in encoding isoeccentric items in the visual periphery. However, under more natural viewing conditions, it is possible that any variability in the periphery would be small, as compared with the differences in encoding precision at different eccentricities (e.g., encoding precision in the fovea vs. the periphery), thus making it a less significant phenomenon.

Finally, some studies have reported so-called misbinding errors in VSTM recall tasks (Bays, Catalao, & Husain, 2009; Emrich & Ferber, 2012) where subjects, instead of reporting the feature value of the target item, mistakenly report the feature value of a distractor. It is not clear whether these types of errors would happen with significant frequency in more natural conditions. Reason for suspicion arises because the frequency of these errors is modified even by seemingly insignificant changes, such as whether subjects make a response by scrolling or pointing (Van den Berg et al., 2012).

A possible objection

It may be objected that addressing the questions raised in the present section does not necessarily require using natural stimuli in VSTM studies but can be addressed equally well, and perhaps even better, with artificial but controlled stimuli. For example, the question about the effects of familiarity can be addressed by familiarizing subjects with a set of artificial stimuli by extensive training prior to the actual experimental sessions. There are several possible responses to this objection. First, although this approach may be feasible for some of the questions raised here, a potential problem with this approach is that even after extensive training, artificial stimuli might still be less privileged for subjects than natural stimuli. For example, Schwarzkopf and Kourtzi (2008) showed that subjects were much better at detecting contours that contained an ecologically valid cue (collinearity of oriented line segments) than those with an ecologically invalid cue (line segments orthogonal to the contour path) even after extensive training with the ecologically invalid cue. Second, the use of natural stimuli is necessary for addressing some of the questions raised above. For example, addressing the question concerning what people can encode about natural textures in VSTM requires knowledge of the statistical structure of natural textures (Portilla & Simoncelli, 2000) and experimentally determining what types of statistics are encoded by subjects using texture stimuli that look reasonably, if not exactly, like natural textures. Lastly, it is important to note that many of the questions raised here would not even be raised were it not for a consideration of whether or not experimental results obtained under ecologically unrealistic conditions generalize to ecologically realistic conditions.

How to extend models of VSTM to natural stimuli

In this section, we discuss how models of VSTM can be extended to ecologically realistic stimuli and some implications of such an extension. We argue that it is possible to move beyond simple, artificial stimuli with unnatural stimulus statistics used in most VSTM studies and do at least partial justice to the richness and complexity of visual stimuli that are encoded in VSTM (see Fig. 1).

The plan of the present section is as follows. We first review evidence indicating that visual perception and VSTM share common neural and representational substrates presumably residing in visual cortical areas (Continuity Between Visual Perception and Visual Memory). We then review experimental and theoretical findings suggesting that the visual system is adapted to the statistical regularities of visual stimuli in the environment and that knowledge of these regularities aids both visual perception and visual memory (Benefits of Learning a Good Model of the Environment). To extend models of VSTM to natural stimuli, we then take our cue from the evidence for common representational substrates for visual perception and VSTM and suggest that knowledge of statistical regularities in the natural environment can be captured with probabilistic generative models. Moreover, we argue that both visual perception and VSTM can be regarded as probabilistic inference problems on the same generative model (Probabilistic Generative Models for Capturing Knowledge of Statistical Regularities in Natural Stimuli). Finally, we discuss some consequences of possible mismatches between the actual stimulus statistics used in an experiment and an observer's internal model of the same stimuli (Model Mismatch, or What Goes Wrong When the Observer's Internal Model Does Not Match the Actual Stimulus Distribution?).

Continuity between visual perception and visual memory

Recent results from neuroimaging and psychophysics provide converging evidence for a continuity between visual perception and VSTM. Several studies have demonstrated that the contents of VSTM during maintenance can be reliably decoded from visual cortical areas that are also involved in the perceptual encoding of the same stimuli (Christophel, Hebart, & Haynes, 2012; Emrich, Riggall, LaRocque, & Postle, 2013; Harrison & Tong, 2009; Riggall & Postle, 2012; Serences, Ester, Vogel, & Awh, 2009). Kang, Hong, Blake, and Woodman (2011) showed that a motion direction maintained in VSTM can interact with a visually perceived motion direction to cause a motion repulsion effect that occurs in the same way as when two visually perceived motion directions interact. Saad and Silvanto (2013) showed that maintenance in VSTM can cause adaptation effects similar to the effects caused by prolonged visual stimulation in the tilt aftereffect. Montaser-Kouhsari and Carrasco (2009) showed that discrimination performances in perceptual and VSTM tasks are affected in very similar ways by heterogeneities at isoeccentric locations in the visual field. These studies suggest that VSTM representations and visual perceptual representations share common substrates residing in visual cortical areas.

