Perception, 2014, volume 43, pages 663–676

doi:10.1068/p7719

Mean size as a unit of visual working memory

Hee Yeon Im[1,2], Sang Chul Chong[2,3]§
[1] Department of Psychological and Brain Sciences, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218-2686, USA; [2] Graduate Program in Cognitive Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, South Korea; [3] Department of Psychology, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, South Korea;
e-mail: scchong@yonsei.ac.kr

Received 12 February 2014, in revised form 24 June 2014

Abstract. Visual environments often contain multiple elements, some of which are similar to one another or spatially grouped together. In the current study we investigated how one can use perceptual groups in representing ensemble features of the groups. In experiment 1 we found that participants' performance improved when items were easily segmented by a grouping cue based on proximity, suggesting that spatial grouping facilitates extracting and remembering ensemble representations from visual arrays consisting of multiple elements. In experiment 2 we found that spatial grouping improved performance only when the grouped subsets were tested for the memory task, whereas it impaired performance when other subsets that were not grouped were tested, suggesting that the benefit from grouping may come from better extraction for storage, rather than later decision processes such as accessibility. Taken together, our results suggest that perceptual grouping of multiple items by proximity facilitates extraction of ensemble statistics from groups of items, enhancing visual memory of the ensembles in a visual array.

Keywords: grouping, mean size, visual working memory

1 Introduction

The human visual system has a limited capacity. Much evidence suggests that the visual system can select and hold only a limited number of objects (three or four) at any one time (Cowan, 2001; Luck & Vogel, 1997; Pashler, 1988; Pylyshyn & Storm, 1988; Scholl, 2001; Simons & Levin, 1997; Vogel, Woodman, & Luck, 2001). This presents a puzzle for understanding our everyday visual experiences. How is it possible for us to acquire a vivid and rich impression of details of the visual world, if no more than three or four items can be held in our memory at once?

One possible answer is that the limited capacity of visual working memory (VWM) is accompanied by impressive representational flexibility. To achieve this flexibility, different types of entities can be represented at different hierarchical levels (Brady, Konkle, & Alvarez, 2011; Feigenson, 2011). These different entity types—such as individual objects, sets, or ensembles—can then function as independent items. As most visual environments are hierarchically structured, hierarchical representation by the visual system may be an adaptive and efficient strategy. Let us take trees as an example. As we navigate through the visual environment, we may perceive, at different times, a single tree, groups of different kinds of trees, or the forest as a whole. We can then represent these different types of information: about a specific tree (eg height, color, or orientation of the tree), about groups of trees (eg average height or approximate number of sets of certain kinds of trees), and about the global properties of the forest (eg area, darkness, or density of the forest). By combining these types of representation or by flexibly shifting between them, the visual system can represent and maintain much more detail about the visual world, overcoming its limited capacity.

§ Corresponding author.

Representations of individual objects, groups, and ensemble statistics provide different aspects of the visual environment, playing distinct but complementary functions to enhance visual cognition. For example, global representation extracted from the whole visual scene (eg texture, ensemble statistics, or global scene layout) can facilitate recognition or detection of individual objects in the scene (for review, see Wolfe, Vo, Evans, & Greene, 2011). Together, for optimal encoding of visual scenes, one should take into account the nature of hierarchical structure that enables compressed memory representations.

Ensemble representation of a group of individual objects is one example of a higher order representation. Ensemble representation, such as the mean size of multiple items with different sizes, is computed from multiple individual measurements by collapsing across them (Ariely, 2001; Chong & Treisman, 2003). Ensemble representation is efficient and robust, providing information about the structure of the visual image that is more easily detected than information about individual elements of the image (Alvarez & Oliva, 2009; Victor & Conte, 2004).

Recent evidence shows that ensemble representations are constrained by the same VWM capacity limit that constrains representations of individual objects. For example, human observers (Feigenson, 2008; Halberda, Sires, & Feigenson, 2006; Poltoratski & Xu, 2013; Zosh & Feigenson, 2012) can represent the approximate numerosity of up to three sets of items at once, but not more than that. These results suggest that ensemble representations function as another ‘unit’ for memory, just as individual objects do.

Representing and storing an ensemble (eg average) of a set of objects can help the visual system to maintain and recall individual features of the elements. At the object level, only a few elements may be remembered; the rest may be missed completely due to the limited memory capacity. When attempting to recall missed objects, one would have to make random guesses, increasing the overall expected error. However, average information about the set can guide one to recall the missed object to some extent by retrieving values biased toward the average of the set. Say you try to remember the different colors of six disks in a visual array and report the colors you remember. If you remember only the colors of three disks and completely missed the others, you cannot avoid making extreme errors for the three forgotten colors (eg by choosing red for a blue disk). However, if you remember that the disks were in ‘cool’ colors on average (even if you cannot remember the exact colors for each disk), it is likely that you will reduce the overall error by choosing six colors from only the continuum of ‘cool’ colors and avoid ‘warm’ colors such as red or orange.

It has been shown empirically that ensemble representations affect memory of individual elements in a visual array in this way. For example, Brady and Alvarez (2011) presented circles that differed in size and color, and they had participants report the size of a designated circle after a one second delay. They found that participants’ responses were biased by the size of other circles in the same color set and by the size of all of the circles, suggesting that the mean size of a memory array influenced performance on the visual memory task about individual circles.

