Journal of Experimental Psychology:
Human Perception and Performance
2015, Vol. 41, No. 5, 1247–1259


0096-1523/15/$12.00 http://dx.doi.org/10.1037/xhp0000086

# How to Break the Configuration of Moving Objects? Geometric Invariance in Visual Working Memory

Zhongqiang Sun, Yuan Huang, Wenjun Yu, Meng Zhang, and Rende Shui
Zhejiang University

Tao Gao
Massachusetts Institute of Technology

Visual working memory is highly sensitive to global configurations in addition to the features of each object. When objects move, their configuration varies correspondingly. In this study, we explored the geometric rules governing the maintenance of a dynamic configuration in visual working memory. Our investigation is guided by Klein's Erlangen program, a hierarchy of geometric stability that includes affine, projective, and topological invariants. In a change-detection task, memory displays were categorized by which geometric invariance was violated by the objects' motions. The results showed that (a) there was no decrement in memory performance until the projective invariance was violated, (b) more dramatic changes (such as a topological change) did not further enlarge the decrement, and (c) objects causing the violation of projective invariance were better encoded into memory. These results collectively demonstrate that projective invariance is the only geometric property determining the maintenance of a dynamic configuration in visual working memory.

*Keywords:* visual working memory, configuration, geometric invariance, motion perception



Humans process dynamic information from multiple objects every day, ranging from chasing prey in ancient times to playing football with teammates in modern times. To evaluate the relationships and dynamic spatial configurations (DSCs) among these objects, the visual system has to store the history of the moving objects for a short period. Presumably, these online storage and evaluation processes are achieved via visual working memory (VWM), which offers a limited capacity for storing and manipulating visual stimuli (Baddeley, 1998; Luck & Vogel, 1997). Compared with the large amount of work on the storage of static features and objects (e.g., Alvarez & Cavanagh, 2004; Shen, Huang, & Gao, 2015; Vogel, Woodman, & Luck, 2006), few studies have explored the memory of motion information. These studies primarily focused on the storage of individual motion directions, such as how the precision of memorized motion direction declines as a function of the set size of the memory display (e.g., Narasimhan, Tripathy, & Barrett, 2009; Shooner, Tripathy, Bedell, & Öğmen, 2010; Zokaei, Gorgoraptis, Bahrami, Bays, & Husain, 2011), how the speed of the motion impacts memory performance (McKeefry, Burton, & Vakrou, 2007), and to what extent the memorized motion direction can be retained over a long period (Blake, Cepeda, & Hiris, 1997). While these previous studies provided a comprehensive picture of how individual items are stored, it is still far from clear how the global configuration of these items is constructed and maintained in VWM.

## Contrasting Static and Dynamic Spatial Configurations

In the current study, we focused on the encoding and maintenance of multiple moving objects' DSCs instead of the storage of each individual object. It has been shown that VWM stores the relationships between individual items on the basis of global spatial configurations (Gmeindl, Nelson, Wiggin, & Reuter-Lorenz, 2011; Jiang, Chun, & Olson, 2004; Jiang, Olson, & Chun, 2000; Olson & Marshuetz, 2005; Woodman, Vecera, & Luck, 2003). The spatial configuration of a static scene is analogous to a spatial reference frame (Olson & Marshuetz, 2005). The effects of this spatial configuration were explored with a change-detection task, in which an array of objects was presented first as the memory display and shortly afterward was presented with the same or a different spatial configuration. Even when the configuration was completely irrelevant to the memory task (e.g., remembering the color or shape; see Jiang et al., 2000), performance was worse when the spatial configuration was distorted. This effect existed even when people were explicitly instructed to ignore the

This article was published Online First June 15, 2015.

Zhongqiang Sun, Yuan Huang, Wenjun Yu, Meng Zhang, and Rende Shui, Department of Psychology and Behavioral Sciences, Zhejiang University; Tao Gao, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

This research was supported by the Center for Brains, Minds and Machines (CBMM), funded by National Science Foundation, the Science and Technology Centers award CCF-1231216, the National Natural Science Foundation of China (31170975, 61431015, 31170975 and 31200786), and the Fundamental Research Funds for the Central Universities.