The idea that perception and short-term memory in general (and visual perception and VSTM in particular) have common neural and representational substrates has a long history in memory research. For example, a prominent theory of working memory holds that short-term maintenance of information consists of temporary reactivation of representations in long-term memory thought to be implemented in posterior cortical areas that are also involved in the initial perceptual encoding (Cowan, 1995; Fuster, 1997; Ruchkin, Grafman, Cameron, & Berndt, 2003). Although not uncontroversial (Baddeley, 2003), there is considerable evidence for this theory, not only for the case of short-term maintenance of visual

information, but also for short-term maintenance of information in other modalities as well (see Postle, 2006, for a review).

Consistent with this evidence for a common neural and representational substrate for visual perception and VSTM, we argue below that both visual perception and VSTM can be formulated as probabilistic inference problems on the same richly structured probabilistic generative models capturing the statistical structure of natural stimuli, with the only difference between the two being that inference in VSTM is based on noisier sensory information due to disruption of the visual signals from the environment. This idea is explained in more detail below (see the section titled Probabilistic Generative Models for Capturing Knowledge of Statistical Regularities in Natural Stimuli).

Benefits of learning a good model of the environment

A crucial observation about the visual system is that, through experience-dependent visual development and learning, people acquire sophisticated internal models of the types of visual stimuli they encounter in their natural visual environment. The natural visual environment is very far from random. It exhibits a rich set of statistical regularities ranging from low-level regularities between luminance values at nearby locations, or between orientations or spatial frequencies at different locations, to mid-level regularities between surface properties or mid-level features of natural objects, to high-level regularities between objects in different natural scene categories. The visual system adapts to these statistical regularities, which makes it more efficient at performing perceptual tasks in the natural environment.

Learning a good internal model of the statistical regularities in the natural environment has several benefits. First, in any task, optimal performance can be achieved only if the observer's internal model matches the actual statistical structure of the stimuli used in the task. For natural tasks involving natural stimuli, this corresponds to the environmental statistics of the relevant stimuli. In many psychophysical tasks, people, in fact, do perform better when visual stimuli are consistent with natural scene statistics than when they are not, suggesting that the visual system is adapted to natural scene statistics (Girshick, Landy, & Simoncelli, 2011; Knill, Field, & Kersten, 1990; Parragha, Troscianko, & Tolhurst, 2000; Stocker & Simoncelli, 2006; Yuille, Fang, Schrater, & Kersten, 2004). In general, the closer the observer's internal model is to natural stimulus statistics, the better the observer's performance will be in natural tasks involving those stimuli. Deviations in the observer's internal model from the natural stimulus statistics (referred to as "model mismatch" in Orhan & Jacobs, 2014) lead to suboptimal performance in natural tasks.

Second, related to the first point, learning the statistical regularities in the natural environment enables the visual system to tolerate significant amounts of noise in visual input. This might explain, for example, the reliable extraction of basic-level category information and substantial amounts of visual detail from very brief presentations of natural images (Fei-Fei, Iyer, Koch, & Perona, 2007; Serre, Oliva, & Poggio, 2007; Sharan, Rosenholtz, & Adelson, 2009), even without the benefit of attention (Li, VanRullen, Koch, & Perona, 2002).

Third, learning the statistical regularities of the natural environment is necessary for efficient encoding of information under resource constraints—that is, for encoding as much information as possible about the environment, subject to inevitable biological constraints on the representational systems of the organism. This idea underlies the influential "efficient coding hypothesis" according to which the goal of early visual processing is to maximize information transmission by reducing the redundancy inherent in natural visual signals due to the rich variety of statistical regularities they exhibit (Barlow, 1961; Geisler, 2008; Simoncelli & Olshausen, 2001).

The idea of efficient coding of information is closely related to the well-known notion of "chunking" in the short-term memory literature (Miller, 1956). Chunking can be considered as a form of redundancy reduction or compression where one forms efficient representations of more frequently occurring, or more familiar, stimuli. This kind of efficient allocation of available resources might lead to an apparent increase in memory performance or in the number of objects represented, without any actual increase in the amount of resources. Two recent studies provided experimental demonstrations of this idea in VSTM. Brady, Konkle, and Alvarez (2009) showed that subjects can quickly learn statistical regularities between color pairs to encode more colors in simple visual displays consisting of such color pairs. Sims, Jacobs, and Knill (2012) showed that subjects can adapt to a decrease in the variance of the stimulus distribution to increase the precision of their memories. Both studies found that these effects were attributable not to an increase in the available resources, but to the efficient allocation of resources due to adaptation to the stimulus statistics. It is important to note that both studies used artificial stimuli with simple forms of statistical regularities, suggesting that these efficient coding effects are likely to play a more significant role in natural stimuli displaying a much richer set of statistical regularities.