Set-based representation of elements is another type of representation that can be maintained in the visual memory as a separate entity. Multiple items can be grouped into a set and treated as one higher unit of memory. This set-based grouping involves binding multiple individuals into a single higher order group based on certain grouping principles such as similarity or spatial proximity, as demonstrated by the early gestalt psychologists (eg Beck, 1982; Koffka, 1935).

Just as with representations of individual objects and ensembles, VWM constrains set representations (Cowan, Nelson, Chen, & Rouder, 2004). The number of individual objects that can be grouped into a set never seems to exceed the working memory capacity limit of three or four items (Cowan et al., 2004; Feigenson & Halberda, 2008). Moreover, sets that are bound from multiple individuals can further be bound into a ‘superset’, and the number of sets that can be bound in this way seems to obey the same limiting principle (Chase & Ericsson, 1981). Taken together, these findings suggest that set-based representation functions in a similar manner as that for individual objects in the VWM.

Previous studies showed that set representation by grouping can increase the amount of information to be maintained in memory at different levels of visual processing. First of all, multiple features can be grouped and stored together as a single 'object'. For example, human observers can remember sixteen different features as well as they can remember only four features, if they form four conjunction objects each containing four features, as though dealing with four single-feature objects (Luck & Vogel, 1997; but see also Wheeler and Treisman, 2002). Moreover, multiple individual items also can be grouped into a set as a higher order unit of memory. For example, Xu and Chun (2007) showed that perceptual grouping enhanced VWM by allowing more visual elements to be remembered. Woodman, Vecera, and Luck (2003) have also shown that perceptual grouping influences what elements are stored in memory such that, when one element of a group was stored in working memory, other elements of the sample gestalt group were likely to be stored as well.

Together, ensemble representation and grouping are processes by which multiple elements are combined into a single higher order description, resulting in a hierarchical structure of VWM. Both processes, by providing representational flexibility, enable larger amounts of information to be encoded into a single higher order individual and maintained in VWM. In order to deal with complex visual scenes containing overflowing information, the visual system can flexibly shift between these different types of representation. Under the framework of this hierarchical structure of VWM, one can reasonably expect the different types of representation to interact closely with each other. To date, the influences of ensemble representation and set-based grouping on the representation of individual objects in visual memory have been explored only separately. In most visual environments, however, grouping and ensemble representation may occur simultaneously. How do these two types of higher order representation interact? Specifically, how does visual grouping affect the efficiency of ensemble representation? Addressing this question would allow us to better understand the hierarchical nature of visual representation of a scene.

Investigating how visual grouping affects ensemble representation would also enable us to understand how ensemble features are represented. Although many recent studies have shown that humans have a remarkable ability to represent ensemble features from briefly presented visual arrays containing multiple similar elements (eg Alvarez & Oliva, 2008; Ariely, 2001; Chong & Triesman, 2003; Im & Halberda, 2013), how they are represented remains controversial. Some authors argue that ensemble representation may be a form of late processing that utilizes high-level features of objects, such as emotions on faces (Haberman & Whitney, 2010), while others argue that ensemble representation may be the same as early texture perception (eg Dakin, Tibber, Greenwood, Kingdom, & Morgan, 2011). To empirically test this, we (Im & Chong, 2009) examined how the Ebbinghaus illusion influenced mean size computation of central circles, surrounded by either larger or smaller circles. Although the physical size of the circles to be averaged remained constant, observers' mean size judgment was biased by the illusion such that, when central circles were surrounded by smaller inducers, observers' mean computation of the central circles was significantly overestimated and vice versa, following the directions of the effect by the Ebbinhaus illusion. The authors suggested that computation of mean size is based on perceived size, after modulation of perceived size of individual items. The current study will add to this finding by addressing the question of whether ensemble representation is affected by the segmentation of groups of individual items in a visual array. If grouping affects the efficiency of ensemble representation, this would suggest that extraction and storage of ensemble representation requires the process of parsing or segmentation of 'groups' of individual elements in a visual array.

There is suggestive evidence that the way in which elements in a visual scene are grouped by perceptual cues affects the extraction of ensemble representation of a group of those elements. For example, it has been shown to be easier to calculate the mean of similar items (Ariely, 2001), as well as that the spatial configuration of elements of a group directly biases estimations of numerosity. For example, globally clustered items appear to be more numerous than the same number of items clustered into multiple subgroups (Frith & Frith, 1972), regularly arranged items look more numerous than randomly distributed items (Ginsburg, 1976; Taves, 1941), and random patterns look more numerous than clustered items (Ginsburg & Goldstein, 1987).

Given that perceptual grouping affects how ensemble representation is extracted, we predict that this relationship should be reflected in visual memory as well. We hypothesize that there is an accumulative facilitation in the representation of visual memory when observers can group items more easily to extract ensemble representations from the visual array. Therefore, more ensembles will be remembered from a visual array in which elements are easily grouped together. To investigate this hypothesis, we presented arrays containing 10 to 25 circles, differentiated by two to five different colors. After the array disappeared, the participants had to report which of the two probed sets had a larger mean size. In experiment 1 we sometimes grouped sets of arrays by proximity, and other times we did not. We found that the overall accuracy of mean size judgments was better for the grouped condition, suggesting that spatial grouping facilitates extracting and remembering ensemble representations from visual arrays consisting of multiple elements. In experiment 2 we grouped only subsets of arrays in some trials in order to further investigate the facilitation effect of grouping on ensemble representation. We found that grouping improved visual memory only when the grouped subsets were task-relevant, whereas it impaired performance when the grouped subsets were task-irrelevant. Taken together, our results suggest that perceptual grouping of multiple items by proximity facilitates visual memory of ensemble representation of groups in a visual array.