Correspondence concerning this article should be addressed to Tao Gao, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Institute of Technology, Cambridge, MA 02139, or to Rende Shui, Department of Psychology and Behavioral Sciences, Xixi Campus, Zhejiang University, 148 Tianmushan Road, Hangzhou, China 310028. E-mail: taogao@mit.edu or rshui@zju.edu.cn

spatial locations, which suggests that the storage of the global spatial configuration is obligatory (Jiang et al., 2000, 2004).

It is rather unclear how to transfer the above conclusion from static displays to dynamic ones. In the case of static displays, the differences between the memory and test displays were introduced by an abrupt change without any intermediate transition. However, in dynamic displays, such an "abrupt change" does not exist, as the configuration changes smoothly along with the trajectory of each individual object. In other words, it is not clear what an "unchanged" configuration means in a dynamic display. Directly applying the conclusion from static scenes to dynamic displays will lead to a naive conclusion that DSCs cannot be maintained in VWM at all, as the spatial location of each object keeps changing.

To encode DSCs in memory, a more sophisticated representation of configuration is required. To the best of our knowledge, only one study has explored the storage of DSCs in VWM (Papenmeier, Huff, & Schwan, 2012). In that study, participants were asked to memorize a scene containing several moving objects. In the test display, an item's motion was "rewound" and replayed. The results showed that change-detection performance was the best when the test item's motion was played with the same context, in which the other objects' motions were copied exactly from the memory display compared with in isolation or with a different context. These results demonstrate that individual motion directions are indeed encoded as a part of a configuration. However, it is unclear what types of visual features are critical in maintaining a dynamic configuration. By "rewinding" the memory display and "playing it back" again, the results of Papenmeier et al.'s (2012) study can be directly compared with those obtained with static features (e.g., Jiang et al., 2000). From this perspective, motion can be treated as a feature that is similar to other static scene attributes, such as color and shape. Nevertheless, motion is different from static features, as it is not only spatially distributed on different objects but also captures how the scene is going to develop over time. A robust dynamic configuration should persistently exist when the objects' movements continue smoothly. Therefore, in this study, we were more interested in exploring the representation of dynamic configuration by playing the motion display forward rather than rewinding it back to the original spatial location in the test displays.

Another related line of research is multiple object tracking (MOT), which focuses on the online perception of moving objects (e.g., Blaxton, Fehd, & Seiffert, 2011; Franconeri, Jonathan, & Scimeca, 2010; Pylyshyn, 2001; Pylyshyn & Storm, 1988; Zhao et al., 2014). A key study on this topic, conducted by Yantis (1992), demonstrated that the configuration of multiple moving objects could be defined as the convex hull over the location of each attentively tracked object. This polygon collapsed when a vertex of the polygon crossed an opposite edge, which caused the relative "ordering" of the points on the perimeter of the polygon to change. While there has been significant progress in understanding the mechanisms of MOT over the past 20 years, it is striking that the geometric rules governing the dynamic configuration of multiple objects were rarely explored further. One motivation of the current work was to revive the geometric perspective pioneered by Yantis (1992) in the context of VWM.

## Geometric Transformations and Dynamic Spatial Configuration

Here, we use geometry as a powerful tool for exploring the nature of DSC. To represent dynamic spatial configuration, the visual system has to balance *dynamic* and *configuration*, which seem to capture opposite properties of a scene. Dynamic implies that the objects' spatial positions are constantly changing, while configuration implies that certain global properties remain invariant over time. The representation of DSC should be a compromise of two opposite demands. On one hand, it needs to be insensitive to small changes such that a DSC persists over time, which is critical to stable and coherent visual experience. On the other hand, it should not tolerate major and dramatic changes such that the destruction of a DSC can reflect the disappearance of important configuration information in the scene. However, the above intuition of DSC alone is far from sufficient, as it is completely unclear what "small changes" and "major and dramatic changes" mean. To address this challenge, we proposed a formal theory based on geometric transformation to explain the construction and deconstruction of DSC.

Such a geometric perspective of DSC has several advantages. First, configuration can be vaguely interpreted as information that does not reside in individual objects. By using geometry, one can avoid intuitively defined configuration and analyze the spatial relationships among individual objects with quantitative precision. Second, a theory of geometry can divide configurations into a hierarchy with several levels of spatial relationships. Such a hierarchy is important both theoretically and methodologically. From a theoretical perspective, each level of the hierarchy can be a candidate representation for DSC. From a methodological perspective, there are explicit algorithms defining the maintenance and destruction of each level of the spatial configuration. These algorithms can be directly employed for designing psychophysical experiments.