Probabilistic generative models for capturing knowledge of statistical regularities in natural stimuli

How can knowledge of statistical regularities in the environment be incorporated into models of VSTM? Here, we argue that knowledge of these statistical regularities can be captured

with richly structured probabilistic generative models and that visual perception and VSTM can both be regarded as probabilistic inference problems on the same generative model, consistent with the evidence for a continuity between visual perception and VSTM (as reviewed in the section titled Continuity Between Visual Perception and Visual Memory).

We first give a brief overview of inference in probabilistic generative models. Figure 2 schematically illustrates a hierarchical probabilistic generative model that can be used to represent the statistical structure of natural scenes and to capture observers' knowledge of this statistical structure (for more concrete examples of hierarchical probabilistic models applied to capture the statistical structure of natural scenes, see Sudderth, 2006). In this model, natural scenes are modeled in terms of hierarchically organized units or variables. Variables at each level are modeled as configurations of variables at the next level below. Thus, scenes are modeled as configurations of different objects, objects as configurations of parts, and so forth, down to the lowest level features. It is important to emphasize that all variables in the model are probabilistic variables, and hence, all are described by probability distributions. This is a crucial property that distinguishes these models from earlier hierarchical models of the structure of natural objects and scenes (Biederman, 1987; Marr, 1982; McClelland & Rumelhart, 1981). The statistical structure of natural scenes can then be captured by a complex joint probability distribution p(Scenes, Objects, . . . , Local features) that describes all the statistical regularities in natural scenes from the lowest to the highest level regularities. For instance, this distribution characterizes what kinds of objects tend to occur in which configurations in natural scenes of a given type, as well as the appearance of those objects. The observer acquires this "prior" distribution, p(Scenes, Objects, . . . , Local features), through extensive prior experience with the environment.

What happens when the observer is presented with a natural scene? It is assumed that when a scene is presented to the observer, the observer does not have direct access to any of the variables represented in the model but, rather, only to noisy measurements or observations of the lowest level variables in the scene. In the schematic model shown in Fig. 2, these correspond to the variables representing the local features. The observer's noise-corrupted measurements or observations of these features are represented by the shaded nodes in Fig. 2. These are the only directly observable variables in the model. All the other variables are latent or unobservable and have to be inferred probabilistically based on the directly observable variables. Intuitively, this means that when presented with a natural scene, the observer does not directly observe the objects in the scene, their locations, parts composing the objects, and so forth but, rather, indirectly (and probabilistically) infers these from noisy observations of lower level features. This is achieved through probabilistic inference by

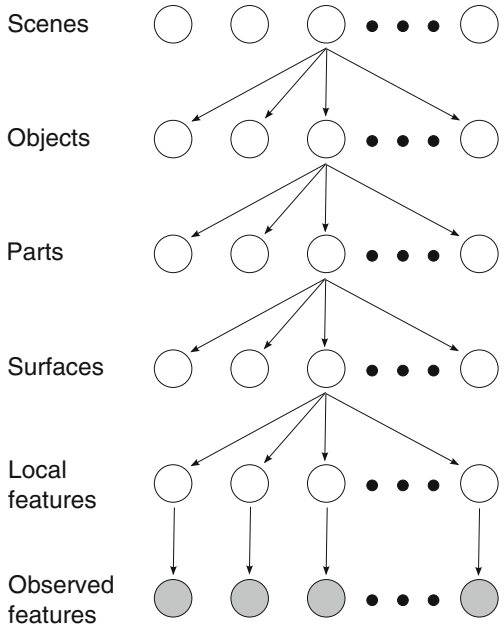

Fig. 2 A schematic illustration of a hierarchical probabilistic generative model capturing the statistical structure of natural scenes. The shaded nodes at the bottom represent the only observable variables in the model. The other nodes all represent latent or unobservable variables

combining the prior distribution p(Scenes, Objects, . . . , Local features) with the noisy measurements of the lowest level features to compute posterior distributions over all the latent variables in the model. The posterior distributions can then be used to make point estimates of the latent variables.

The following properties apply generally to all hierarchical probabilistic generative models (hence, they do not depend on the particulars of the model shown in Fig. 2):

Property 1: Measurement noise degrades the quality of all posterior distributions in the model. Increasing the measurement noise leads to posterior distributions with lower precision.

Property 2: Measurement noise affects the representations (i.e., posterior distributions) of lower level variables more than those of higher level variables.

Property 3: Representations of higher probability regions under the prior are enhanced relative to the representation of low-probability regions; that is, stimuli that are more likely according to the prior lead to posteriors with higher precision.