2 Experiment 1

In experiment 1 we investigated how many mean sizes observers could store in visual memory and how this limit on the number of mean representations was affected by perceptual grouping. We presented a stimulus display consisting of two to five spatially intermingled subsets of differently colored circles. After viewing the display, participants were presented with two probes corresponding to two specific subsets from the original display and asked to judge which of the two probed subsets had the larger mean size in the previous display. We used two different time intervals (0 and 1 s) between the stimulus display and the probe display to investigate how many mean sizes could be retained in visual memory from the visual array of spatially intermingled elements. We compared participants' accuracy on this spatially intermixed array with that on the visual array in which each group of elements was spatially grouped by proximity, and we investigated whether perceptual grouping improved participant accuracy in the visual memory task.

2.1 *Method*

2.1.1 *Participants.* Forty-four undergraduate students from Yonsei University participated in the experiment (twenty nine for course credit and fifteen for monetary compensation). The no-interstimulus interval (no-ISI) condition included fourteen participants, the 1 s interval condition included fifteen participants, and the grouped condition included fifteen participants. All participants had normal or corrected-to-normal vision and were unaware of the purpose of the study. The Institutional Review Board of Yonsei University approved the experimental protocol, and signed informed consent forms were obtained from all participants.

2.1.2 *Apparatus and stimuli.* The stimuli were created using MATLAB and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and presented on a linearized Samsung 21″ monitor driven by a Pentium VI computer. The frame rate of the monitor was 85 Hz. The participants were positioned approximately 90 cm from the display with their heads fixed on a forehead-and-chin rest. At this distance, 1 pixel on the display was approximately 0.013 deg.

The display contained two to five sets, with five circles of different sizes in each set (figure 1). To distinguish different sets, we used five colors (red, blue, yellow, green, and cyan). The luminance was 20.34 cd m^{-2} for red, 10.53 cd m^{-2} for blue, 92.87 cd m^{-2} for yellow, 73.25 cd m^{-2} for green, and 83.06 cd m^{-2} for cyan. Note that we randomly assigned five colors to each set in each trial. Thus, differences in luminance are not likely to have influenced performance. The luminance of the gray background was 56.01 cd m^{-2}. For each trial, the mean size of the smallest set was first randomly chosen from a uniform distribution ranging from 0.85 deg to 0.94 deg. On the basis of the smallest set size, mean sizes of other sets were determined in a multiplicative step of 10% size difference, yielding an overall range of mean sizes of subsets from 0.85 deg to 1.59 deg.

The array of possible circle locations was specified by an invisible 6 × 6 grid (each cell subtending 2.42 deg × 2.42 deg). The location of each circle was randomly determined with the constraint that the circles never overlap. However, in the grouped condition, all circles within each subset presented on the screen were aligned vertically (figure 1b). Each circle was presented with a slight spatial jitter within a range of 0.38 deg from the center in a cell. The probes for the mean discrimination task were 0.38 deg × 0.76 deg filled rectangles. Each probe had a different color that was chosen from one of the colors of sets in the original display. The probes were presented at the center of each side of display.

2.1.3 *Design.* Experiment 1 had a 3 × 4 design with three different display types (no-ISI, 1 s ISI, and grouped conditions) and four different numbers of sets (from two to five). The type of display was a between-subjects variable. In the no-ISI condition there was no delay between the stimulus and probe arrays, but in the 1 s ISI condition there was a delay of 1 s after the stimulus array. In the grouped condition the ISI was 0, and all subsets were grouped by proximity. The number of sets was a within-subjects variable. The order of trials within each block was randomly selected under the constraint that all conditions were presented once before they were repeated. Each participant performed 20 trials (4 set sizes × 5 repetitions) in the practice session and 240 trials (4 set sizes × 60 repetitions) in the experimental session.

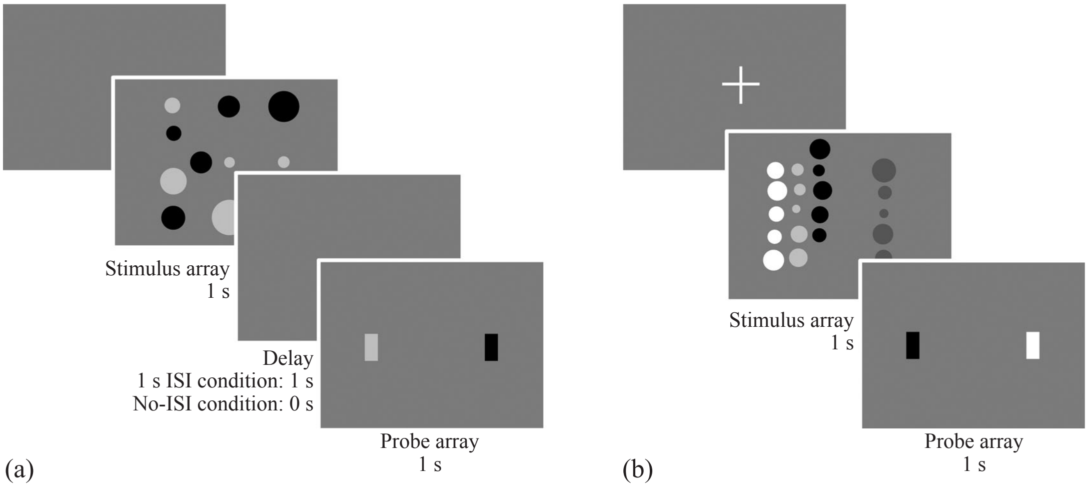

Figure 1. Stimuli and procedure of experiment 1: multiple sets (up to five) of five circles were presented for 1 s. The task was to judge which probed set had the larger mean size. (a) depicts the nongrouped condition, and (b) the grouped condition.