More generally, geometry is important for visual perception due to the deep connections between vision and graphics. Computer graphics start from creating a 3D scene and then projecting it onto a 2D image (i.e., the rendering process) by using geometric rules. In contrast, vision starts from 2D images on the retinas and then reconstructs 3D scenes given 2D images. In other words, vision can be understood as a type of "inverse graphics." This perspective has a long tradition in psychology, since von Helmholtz (1867), and has been supported by many empirical studies (e.g., Feldman et al., 2013; Scholl, 2005). Recently, this perspective has also been successfully implemented in computer vision for image parsing and recognition (Kulkarni, Yildirim, Kohli, Freiwald, & Tenenbaum, 2014). Employing geometric rules to understand DSC fits well with this "inverse graphics" perspective of vision.

Throughout the history of human vision research, geometry has mostly been used in studies of shape constancy and object recognition. The studies of shape constancy concern the perception of the physical shape regardless of the transformation of the retinal image. This line of research explores how humans maintain a constant shape representation during the change of viewpoint of the observer or the movement of objects. Some researchers found that the shape representation is based on geometric invariants at different levels of structural stability. These geometric invariances include Euclidean properties and affine properties such as copla-

narity (e.g., Tittle, Todd, Perotti, & Norman, 1995; Todd & Bressan, 1990). In addition, in studies of object recognition, one seminal study showed the importance of topological structure in visual perception (Chen, 2005). This topology theory proposes that the global nature of perceptual organization can be described in terms of topological invariants, which include properties such as connectedness, continuity, and boundary that are preserved under continuous deformations, including stretching and bending. More importantly, Chen (2005) systematically introduced Klein's Erlangen program[1] into the study of visual perception. He found that the relative perceptual salience of different geometric properties is remarkably consistent with the hierarchy of geometries according to Klein's Erlangen program, which stratifies geometries in terms of their relative stability over transformations (Chen, 2001, 2005; Todd, Chen, & Norman, 1998).

There are three types of geometric invariance in Klein's Erlangen program, which differ in the degree of freedom in geometric transformation. The most constrained one is *affine* transformation, allowing translation, rotation, reflection, and stretch, as well as any combination of these. It distorts angles, distances, and areas while retaining the other properties (e.g., parallel lines remain parallel). In this case, a square can be affinely transformed to any kind of rectangle or even parallelogram but not to a trapezium. *Projective* transformation allows a much larger class of transformations than affine transformation does. It introduces further distortions (e.g., without parallelism) but preserves more basic properties such as the collinearity of points (e.g., points remain points and lines remain lines) and incidence (that is, whether a point lies on a line). More importantly, under projective transformation, a convex polygon can be projectively transformed to any other convex polygon but cannot be transformed to a concave one and vice versa. The preservation of concavity is guaranteed by a concept called the *cross-ratio*[2] (e.g., Anderson, 2006). Finally, as the most radical transformation, *topological* transformation includes continuous deformations such as stretching and bending, while other properties such as connectedness, continuity, and boundary are preserved. A typical type of topological invariant is the number of bounded holes in the figure (transforming a 0 to an 8 violates topological transformation). Therefore, geometric stability increases from affine to projective to topological invariants, along with a reduced number of constraints. While it is relatively easy to break invariance defined by affine transformation, it requires rather dramatic geometric changes to break topological invariance.

## Current Study

Based on studies of subjective contours (e.g., Grossberg, 2014; Grossberg & Mingolla, 1985; Kanizsa, 1974), we assume that observers can connect multiple moving objects by imaginary lines. These lines define a virtual polygon to which different types of geometric transformations can apply. In the context of VWM, we further assume that manipulating geometric properties should be able to impact memory performance by mediating the stability of the DSC.

To illustrate the scope of possible results by using this paradigm, we indulged in a few extreme hypotheses herein as follows:

> *Extreme Hypothesis 1:* VWM would encode each individual motion direction in isolation, and none of the transformation types would impact memory performance.

> *Extreme Hypothesis 2:* Configuration could exist but in a very fragile state that could be disrupted by even affine transformation, producing a significant drop in memory performance.

> *Extreme Hypothesis 3:* The configuration would be highly robust such that only a topology change (the most dramatic change) could break it.