We suggest that both visual perception and encoding in VSTM can be understood as probabilistic inference problems, similar to the inference problem demonstrated in the schematic model in Fig. 2. In both cases, the visual system probabilistically infers the values of certain latent variables given noisy measurements or observations of some basic visual features. In different applications, the latent variables of interest might differ. The observable variables might also differ. In

some cases, as in Fig. 2, the observable variables might correspond to some low-level image features and their locations, whereas, in other cases, they might correspond to higher level, more abstract visual features. More realism can be introduced into the model by taking into account the fact that perceptual resolution in the visual periphery—or in other words, the precision of measurements obtained from the periphery—is lower than in the fovea.

Crucially, we propose that differences between visual perception and VSTM can be captured entirely by the quality (or precision) of the observable variables. This is the main claim of this section. In general, the quality of the observable variables might depend on the presentation duration of the scene (with longer presentation durations leading to more precise measurements; Bays, Gorgoraptis, Wee, Marshall, & Husain, 2011), as well as the length of the potential delay interval during which visual information has to be maintained in memory (precision of measurements degrading with the length of the delay interval).[3] For visual perception, the scene remains visible throughout (and there is no delay interval); hence, the measurements of the observable variables are higher in quality or less noisy. For VSTM, the scene is removed for a certain delay interval after a brief presentation, so the measurements of the observable variables are expected to be noisier. This greater uncertainty in the values of the observable variables "percolates" through the model (by Property 1 above) and can have a large effect on the posterior distributions over the latent variables depending on the strength of the prior (for example, how strong the statistical regularities in natural scenes of this type are), as well as on the noise level (which, as mentioned, depends on the presentation time, the length of the delay interval, and possibly other task variables). Other than this difference in the precision of the measurements of observable variables, however, the representation of prior knowledge about the scene and the probabilistic inference process involved in determining the values of the latent variables in the scene are identical in visual perception and VSTM. That is, there are no qualitative differences between visual perception and VSTM, just quantitative differences (that may still have a large impact on behavior depending on the strength of the prior, as mentioned above). This account is consistent with the results reviewed above suggesting common representational (and neural) substrates for visual perception and VSTM.

[3] See Magnussen and Greenlee (1999) for evidence against degradation of memory precision with delay interval. However, these researchers only consider the encoding of single items in VSTM. Results may be different for multiple items. For example, Zhang and Luck (2009) presented evidence for deterioration of memory performance with delay interval in a recall task with three items. Zhang and Luck (2009) interpreted this result as an increase in the probability of failure to encode items, rather than as a gradual decrease in memory precision.

It seems clear that if the study of VSTM is to be extended to a richer set of stimuli that are more representative of the natural visual environment (Fig. 1), models like the one shown in Fig. 2 that can capture the rich statistical regularities in natural environments, and people's prior knowledge of those regularities, will be essential (see Brady et al., 2011, for a similar point).

Some experimental details

How can hierarchical probabilistic models such as the one shown in Fig. 2 be used to model and interpret data from recall or recognition experiments using natural scenes? In experiments using natural scenes, memory for lower level variables, such as the color of an object, can be probed directly as in standard VSTM recall studies. Memory for higher level variables, such as objects, can be probed using either a change detection or a two-alternative forced choice (2AFC) task. In the former case, for example, a randomly chosen object might disappear on half of the trials, and subjects might be asked to indicate whether any change has taken place in the scene. In this case, the probabilistic model would need to be duplicated, one copy for the target scene, the other one for the probe scene, with the two copies connected by a binary variable *change* indicating whether a change has taken place or not (as in Ma & Huang, 2009). The probabilistic inference of latent variables including the variable of interest, *change*, would proceed exactly as before. In the latter case (a 2AFC task), after the target scene is presented, a certain location in the scene might be probed, and subjects might be asked to indicate which one of the two alternative objects was present at that location in the target scene. This can be modeled by evaluating the probabilities of the two alternative objects under the posterior distributions over all the variables corresponding to the object at the target location in the original scene. The alternative with the higher posterior probability can then be chosen as the model's response.

Implications for the debate on the nature of capacity limits

Considered from a perspective that takes into account the richness and complexity of stimuli that can be encoded in VSTM (Fig. 1), current discrete or continuous resource models that dominate the discussion in the VSTM literature prove to be too simplistic to capture the nature of capacity limitations in VSTM. Richly structured probabilistic generative models, such as the one shown in Fig. 2, that can capture subjects' knowledge of statistical regularities in natural stimuli are better suited to modeling what people can and do encode about such stimuli in VSTM. In the model shown in Fig. 2, there are two distinct limitations on the model's performance. The first one is the accuracy of the model's prior. This factor is generally ignored in the models of capacity limitations in

VSTM. As was argued above, the closer the prior is to the actual environmental statistics of the relevant variables, the better the model's performance will be. Crucially, differences in the priors of two models will be reflected in the models' performances even when the noise in the observable variables is the same in both models. The second factor limiting the performance of a model is the quality of the observable variables, with less noise in the observable variables yielding better performance in general.