2.1.4 *Procedure.* Figure 1 shows a sample trial in experiment 1. Each trial began with a fixation cross. A stimulus array was then presented for 1 s, followed by a gray screen with two color probes and a fixation cross at the center of the screen. In the 1 s ISI condition a gray screen was presented for 1 s before the probe screen. The stimulus array was presented for 1 s in order to allow enough time for the participants to perceive every circle presented on the screen (Eng, Chen, & Jiang, 2005). The probe array remained visible until the participants responded. Because the probes appeared after the stimulus presentation, participants had no prior knowledge as to which of the sets they would have to report on. Therefore, in order to respond correctly, the participants were required to retain memory of all of the sets that had been previously presented. When they thought that the left probe indicated the set with the larger mean size, they pressed '1'; otherwise, they pressed '2'. Feedback was provided for incorrect responses.

2.2 *Results and discussion*

The results of experiment 1 are shown in figure 2. A repeated-measures ANOVA revealed the significant main effect by the number of sets ($F_{3,123} = 63.34, p < 0.01$) on participants' accuracy in the mean size judgment task. Specifically, as the number of sets of circles presented on the visual array increased, participants' accuracy decreased. Despite the decreasing accuracy with the number of sets, the overall accuracy was significantly higher than chance for all set sizes (all *p*s < 0.01). The main effect by the display type (no-ISI, 1 s ISI, and group) was also significant ($F_{2,41} = 24.50$, $p < 0.01$), but the interaction between the number of sets and the display type was not significant ($F_{6,123} = 0.76$, $p = 0.60$).

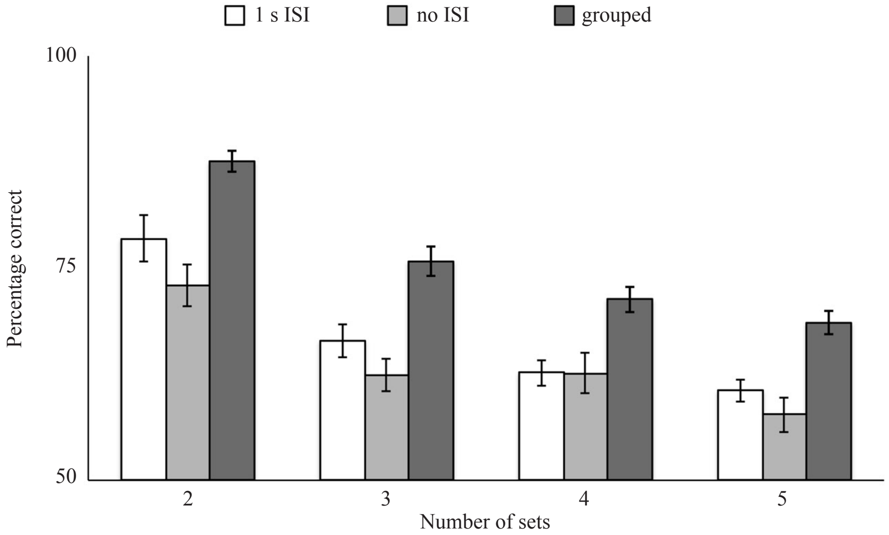

Figure 2. Results of experiment 1: accuracies of mean size judgments are shown depending on the conditions. The error bars indicate the standard errors.

We first asked whether the memory delay period degraded participants' performance in the mean size judgment task. The a posteriori contrast analysis showed that participants' accuracy for the no-ISI and 1 s ISI conditions did not significantly differ ($F_{1,27} = 2.68$, $p = 0.12$), suggesting that representation of mean sizes was not impaired by the memory delay. This is consistent with the finding of a previous study (Chong & Treisman, 2003). We then examined whether group manipulation improved participants' performance on the mean size judgment task. We found that participants' accuracy in the grouped condition was significantly higher than that in the other two conditions, in which subsets of circles were not spatially grouped (both *p*s < 0.01). This result suggests that spatially grouping subsets helps observers to represent the mean size of multiple groups of objects. This is consistent with previous findings on the effect of grouping on visual memory (Woodman et al., 2003; Xu, 2006). When parsing and segmenting groups of elements of a visual array becomes easy, observers' ability to extract and remember ensemble statistics of the groups from the array becomes more accurate and efficient.