Based on the aforementioned studies, we expected that the manipulation of geometric transformation could impact memory performance and were interested in how it does so.

Breaking geometric invariances can impact VWM performance in two qualitatively different ways. The first is the continuous decay hypothesis, which asserts that all types of geometric invariances engage in the maintenance of DSC. The stability of DSC should drop monotonically by gradually breaking different levels of geometric invariances from affine to topology. The other is the abrupt collapse hypothesis, which asserts that DSC is only defined on a given level of geometric transformation. Any variances below this level can be tolerated without imposing any decrement in memory performance, while any variance above this level cannot impose any further decrement.

To address these alternative hypotheses, we manipulated DSCs with different geometrical invariances in a series of experiments given Klein's Erlangen program. We focused on how memory performance changed as a function of breaking different levels of geometric invariances. In addition to overall performance, we were also interested in whether all the vertices of the DSCs were treated equally given a certain geometric transformation, as some vertices may play more important roles in a transformation than others.

Experiments 1 and 2 investigated whether any of the aforementioned hypotheses could successfully explain memory performance as a function of geometric transformations. Experiment 3 further explored the asymmetry of processing different moving objects, which are identical except for their roles in a given geometric transformation. Finally, Experiment 4 demonstrated that all the previous effects could be strengthened by linking each object to form a real polygon.

## General Method

We modified the change-detection task from VWM studies (Jiang et al., 2000; Luck & Vogel, 1997; Phillips, 1974) to explore the storage of multiple objects' motion directions. A memory display was presented first, in which four dots moved in different directions for 500 ms, followed by a 1,000-ms retaining display, in which the four dots stopped moving and remained static. Then all the dots started to move again from the location where they stopped before for another 500 ms as the test display. In half of the

[1] Klein's Erlangen program was an influential research program published in 1872 by Felix Klein. This manifesto classified and characterized geometries on the basis of projective geometry and group theory. Projective geometry was emphasized as the unifying frame for all other geometries considered by him. In particular, Euclidean geometry was more restrictive than affine geometry, which in turn is more restrictive than projective geometry.

[2] The cross-ratio is a number associated with a list of four collinear points, particularly points on a projective line. Given four sequential points A, B, C, and D on a line, their cross-ratio is defined as (A, B; C, D) = (AC × BD)/(BC × AD).

trials, all the dots' motion directions in the test display were identical to those in the memory display. In the other half of the trials, one randomly selected dot moved with a different motion direction in the test display. The participant was asked to store the motion directions in the memory display and to detect whether any motion direction was changed in the test display.

The configuration was defined as the virtual polygon with the four dots being its vertices. In all cases, the boundary of the virtual polygon overlapped with the convex hull of the four dots. The shape of the virtual polygon was gradually transformed to a new one by the motion of each dot in the memory display. The type of configuration change could be classified given what geometric constraints were satisfied or violated during such a transformation.

We set four main geometric transformation conditions that could occur during the memory display, starting from affine transformation[3] (see detailed videos on the website http://www.psych.zju.edu.cn/english/redir.php?catalog_id=14613).

The geometric transformation of the polygon was achieved by matrix multiplication in MATLAB 7.6 (The MathWorks, Natick, Massachusetts). The coordinates of four initial vertices ($x, y$) were changed to four new coordinates ($x', y'$) by the corresponding transformation matrix. Taking affine transformation as an example, it includes translation transformation, scale transformation, rotation transformation, and shear transformation.[4]

### Affine Condition (Satisfying Affine Transformation)

The virtual polygon during the memory display could only be changed by affine transformations, including translation, rotation, scaling, and shearing, as well as any combination of these.

### Nonaffine Condition (but Satisfying Projective Transformation)

In order to generate the movement traces, we initially programmed a new invisible polygon that was affinely transformed from the original virtual polygon. The line segment between each dot's original location and the vertex in the new polygon nearby was set as the moving trajectory for this dot. Three dots moved along their own trajectories (starting from their original locations) while the fourth dot's destination was changed to a new location 2.9° from the vertex in the new polygon (the changing direction was randomly determined). Consequently, the convex hull of the four dots after their movements could not be affinely transformed from the original polygon.