It may be possible to retain some aspects of both discrete and continuous resource models of capacity limitations in complex probabilistic generative models such as the one shown in Fig. 2. For example, the model in Fig. 2, on the one hand, posits variables corresponding to objects, and, as was noted in Property 2 above, these higher level variables in general tend to be less affected by measurement noise in the observable variables, having a more all-or-none character than the lower level variables, such as the features making up the objects. This property resembles the encoding of objects in the item-limit models of VSTM (e.g., Luck & Vogel, 1997). On the other hand, the representation of all variables in the model, including those of objects or object categories, is probabilistic (in terms of a posterior distribution over the relevant variable), and the quality of these probabilistic representations, especially that of lower level features, is degraded with measurement noise in the observable variables, resembling the encoding of features in continuous resource models.

Other types of visual information

So far, we have only discussed, in broad outlines, how the encoding of natural scenes in VSTM can be modeled with hierarchical probabilistic generative models. How about other types of visual information—for example, visual textures (Fig. 1b)?

Probabilistic generative models provide a rich, versatile, and quantitative format for representing people's prior information or expectations in a wide range of different perceptual domains (Kersten & Yuille, 2003; Knill & Richards, 1996). In different cases, this prior information can be regarded either as the product of lifelong adaptation to statistical properties of the natural environment or as shorter-term adaptations acquired by subjects during the course of an experiment (Brady et al., 2009; Sims et al., 2012). For different types of visual information, people's prior expectations can be captured by different probabilistic generative models. For example, a generative model that can be used for modeling people's prior information about visual textures (Balas, 2006; Balas, Nakano, & Rosenholtz, 2009; Portilla & Simoncelli, 2000) will be different from a generative model that may be used to capture their prior information about the shapes of objects belonging to different object categories (Sudderth, 2006) or their prior information about the structure of different natural scene categories (Sudderth, Torralba, Freeman, & Willsky, 2008). Crucially, using probabilistic generative models—commonly employed for representing prior information or expectations in visual perception (Kersten & Yuille, 2003; Knill & Richards, 1996)—to capture the representational structure of VSTM would unify the representational formats used in visual perception and VSTM (see the section titled Continuity Between Visual Perception and Visual Memory).

Recent evidence suggests that, even in experiments using simple, artificial items with unnatural stimulus statistics (e.g., oriented Gabors with independently drawn orientations), subjects' internal models display a rich, hierarchical structure (Brady & Tenenbaum, 2013; Orhan & Jacobs, 2013) that does not match the simple, unstructured models used by the experimenter to generate the stimuli. We recently proposed a probabilistic clustering model that attempts to characterize aspects of the internal model that subjects use to encode simple displays typically used in VSTM experiments (Orhan & Jacobs, 2013). This model implements the assumption that stimuli are generated in clusters, where the number of clusters and the assignment of stimuli to clusters are uncertain and have to be inferred probabilistically from the display, even if the actual generative process used by the experimenter does not involve any clusters. The model predicts biases in memory estimates of stimuli based on their VSTM representations, as has been observed previously in a number of VSTM studies (Brady & Alvarez, 2011; Huang & Sekuler, 2010). Due to the coupling of stimuli through the probabilistic clustering assumption, the model also predicts dependencies between memory estimates of different stimuli that should decrease with the difference between their feature values. We experimentally confirmed this novel prediction of the model using both continuous recall (delayed estimation) and change-detection tasks with small set sizes.

Similarly, Brady and Tenenbaum (2013) recently showed that hierarchical models incorporating the encoding of summary statistics of stimuli in a given display, as well as the perceptual grouping of stimuli, are necessary to account for display-by-display performance in change-detection tasks. Brady and Tenenbaum demonstrated that this is true even in experiments where stimulus values are drawn independently on each trial and, hence, contain no statistical structure.

Model mismatch, or what goes wrong when the observer's internal model does not match the actual stimulus distribution?

What happens when an observer's internal model does not match the actual stimulus statistics used in an experiment? Such mismatches can arise because, as we argued above (Benefits of Learning a Good Model of the Environment), the visual system is adapted to the statistical structure of

the natural environment. However, most VSTM studies use stimuli that are unstructured, such as stimuli in which objects' feature values are drawn independently from a uniform distribution. In general, if an observer's internal model is adapted to represent stimuli with a statistical structure that differs from the actual statistical structure of the stimuli used in an experiment and the observer's internal model has a limited ability to adapt to the statistical structure of the experimental stimuli during the course of an experiment, then this creates a "model mismatch." Importantly, model mismatch negatively impacts the observer's performance in a VSTM task.