Next, we analyzed how many mean sizes were stored depending on grouping. We assumed for an ideal observer that the observer would choose the correct answer only when the observer either represented both of the test sets or made a lucky guess. If we further assume that the observer's capacity limit is two sets, therefore, the expected accuracy levels are computed as follows: accuracy = $1P + 0.5(1 - P)$, where P is the probability that the two represented sets are tested from N sets displayed [ie $1/(N$ choose 2)]. At the capacity limit of two sets, if two random sets are sampled on every trial from displays with two, three, four, and five sets, the predicted accuracy levels are 100.00, 66.68, 58.33, and 55.00%, respectively. Furthermore, if we assume that the observer's capacity limit is three sets, the predicted accuracy levels are 100.00, 100.00, 62.50, 55.50%, and so forth. We can then fit the lower bound on the number of sets which would be needed for each participant to attain their observed accuracy levels at the set sizes of four and five. This analysis provided 2.5 sets on average when circles were spatially intermixed in a visual array, and 3.5 sets on average when circles were spatially grouped. Therefore, it appears that the participants correctly remembered at least 2.5 sets without grouping and 3.5 sets with grouping. Grouping enhanced participants' memory by at least one set.

Note that the capacity limits for individual participants were predicted based on the lower bound because the accuracy prediction assumes an ideal observer with no internal noise. For an ideal observer, when capacity limit is larger than the number of sets presented on a visual array, accuracy is predicted to be 100%. However, human observers' performance never approaches 100% even when a visual array contains only two sets, unlike the prediction. This is because human observers' performance is susceptible to internal noise that is inherently embedded in the system for ensemble representation (Ariely, 2008; Im & Halberda, 2013). Therefore, in order to accurately fit human observers' performance, one should measure the level of internal noise of an individual observer. When estimates of internal noise for individual participants are not available, the lower bound may instead provide a quick and easy way to predict *at least* how many sets would need to be represented in order for an individual participant to attain a comparable (or even higher) level of accuracy (Haberman & Whitney, 2010; but see Im & Halberda, 2013). Therefore, this is a conservative criterion for our purpose.

Moreover, it is also worth noting that these estimates of the capacity limits are based on the notion of a high-threshold model by assuming two discrete states: (1) answering correctly when the two probed sets were represented, or (2) otherwise, guessing blindly. It remains controversial, however, whether internal representation is fixed or continuously degraded. Future research should empirically investigate the nature of ensemble representation. Whether discrete or continuous, our main focus in the current study still holds: spatial grouping enhanced participants' performance on the mean size comparison task by facilitating parsing of groups of items.[1]

[1] Although both models assuming fixed and graded representations will predict enhancement of participants' performance by spatial grouping, explanations for the enhancement can be different between the two models. In the fixed representation model, for example, participants' improved performance by spatial grouping is due to the increased capacity limit of the number of sets to be remembered (see our results); on the other hand, the graded representation model predicts that the increased fidelity of representations of the grouped sets is the reason for the enhancement.

3 Experiment 2

In experiment 1 we found that spatially grouping elements helped participants to improve their visual memory. When elements of each subset were spatially grouped by proximity, participants' accuracy on the mean size judgment task was improved by one set. We hypothesized that spatial grouping facilitated parsing of groups of elements in a visual array, allowing observers to extract and encode the groups more accurately and efficiently. In experiment 2 we tested this by presenting participants with visual arrays in which only two subsets (out of two or five sets) were spatially aligned. First of all, if spatial grouping enhances the efficiency of mean size extraction, we expect to see increased accuracy when groups of items are spatially aligned in a visual array even when the number of groups the visual array contains does not exceed the capacity limit of VWM (eg two groups of circles). Second, when there are more groups of items than the capacity limit (eg five groups of circles), we expect that groups of items that are spatially aligned will be more easily segmented and combined for mean size extraction, and thus will be prioritized for selection and encoding, over others that are randomly intermixed. To foreshadow our main results for experiment 2, we found that participants' accuracy was higher for visual arrays in which two groups were spatially aligned than for those in which two groups were intermixed randomly, suggesting that efficient parsing of groups enhances the fidelity of mean size representation. Moreover, we found that the grouping cue enhanced the accuracy of the mean comparison task only when two spatially grouped sets were the target sets among five sets. When spatially grouped sets were nontargets, however, the grouping cue significantly impaired the accuracy of the mean size comparison of the two other sets that were not spatially aligned, suggesting that grouped sets were prioritized for selection and encoding processes for limited VWM.

3.1 *Method*

3.1.1 *Participants.* Fifteen new undergraduate students from Yonsei University participated in the experiment for course credit. All participants had normal or corrected-to-normal vision and were unaware of the purpose of the study.

3.1.2 *Apparatus and stimuli.* The stimuli and the apparatus were similar to those for experiment 1, except that we varied the spatial configuration of circles for the grouped conditions by spatially aligning only two groups of circles (out of two or five groups) in visual arrays.

3.1.3 *Design.* The current study had two independent variables: the number of sets presented in the visual array (either two or five sets), and the type of spatial configuration: TG (target grouping), NG (nontarget grouping), and intermixed. In the TG condition two sets were spatially grouped by proximity, and the grouped sets were always selected as targets for the mean size comparison task and probed by a postcue. However, in the NG condition the probes were always selected from the nongrouped sets for the mean size comparison task. Note that the NG condition at a set size of five allowed us to ensure that participants could not expect any causal relation between the grouping cue and the to-be-probed sets. Because two distractor subsets were grouped in this condition, participants could not use the strategy of always selecting only the two grouped sets and focusing on them to perform the mean size comparison task. In the intermixed condition there was no spatial grouping by proximity and all the groups of circles were spatially intermixed. Because the NG condition was not applicable to the set size of two, there were two different grouping conditions for this set size (2 TG and 2 intermixed) and three different grouping conditions for the set size of five (5-TG, 5-NG, and 5-intermixed). The five different types of grouping are graphically demonstrated in figure 3a. The order of trials within each setting was randomly selected under the constraint that all conditions were presented once before they were repeated. There were 30 trials (5 conditions × 6 repetitions) in the practice session and 240 trials (5 conditions × 48 repetitions) in the experimental session. Feedback was provided for incorrect responses.