### Nonprojective Condition (but Satisfying Topological Invariance)

The memory display started with a convex polygon. During the motion, Dot A crossed a diagonal of the polygon and stopped, which changed the convex polygon into a concave one. Dot A was also labeled as the *critical* dot, while the others were labeled as *noncritical* dots. The configurations before and after the memory display were not projectively equivalent, although they had the same topological property.

### Nontopological Condition (and Not Satisfying Any Type of Geometric Invariance)

The initial spatial configuration was a convex polygon. In the memory display, Dot A crossed the diagonal of the polygon and an opposite edge sequentially, changing the convex polygon into a self-intersecting convex polygon. This dot and its adjacent Dot B in the opposite edge were labeled as critical dots because these two dots' movements caused this configuration change. The other two dots in the display were labeled as noncritical dots. All dots had equal possibilities to change their directions. The self-intersecting polygon violated all geometric invariances.

A schematic illustration of a single trial is depicted in Figure 2. The stimuli were presented on a 17-in. monitor (34.4cm * 25.8cm, 100-Hz refresh rate) with a black (RGB 0, 0, 0) background. At the beginning of each trial, a white fixation cross was presented for 500 ms. After that, four white dots (RGB 255, 255, 255; diameter of 0.5°) were randomly displayed on the boundary of an invisible circle (diameter of 5.2°) with the constraint that (a) the center-to-center distance between any two dots is within the range of 2.6 to 4.5° and (b) the invisible envelope of the four dots could form a convex polygon (quadrilateral). The orientation of each dot's motion was randomly selected from a set of 12 directions: from 0 to 330° (relative to the positive direction of the horizontal axis) in 30° steps. Each direction was different from the others.

Each trial started with a display of four static dots for 500 ms (preview display). These dots then moved in straight lines with a speed of 5.8°/s[5] for 500 ms (memory display), followed by a 1,000-ms display in which the dots remained static at the last positions of their previous motions (retaining display). Then, these dots moved again for another 500 ms as a test display and remained in the display until a response was initiated. In 50% of the trials, the moving directions of all dots in the memory display were the same as those in the test display. In the remaining 50% of trials, the moving direction of one dot changed 45° either clockwise or counterclockwise from that in the memory display.

[3] In our paradigm, each object moved separately and independently, preventing us from including Euclidean invariance, which requests identical motion directions among objects. Therefore, we started from affine transformation.

[4] $[x' \ y' \ 1] = [x \ y \ 1] \times R(\Theta) \times S(s_x,s_y) \times Sh(sh_x,sh_y) \times T(t_x,t_y)$; $T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ tx & ty & 1 \end{bmatrix}$; $t_x/t_y$ specifies the displacement along the $x$-axis/$y$-axis, ranging from $-1$ to 1; $S = \begin{bmatrix} Sx & 0 & 0 \\ 0 & Sy & 0 \\ 0 & 0 & 1 \end{bmatrix}$; $S_x/S_y$ specifies the scale factor along the $x$-axis/$y$-axis, ranging from 1.2 to 1.6; $Sh = \begin{bmatrix} 1 & Shy & 0 \\ Shx & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$; $Sh_x/Sh_y$ specifies the shear factor along the $x$-axis/$y$-axis, ranging from $-1.6$ to 1.6; $R = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$; $\theta$ specifies the angle of rotation, ranging from 0 to $\pi/3$.

[5] We used this relatively slow motion direction to reduce the perceptual difficulty of the task so as to concentrate on VWM. A previous study has shown a crowding effect of motion perception when the motion speed is 14°/s, but not when the speed is 7°/s (Alvarez & Franconeri, 2007).

Participants responded by pressing one of two buttons ("F" and "J" on the keyboard) in each trial to indicate whether any dot's moving direction had changed. They were instructed to respond as accurately as possible without worrying about the response time. Both response accuracy and reaction time (RT) were recorded. Before the actual experiment, a practice session of at least 2 min (20 trials) was completed to ensure that the participants understood the instructions.

The sample size in the current study was determined by a power analysis based on predicted effect size using G*Power 3 (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007). According to the effect size ($\eta_p^2 = .17$) obtained from the pilot experiment, the analysis suggested a sample size of 16. This sample size was adopted by all the following experiments.

Each participant provided written and informed consent before the experiment, and the procedures were in compliance with the Code of Ethics of the World Medical Association (Declaration of Helsinki), as well as approved by the Research Ethics Board of Zhejiang University.