A consideration of model mismatch might affect the interpretation of results obtained from VSTM studies (Orhan & Jacobs, 2014). For example, the decline in precision with set size found in standard VSTM studies is generally interpreted as a signature of VSTM resource or capacity limits. However, in a computational study, we have recently demonstrated that the same result can be accounted for as a consequence of model mismatch without assuming any resource limitations (Orhan & Jacobs, 2014). The basic idea is illustrated in Table 1 with a toy example.

In this example, we imagine a world consisting of objects that can take one of two values: $s_i = 0$ or $s_i = 1$ (the example can be easily generalized to objects that can take a continuum of values). Suppose, first, that only one object, s_1, is presented to the observer. We assume that $s_1 = 0$ and $s_1 = 1$ are equally likely in the environment, and the observer allocates k units of resources for encoding either value of s_1, where the resources can be thought of as the number of spikes, for instance (Ma & Huang, 2009). The expected amount of resources allocated for representing a single item is thus equal to k.

Now imagine that a second object, s_2, is presented to the observer alongside the first object, s_1. Suppose that there are correlations in the observer's natural environment such that stimulus configurations where both objects have the same value ($s_1 = 0$, $s_2 = 0$ or $s_1 = 1$, $s_2 = 1$) are more common than stimulus configurations where the objects have different values ($s_1 = 0$, $s_2 = 1$ or $s_1 = 1$, $s_2 = 0$). In particular, suppose that each of the more frequent stimulus configurations occurs with probability .4, and that each of the less frequent configurations occurs with probability .1. Critically, we assume that the observer's visual system adapts to this statistical regularity in the environment by allocating more resources for representing the more frequent configurations and fewer resources for representing the less frequent configurations (similar to the observation made in Property 3 above; see also Ganguli & Simoncelli, 2010). For concreteness, we suppose the observer allocates $2.4k$ units of resources for representing each of the more frequent configurations and $0.4k$ units of resources for representing the less frequent ones. The expected total amount of resources allocated for representing a two-object display is then $(0.4 \times 2.4k) + (0.4 \times 2.4k) + (0.1 \times 0.4k) + (0.1 \times 0.4k) = 2k$—that is, twice the amount of resources expended for representing a single object. Therefore, the resources increase linearly with set size. In other words, the amount of resources per item stays the same.

Table 1 Possible configurations for one ($N = 1$) and two ($N = 2$) objects (left), resources allocated to each configuration and probabilities of configurations (right)

$N = 1$	
$s_1 = 0$	k
	$p_0 = 0.5$
$s_1 = 1$	k
	$p_1 = 0.5$
(Expected) total resources: k	
(Expected) resources per item: k	
$N = 2$	
$s_1 = 0, s_2 = 0$	$2.4k$
	$p_{00} = 0.4$
$s_1 = 1, s_2 = 1$	$2.4k$
	$p_{11} = 0.4$
$s_1 = 0, s_2 = 1$	$0.4k$
	$p_{01} = 0.1$
$s_1 = 1, s_2 = 0$	$0.4k$
	$p_{10} = 0.1$
(Exp.) total resources: $2k$	
(Exp.) resources per item: k	
$N = 2$ (mismatched stimuli)	
$s_1 = 0, s_2 = 0$	$2.4k$
	$p_{00} = 0.25$
$s_1 = 1, s_2 = 1$	$2.4k$
	$p_{11} = 0.25$
$s_1 = 0, s_2 = 1$	$0.4k$
	$p_{01} = 0.25$
$s_1 = 1, s_2 = 0$	$0.4k$
	$p_{10} = 0.25$
(Exp.) total resources: $1.4k$	
(Exp.) resources per item: $0.7k$	

But now consider what happens when we present the observer with stimuli drawn from a distribution that his or her visual system is not adapted to. Assume, for example, that all four configurations are presented to the observer with equal probability. The expected total amount of resources allocated for representing a two-object display is now $(0.25 \times 2.4k) + (0.25 \times 2.4k) + (0.25 \times 0.4k) + (0.25 \times 0.4k) = 1.4k$, and the expected resources per item is therefore $0.7k$. This represents a decline in the amount of resources per item from $N = 1$ (i.e., a set size effect).

This example has important implications for our understanding of VSTM. Recall that set-size effects have conventionally been interpreted as indicating that VSTM has a fixed amount of memory resources. We claim, however, that there is

an alternative interpretation of set size effects. As our example illustrates, these effects might, instead, be a consequence of using a stimulus distribution that does not match the natural stimulus distribution that the observer is adapted to.

To reiterate our argument, even if VSTM resources increase linearly with set size (meaning that resources are effectively unbounded over the range of set sizes tested), the visual system might be allocating most of these resources to stimulus configurations that are common in the natural environment but not proportionately common in the experiment. This leads to inefficient use of resources (with respect to the stimulus distribution used in the experiment), which may cause a decline in memory precision in the experiment (even when the amount of resources increases proportionally to the set size). Thus, we suggest that it is an open empirical question as to what extent set size effects observed in VSTM studies are caused by genuine resource limitations versus inefficient use of resources due to model mismatch.