3.1.4 *Procedure.* The procedure was identical to the no-ISI condition in experiment 1.

3.2 *Results and discussion*

The results are shown in figure 3b. The overall accuracy of the mean size comparison was significantly higher than chance (50%) in all of the conditions (all *p*s < 0.05). As in experiment 1, we first found facilitation by spatial grouping. Participants' accuracy in the TG condition, in which two grouped subsets were probed as targets for the mean size judgment task, was significantly higher than that in the intermixed condition, in which none of the subsets was grouped. This was true for set sizes of two and five (both *p*s < 0.01). Spatially grouping elements of subsets in the visual array enhanced accuracy on the mean size judgment task of the sets. Moreover, the accuracy of the mean size comparison in the target grouping condition with a set size of five did not differ from that in the intermixed condition with a set size of two ($t_{14} = 1.87$, $p = 0.08$).

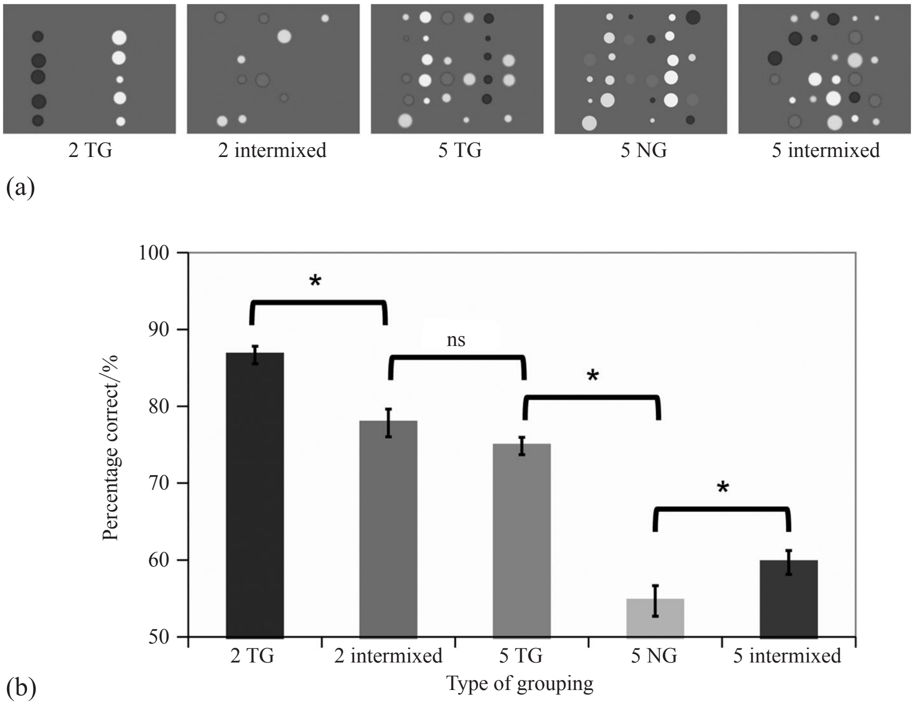

Figure 3. Experiment 2: (a) the five conditions; (b) the results. Accuracies of mean size judgments are shown depending on the conditions. Notes: * indicates a significant difference; the error bars indicate the standard errors; TG = target grouping; NG = nontarget grouping.

However, the spatial grouping of subsets helped only when the grouped subsets were targets for the mean size comparison task. When the grouped subsets were non-targets for the mean size judgment task, participants' accuracy in the mean size judgment task was impaired by grouping. Accuracy in the NG condition with a set size of five was significantly worse than that in both the TG and intermixed conditions (both *p*s < 0.05). Spatially grouping subsets in the visual array improved observers' accuracy in the mean size comparison task only when the grouped subsets were targets for the mean size comparison task. When the grouped subsets were among nontargets for the mean size judgment task, however, spatial grouping impaired observers' performance. Our findings suggest that spatial grouping facilitates parsing and selection of grouped subsets to be prioritized over nongrouped subsets for memory representation in VWM. These findings together provide evidence that the visual system can select and retain ensembles of multiple elements as higher order units for memory, just as it can individual objects.

One might argue that participants could have used the explicit strategy of always trying to extract only two grouped sets, whenever they are presented. If that were the case, however, participants' accuracy for the TG condition with five sets (5 TG) should be comparable with that for the TG condition with two sets (2 TG), rather than with that for the intermixed condition with two sets (2 intermixed). In addition, the accuracy for the nongrouped condition with five sets (5 NG) should be as low as the chance level. Although we did observe impairment, participants' accuracy for the 5 NG condition was significantly higher than the level of chance.

4 General discussion

The goal of this study was to examine how two different higher order representations built from multiple individual objects, ensemble statistics, and perceptual grouping play a joint role in maintaining information in VWM. Specifically, we investigated additional facilitation when both representations were available for VWM. We first showed that simply grouping elements by spatial proximity enhanced participants' performance on the mean size comparison task. The estimated number of sets remembered by participants was approximately 2.5 on average when the individual objects were spatially intermixed; however, when the objects were spatially grouped (experiment 1), participants seemed to remember at least one more set.