## Experiment 1: Breaking Affine Transformation

We began by exploring whether breaking affine geometric transformation can disrupt the memory of dynamic configuration. Experiment 1 contained two different type-of-transformation conditions for the movement in the memory display: the affine transformation condition (affine) and the nonaffine transformation condition (nonaffine).

### Method

Sixteen undergraduates (seven females, 18–26 years of age) from Zhejiang University participated in this experiment. All had normal color vision and normal or corrected-to-normal visual acuity.

In the affine condition, the configuration of the four dots changed smoothly during the memory display with the constraints that (a) the virtual polygon was convex during the entire memory display and (b) the virtual polygons at any time point were affinely equivalent. In the nonaffine condition, the virtual polygons were all convex, but two virtual polygons before and after the movement during the memory display were not affinely equivalent.

Each participant performed 120 trials per condition, resulting in a total of 240 trials presented in a randomized order. The experiment was divided into four blocks with 2-min breaks between them.

### Results and Discussion

The results of Experiment 1 are shown in Figure 3. No significant difference in accuracy was found between the affine condition (80.38% ± 7.27%) and the nonaffine condition, 81.19% ± 6.78%; $t(15) = -0.76$, $p > .05$.

We also calculated the $d'$ value for each condition. The difference in sensitivity between affine (2.00 ± 0.66) and nonaffine (1.96 ± 0.63) was not significant, $t(15) = 0.36$, $p > .05$.

These results suggested that merely breaking affine invariance had no significant impact on the encoding dynamic configuration in working memory, which did not support the continuous decay hypothesis. We will further explore the abrupt collapse hypothesis in Experiment 2.

## Experiment 2: Breaking Projective and Topological Invariances

By continuously following the hierarchy of Klein's Erlanger program, we further broke either projective or topological geometric transformation in different conditions such that they could be compared directly. In addition to analyzing the overall performance, we also explored the effects of global geometric transformation on the memory of individual items. When transforming a convex polygon into a concave or a self-crossing one (see the Method section below), the role of each item is not equally important. The critical dots were defined as the ones whose motions caused a violation of certain geometric invariances.

### Method

A new group of 16 undergraduates (six females, 18–24 years of age) from Zhejiang University participated in this experiment. All had normal color vision and normal or corrected-to-normal visual acuity.

The displays can be divided into three conditions. The nonaffine condition was identical to the nonaffine condition in Experiment 1, in which affine transformation was violated, but projective and topological invariances were satisfied. The critical dot was not defined in this condition. In the nonprojective condition, the configuration of the four dots changed smoothly from a convex to a concave polygon. Two polygons before and after the memory display were not projectively equivalent. In this case, projective invariance was violated while topological invariance was satisfied, and the critical dot was the one that fell inside the convex hull instead of on the boundaries. In the nontopological condition, the configuration changed from a convex to a self-intersecting convex polygon, in which one of the dots crosses its opposite edge. In this case, topological invariance was not satisfied, and all the geometric invariances in the hierarchy were violated. Two dots that moved toward each other and finally made two opposite sides intersect were determined to be critical dots (see Figure 1). These transformations were covered in the Introduction.

Each participant performed 120 trials per condition, resulting in a total of 360 trials in which all trials were presented in a randomized order. The experiment was divided into six blocks with 2-min breaks between them.

### Results and Discussion

The averaged accuracy and $d'$ are depicted in Figures 4A and 4B. The most important result is the abrupt drop of performance when breaking the projective invariance. Confirming this observation, a one-way repeated-measures analysis of variance (ANOVA) revealed a significant main effect of the type of transformation, $F(2, 30) = 36.71$, $p < .01$, $\eta_p^2 = 0.71$. Bonferroni-corrected post hoc contrasts confirmed that the accuracy of the nonaffine condition (79.19% ± 6.48%) was higher than those of the nonprojective (70.25% ± 8.18%, $p < .01$) and nontopological (69.63% ± 5.89%, $p < .01$) conditions. Interestingly, there was no difference between the nonprojective and nontopological conditions ($p > .05$).