The toy example described above also suggests that using artificial stimuli and unnatural stimulus statistics might lead researchers to underestimate the capacity of VSTM. In information theory, the capacity of an information channel is defined as the maximum mutual information that can be achieved between an input ensemble X and the output of the channel Y, where the maximum is taken with respect to all possible probability distributions over the input ensemble (MacKay, 2003). If the VSTM system is abstractly conceived of as an information channel (Sims et al., 2012), then X might correspond to all possible visual stimuli, and Y might correspond to all possible responses of neurons or neural populations underlying the VSTM system, presumably the visual cortical areas where the initial perceptual encoding takes place (see the section titled Continuity Between Visual Perception and Visual Memory). If it is assumed, as seems plausible, that the responses of these visual cortical areas are adapted to natural stimulus statistics (Benefits of Learning a Good Model of the Environment), then the optimal input distribution, $P(X)$, will correspond to the statistics of the visual stimuli X in the natural visual environment. Any other input distribution will drive the information rate below the capacity of the channel. Using artificial stimuli with unnatural stimulus statistics in VSTM studies might yield a similar underutilization of VSTM, leading researchers to underestimate its capacity.

A possible objection

It could be argued that demonstrations of change blindness with natural scenes (Rensink, 2002; Simons & Levin, 1997) provide evidence against the suggestion that the capacity estimates for VSTM would be significantly boosted with the use of natural stimuli. However, the precise causes of change blindness in natural scenes and its implications for our understanding of VSTM are currently a matter of debate. For instance, researchers have suggested that failure to detect (sometimes quite salient) changes in natural scenes might be due not to apparent capacity limitations in visual attention or in VSTM, but, rather, to simpler factors that might have nothing to do with capacity limitations of visual attention or VSTM, such as insufficient time given subjects to encode the details of the scene (Brady, Konkle, Oliva, & Alvarez, 2009) or a lack of fixation near target objects (Hollingworth, 2006; Hollingworth & Henderson, 2002). Therefore, a failure to detect changes in natural scenes in change blindness studies does not necessarily imply a severe capacity limitation in visual attention or in VSTM.

Conclusion

Recent findings from neuroimaging and psychophysics suggest common neural and representational substrates for visual perception and VSTM. Visual perception is adapted to a rich set of statistical regularities present in the natural visual environment. The continuity between visual perception and VSTM suggests that VSTM is adapted to these statistical regularities too. Thinking about the operation of VSTM in such environments may lead to a reevaluation of the results obtained from experiments using artificial stimulus statistics and tasks and may inspire the development of ecologically more realistic, complex models incorporating prior knowledge of natural scene statistics.

Acknowledgments This work was supported by research grants from the National Science Foundation (DRL-0817250) and the Air Force Office of Scientific Research (FA9550-12-1-0303).

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale: Erlbaum.

Baddeley, A. D. (2003). New data: Old pitfalls. *Behavioral and Brain Sciences, 26,* 729–730.

Balas, B. (2006). Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Research, 46,* 299–309.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistical model of peripheral vision explains visual crowding. *Journal of Vision, 9,* 1–9.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5,* 617–629.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory Communication* (pp. 217–234). Cambridge: The MIT Press.

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), 1–11.

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011a). Temporal dynamics of encoding, storage and reallocation of visual working memory. *Journal of Vision, 11*(10, 6), 1–15.

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

Bays, P. M., Wu, E. Y., & Husain, M. (2011b). Storage and binding of object features in visual working memory. *Neuropsychologia, 49,* 1622–1631.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science, 22*(3), 384–392.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009a). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Pscyhology: General, 138,* 487–502.

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), 1–34.

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

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher-order regularities into working memory capacity estimates. *Psychological Review, 120*(1), 85–109.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50,* 255–272.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217.

Christophel, T. B., Hebart, M. N., & Haynes, J. D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *Journal of Neuroscience, 32,* 12983–12989.

Cowan, N. (1995). *Attention and memory: an integrated framework*. New York: Oxford University Press.

Emrich, S. M., & Ferber, S. (2012). Competition increases binding errors in visual working memory. *Journal of Vision, 12*(4, 12), 1–16.

Emrich, S. M., Riggall, A. C., LaRocque, J. J., & Postle, B. R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *Journal of Neuroscience, 33,* 6516–6523.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7*(1, 10), 1–29.

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision, 11*(12), 1–12.

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of working memory. *Nature Communications, 3,* 1129.

Fuster, J. M. (1997). Network memory. *Trends in Neurosciences, 30,* 451–459.