In previous studies grouping items improved the VWM capacity only when certain constraints were satisfied. Woodman et al. (2003) found that grouping increased the probability of storing grouped features. Yet they did not investigate whether grouping increased the overall VWM capacity. Xu (2006) showed that grouping increased the capacity of visual memory only when both proximity and connectedness were used for grouping items. In addition, Xu and Chun (2007) showed that grouping increased the capacity only when the set size was close to the capacity limit of visual memory. Wheeler and Treisman (2002) showed that within-feature grouping did not improve the capacity of visual memory as much as between-feature grouping. Likewise, only repeated target location was found to improve the performance of working memory (Olson & Jiang, 2005). Because visual environments contain many regularities already (Kersten, 1987), it may be difficult to enhance the capacity of visual memory by providing additional structure through grouping. In our study, however, by rendering group structure in a visual array salient enough (ie by using both proximity and color as cues for grouping), we show that efficient grouping increased the capacity of visual memory by one set in a visual array, which provides compressed memory representation of about five more elements of the array.

Increasing the VWM capacity limit by one set is not trivial (Chen, Eng, & Jiang 2006; Olson & Jiang, 2005). For example, Chen et al. (2006) investigated whether training changes the capacity of visual memory. Using unfamiliar random polygons for stimuli in a change-detection task, they found that the VWM capacity was the same for both familiar and unfamiliar shapes, although training improved the familiarity of trained shapes. Our study, however, by introducing statistical structure and emphasizing to-be-remembered sets by grouping, was able to show a one-set increase in VWM capacity. Representing and maintaining one more set of multiple objects could increase the overall probability that all of the elements (eg the sizes of the five circles in the current experiments) of the set will be encoded and retrieved to some extent for further processing.

We found that our participants could maintain up to at least 2.5–3.5 sets of multiple items (which in fact includes 13–18 individual elements). A previous study on enumeration of multiple groups of elements (Halberda, Sires, & Feigenson, 2006) found a similar three-set limit, but including two subsets and one superset (total number of dots). The number of ensembles that can be maintained at once seems to be constrained by the comparable range of capacity limits that also apply to individual objects. For example, when memory for simple, highly discriminable colored squares is tested, typical adult observers have a capacity of only three to four objects (Vogel & Awh, 2008). When object complexity increases, however, only up to two to three objects can be remembered (Alvarez & Cavanagh, 2004), but not more than that. The three-item limits have also been observed in studies of object-based attention (Pylyshyn & Storm, 1988; Scholl, 2001) as well as in studies of VWM (Alvarez & Cavanagh, 2004; Luck & Vogel, 1997, 2013; Vogel & Awh, 2008). Together, ensembles of multiple items can function as higher order units for VWM, just as individual objects can. Sets of multiple items can be selected, encoded, and retrieved as discrete units for VWM, constrained by the same capacity limit of VWM. This aspect allows for hierarchical representation in which the visual system can compress much more information than it can hold at any given time and flexibly shift between sets and elements for retrieval, just as we make chunks of items to encode and retrieve together (eg remembering a phone number with three separate chunks). We suggest that such representational flexibility can allow us to bypass the strict constraints imposed by VWM.

Spatially grouped sets seem to be prioritized for selection and encoding for memory representation in VWM. It is important that the benefit from grouping we found in the current study is based on ensembles of multiple items rather than on individual objects. For individual objects the object benefit has been previously observed: two features from the same object are remembered much better than the same two features located on two separate objects (eg Luck & Vogel, 1997). The object benefit becomes stronger when two parts of an object are grouped more strongly (Xu, 2006). In the current study we show that selection and memory of sets of multiple items become much better when elements are spatially grouped together. The grouped sets were better extracted and better remembered, whereas other ungrouped sets were not.

Where does the ensemble-based benefit from spatial grouping occur in VWM? Is it at the encoding, maintenance, or retrieval stage? The previous studies on object benefit in memory tasks and visual search tasks showed an object benefit in which detection and memory are improved when two parts are grouped into a single conjoined object compared with when they are separate parts (Goldsmith, 1998; Xu, 2006). It has been suggested that such an object benefit is more likely due to facilitation in the encoding process. The current study adds to these previous findings of the object benefit, by further showing that efficient grouping facilitates encoding of ensembles of discrete objects, such that ensembles can function as higher order units for selection and memory. In addition, a previous study on mean size discrimination showed that participants' accuracy and latency in discriminating mean sizes of two sets of elements were systematically improved as spatial separation between the two sets increased, suggesting that separation between sets led to efficient parsing and encoding of ensembles (Ly, Im, & Halberda, 2009). This mean size discrimination task shares only an encoding process, but not maintenance and retrieval processes, with our VWM task. The presence of benefit by spatial grouping therefore seems to occur at an encoding stage to facilitate participants' performance for both the mean size discrimination task and our VWM task. Together, we suggest that the ensemble-based benefit by grouping we observed in the current study may also occur in the encoding stage, just as the object benefit in VWM does. More research, however, is needed to verify this hypothesis.