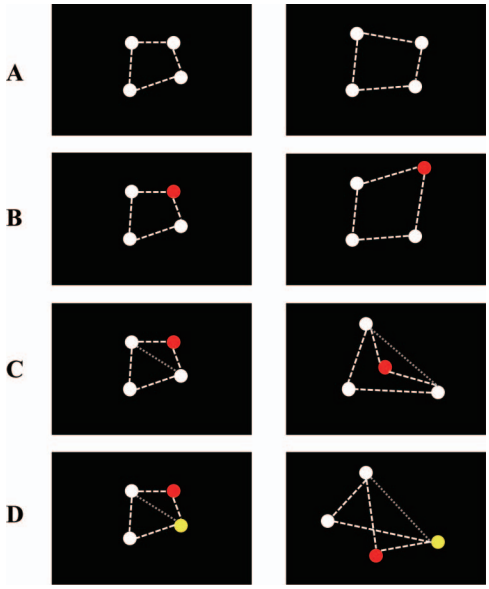


*Figure 1.* Four invariance conditions. Affine invariance condition (affine; baseline) (A). Different levels of form stability in ascending order from B to D; they differ in affine geometry, projective geometry, and topological geometry, respectively (B, C, and D). These constitute a hierarchy of geometries according to Klein's Erlangen program. All colored dots represent critical dots in Experiments 2 and 3. The thin dashed lines in C and D indicate the diagonal lines, and both the large dotted lines and thin dashed lines did not appear during the experiment. See the online article for the color version of this figure.

Analysis of the $d'$ revealed the same pattern of results. A significant main effect was found for sensitivity, $F(2, 30) = 43.88$, $p < .01$, $\eta_p^2 = 0.75$. The sensitivity was significantly higher in the nonaffine condition (1.88 ± 0.57) than in the nonprojective condition (1.22 ± 0.52, $p < .01$) or the nontopological condition (1.08 ± 0.36, $p < .01$), while no difference was found between the latter two conditions ($p > .05$). These results indicated that memory was more accurate in the nonaffine condition than in the nonprojective and nontopological conditions.

The averaged hit rates of detecting the change of critical and noncritical dots in the test display are depicted in Figure 4C. Paired $t$ tests revealed a significantly higher hit rate for critical dots (76.25% ± 14.11%) than for noncritical dots, 51.38% ± 17.63%; $t(15) = 4.97$, $p < .01$, in the nonprojective condition. The same pattern was found in the nontopological condition—critical dots: 70.00% ± 10.79%; noncritical dots: 54.88% ± 12.13%; $t(15) = 5.97$, $p < .01$.

Taken together, this experiment provided clear evidence that projective invariance is the most critical geometric property in representing a dynamic configuration in VWM. This could be demonstrated by three discoveries: (a) breaking projective invariance can cause a significant drop in memory performance, (b) breaking a more constrained invariance (i.e., the affine transformation) will not impact the memory performance, and (c) further breaking a more liberal invariance (i.e., the topological invariance) will not further impair the performance. These results were in line with the abrupt collapse hypothesis but were against the continuous decay hypothesis.

The higher hit rate for the critical dots has two implications. It first suggests that the inferior performance of the nonprojective and nontopological conditions cannot be explained by low-level factors such as the "density" or "crowdedness" of the display. An additional analysis of the motion trajectory for the nontopological condition showed that critical dots were the most crowded objects in the display, as their average distance to other items (4.85° ± 0.58°) was actually significantly shorter than that of the noncritical dots, 5.78° ± 1.08°; $t(14) = -3.03$, $p < .01$. A "crowding" effect (e.g., Alvarez & Franconeri, 2007; Ma, McCloskey, & Flombaum, 2014) predicts lower performance for detecting changes of these dots, which is opposite to our results. Second, it also cannot be explained by a "proximity" effect. Additional analyses for the nonaffine condition showed that there is no correlation between the hit rates and the averaged distance between the test item and the other items, $r = .08$, $p > .05$. These results collectively suggest that the higher hit rate for the critical dots can be attributed to their special roles in violating projective and topological transformations.

## Experiment 3: Geometry Transformation, Not Visual Acuity

Here, we further explored whether the superior performance in detecting a direction change in the critical dot(s) was due to some other low-level sensory factor, such as visual acuity, instead of high-level processing of geometric invariances. As the critical dot should move toward the geometric center of the virtual polygon, and since the memory display was also presented around the observer's fixation, the critical dot was moved closer to the fixation point, which could have led to a higher hit rate due to better visual acuity.