Ganguli, D., & Simoncelli, E.P. (2010). Implicit encoding of prior probabilities in optimal neural populations. *Advances in Neural Information Processing Systems*, 23.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology, 59,* 167–192.

Geisler, W. S., Perry, J. S., & Ringach, D. (2009). Natural systems analysis. *Visual Neuroscience, 26,* 1–3.

Giese, M., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4,* 179–192.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience, 14*(7), 926–932.

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458,* 632–635.

Hayhoe, M. (2000). Vision using routines: a functional account of vision. *Visual Cognition, 7,* 43–64.

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science, 1,* 189–202.

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.

Hollingworth, A. (2006). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual Cognition, 14,* 781–807.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 113–136.

Huang, J., & Sekuler, R. (2010). Distortions in recall from visual memory: Two classes of attractors at work. *Journal of Vision, 10*(24), 1–27.

Kang, M.-S., Hong, S. W., Blake, R., & Woodman, G. F. (2011). Visual working memory contaminates perception. *Psychonomic Bulletin & Review, 18,* 860–869.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology, 13,* 1–9.

Knill, D. C., Field, D. J., & Kersten, D. (1990). Human discrimination of fractal images. *Journal of the Optical Society of America A, 7,* 1113–1123.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Land, M. F., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception, 28,* 1311–1328.

Lazebnik, S., Schmid, C., & Ponce, J. (2005). A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27,* 1265–1278.

Li, F.-F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, 99,* 8378–8383.

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

Ma, W. J., & Huang, W. (2009). No capacity limit in attentional tracking: evidence for probabilistic inference under a resource constraint. *Journal of Vision, 9*(11), 1–30.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Magnussen, S., & Greenle, M. W. (1999). The psychophysics of perceptual memory. *Psychological Research, 62,* 81–92.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: part 1. An account of basic findings. *Psychological Review, 88*(5), 375–407.

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.

Montaser-Kouhsari, L., & Carrasco, M. (2009). Perceptual asymmetries are preserved in short-term memory tasks. *Attention, Perception, & Psychophysics, 71,* 1782–1792.

Neisser, U. (1976). *Cognition and reality: principles and implications of cognitive psychology.* WH Freeman.

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review, 120*(2), 297–328.

Orhan, A.E., & Jacobs, R.A. (2014). Are performance limitations in visual short-term memory tasks due to capacity limitations or model mismatch? Manuscript under review.

Parragha, C. A., Troscianko, T., & Tolhurst, D. J. (2000). The human visual system is optimised for processing the spatial information in natural visual images. *Current Biology, 10,* 35–38.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40,* 49–71.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience, 139,* 23–38.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*(2), 275–280.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53,* 245–277.

Riggall, A. C., & Postle, B. R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. *Journal of Neuroscience, 32*(38), 12990–12998.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences, 105,* 5976–5979.

Ruchkin, D. S., Grafman, J., Cameron, K., & Berndt, R. S. (2003). Working memory retention systems: A state of activated long-term memory. *Behavioral and Brain Sciences, 26,* 709–728.

Saad, E., & Silvanto, J. (2013). How visual short-term memory maintenance modulates subsequent visual aftereffects. *Psychological Science, 24,* 803–808.

Schwarzkopf, D. S., & Kourtzi, Z. (2008). Experience shapes the utility of natural statistics for perceptual contour integration. *Current Biology, 18,* 1162–1167.

Serences, J. T., Ester, E., Vogel, E., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science, 20*(2), 207–214.

Serre, J. T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, 104,* 6424–6429.

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2009). Material perception: What can you see in a brief glance? *Journal of Vision, 9*(8), 784.

Simoncelli, E. P., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24,* 1193–1216.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1,* 261–267.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review, 119,* 807–830.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9*(4), 578–585.

Sudderth, E. B. (2006). *Graphical models for visual object recognition and tracking (Unpublished doctoral dissertation)*. Cambridge: MIT.

Sudderth, E. B., Torralba, A., Freeman, W., & Willsky, A. (2008). Describing visual scenes using transformed parts and objects. *International Jouranl of Computer Vision, 77,* 291–330.

Van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, 109*(22), 8780–8785.

Velisavljević, L., & Elder, J. H. (2008). Visual short-term memory for natural scenes: Effects of eccentricity. *Journal of Vision, 8*(4, 28), 1–17.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), 1120–1135.

Xiao, J., Hays, J., Ehinger, K., Oliva, A., & Torralba, A. (2010). *SUN database: Large scale scene recognition from abbey to zoo. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos: IEEE Computer Society.

Yuille, A.L., Fang, F., Schrater, P., & Kersten, D. (2004). Human and ideal observers for detecting image curves. *Advances in Neural Information Processing Systems*, 17.

Zhang, P. H., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453,* 233–235.

Zhang, P. H., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science, 20*(4), 423–428.