Ensemble representation has mostly been explored in the context of global, nonselective visual processing. Many previous studies emphasized only that ensemble representation is an efficient global structure that can deal with multiple objects, as if it is always available for free. However, visual environments often contain multiple sets of items that are spatially intermixed. For ensemble representation to be an efficient structure to summarize visual environments, the visual system should be able to segment sets of items and represent multiple ensemble features separately. Here, we show that one can extract multiple ensemble representations from visual arrays containing spatially intermixed items, but this ability is also constrained by the limited capacity of the visual system. Importantly, we also show that ensemble representation requires parsing and segmentation of groups as higher order units for selection and memory. Parsing and segmenting meaningful groups helps the visual system to understand the visual environment as a meaningful and coherent structure. When parsing and segmenting items into groups of elements is made efficient and easier by the aid of grouping, ensemble representations of the groups are more likely to be selected, extracted, and better maintained.

The visual environment is highly structured. VWM mirrors this structure through the use of structured representations (Brady et al., 2011). We found that ensemble representation and set representation by grouping, two such structures, help each other to achieve efficient representation in visual memory. Specifically, set representation by grouping biases the selection and extraction of sets into ensemble representation for further storing and maintaining in VWM.

Acknowledgment. This work was supported by a National Research Foundation of Korea grant funded by the Korea government (MEST, No. 2011-0025005).

References

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

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, **19**, 392–398.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the USA*, **106**, 7345–7350.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, **12**, 157–162.

Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & Psychophysics*, **70**, 1325–1326.

Beck, J. (Ed.). (1982). *Organization and representation in perception*. Hillsdale, NJ: Erlbaum.

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.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, **11**(5):4, 1–4.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433–436.

Chase, W. G., & Ericsson, K. A., (1981). Skilled memory. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 141–189). Erlbaum, Hillsdale, NJ.

Chen, D., Eng, H. Y., & Jiang, Y. (2006). Visual working memory for trained and novel polygons. *Visual Cognition*, **14**, 37–54.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, **43**, 393–404.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, **24**, 85–87.

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.

Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences of the USA*, **108**, 19552–19557.

Eng, H. Y., Chen, D., & Jiang, Y. (2005). Visual working memory for simple and complex visual stimuli. *Psychonomic Bulletin & Review*, **12**, 1127–1133.

Feigenson, L. (2008). Parallel non-verbal enumeration is constrained by a set-based limit. *Cognition*, **107**, 1–18.

Feigenson, L. (2011). Objects, sets, and ensembles. In S. Dehaene, & E. Brannon (Eds.), *Attention and performance* (Vol. XIV, pp. 13–22). Oxford: Oxford University Press.

Feigenson, L., & Halberda, J. (2008). Conceptual knowledge increases infants' memory capacity. *Proceedings of the National Academy of Sciences of the USA*, **105**, 9926–9930.

Frith, C., & Frith, U. (1972). The solitaire illusion: An illusion of numerosity. *Perception & Psychophysics*, **11**, 409–410.

Ginsburg, N. (1976). Effect of item arrangement on perceived numerosity: Randomness vs regularity. *Perceptual & Motor Skills*, **43**, 663–668.

Ginsburg, N., & Goldstein, S. R. (1987). Measurement of visual cluster. *The American Journal of Psychology*, **100**, 193–203.

Goldsmith, M. (1998). What's in a location? Comparing object-based and space-based models of feature integration in visual search. *Journal of Experimental Psychology: General*, **127**, 189–219.

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

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, **17**, 572–576.

Im, H. Y., & Chong, S. C. (2009). Computation of mean size is based on perceived size. *Attention, Perception, & Psychophysics*, **71**, 375–384.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception, & Psychophysics*, **75**, 278–286.

Kersten, D. (1987) Predictability and redundancy of natural images. *Journal of the Optical Society of America A*, **4**, 2395–2400.

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

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, **390**(6657), 279–281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, **17**, 391–400.

Ly, R., Im, H. Y., & Halberda, J. (2009). Spatial overlap of collections affects the resolution of ensemble features. *Journal of Vision*, **9**(8), 921 (Abstract).

Olson, I. R., & Jiang, Y. (2005). Associative learning improves visual working memory performance. *Journal of Experimental Psychology: Human Perception and Performance*, **31**, 889–900.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, **44**, 369–378.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437–442.

Poltoratski, S., & Xu, Y. (2013). The association of color memory and the enumeration of multiple spatially overlapping sets. *Journal of Vision*, **13**(8):6, 1–11.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, **3**, 179–197.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, **80**, 1–46.

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

Taves, E. H. (1941). Two mechanisms for the perception of visual numerousness. *Archives of Psychology*, **37**, 1–47.

Victor, J. D., & Conte, M. M. (2004). Visual working memory for image statistics. *Vision Research*, **44**, 541–556.

Vogel, E. K., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, **17**, 171–176.

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.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, **131**, 48–64.

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

Wolfe, J. M., Vo, M. L., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Science*, **15**, 77–84.

Xu, Y. (2006). Encoding objects in visual short-term memory: The roles of feature proximity and connectedness. *Perception & Psychophysics*, **68**, 815–828.

Xu, Y., & Chun, M. M. (2007). Visual grouping in human parietal cortex. *Proceedings of the National Academy of Sciences of the USA*, **104**, 18766–18771.

Zosh, J. M., & Feigenson, L., (2012). Memory load affects object individuation in 18-month old infants. *Journal of Experimental Child Psychology*, **113**, 322–336.