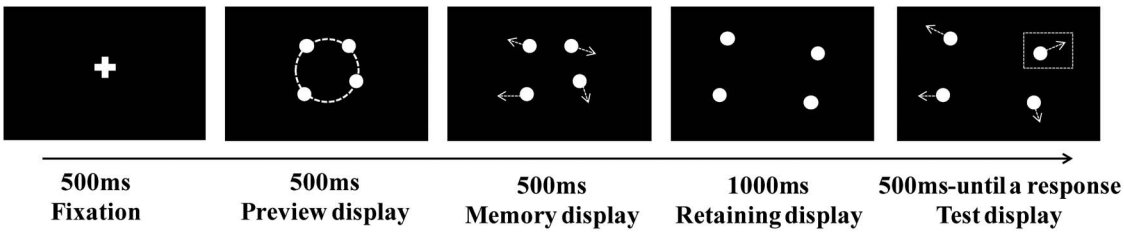


*Figure 2.* A timeline of one trial. The dashed rectangle denotes the item whose moving direction was changed and did not appear during the experiment.

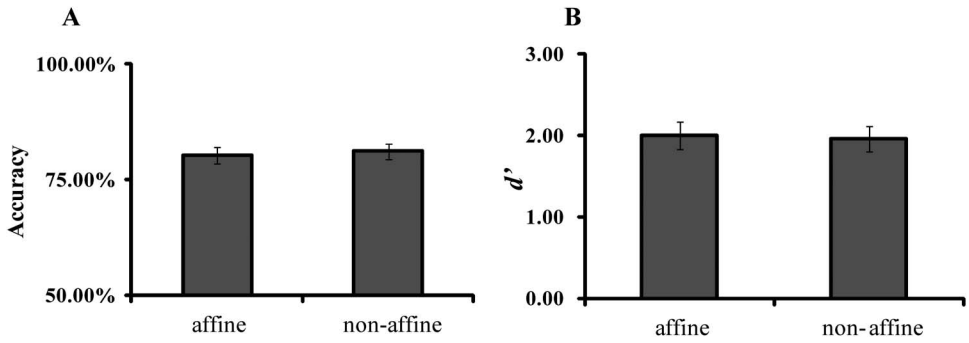


*Figure 3.* Results of Experiment 1. The accuracies of affine and nonaffine conditions (A). *d′*s of affine and nonaffine conditions (B).

We tested this alternative hypothesis by shifting the fixation point away from the center of the memory display by (a) moving all the dots out of the center of the visual field and (b) ensuring that the critical dot was not closer to the fixation point than the others.

## Method

A new group of 16 undergraduates (seven females, 19–25 years of age) from Zhejiang University participated in this experiment.

The ability of attention manipulation and visual information processing would decline along with the target's away from the center of the visual field (Findlay, 1982; Pollatsek, Lesch, Morris, & Rayner, 1992; Posner, 1980; Rayner, McConkie, & Zola, 1980; Rayner & Morris, 1992). We manipulated the position of the fixation cross to 4.4° either left or right from the center of the screen, aligned with the horizontal meridian (see Figure 5). Participants were required to look at the fixation during the whole trial. The four dots were located randomly on the boundary of an invisible circle with this fixation as its center.

The other aspects of Experiment 3 were identical to those of Experiment 2.

## Results and Discussion

The results here were a replication of those in Experiment 2 (see Figure 6). A one-way repeated-measures ANOVA showed that the main effect of the type of transformation was significant, $F(2, 30) = 39.88$, $p < .01$, $\eta_p^2 = 0.73$. Post hoc contrasts revealed that participants' performance in the nonaffine condition (80.23% ± 6.16%) was significantly better than that in the nonprojective condition (71.18% ± 6.28%, $p < .01$) or the nontopological condition (70.14% ± 5.51%, $p < .01$), while no difference was found between the latter two conditions ($p > .05$).

We also considered sensitivity ($d'$ value). One-way ANOVA showed a main effect of type of transformation, $F(2, 30) = 33.59$, $p < .01$, $\eta_p^2 = 0.69$. It was significantly higher in the nonaffine condition (1.97 ± 0.61) than in the nonprojective condition (1.23 ± 0.41, $p < .001$) and the nontopological condition (1.12 ± 0.34, $p < .001$).

We further adopted a paired $t$ test to compare the distance between the critical dots and the fixation to that between the noncritical dots and the fixation, and no significant difference was found in either the nonprojective condition—critical dots: 4.16° ± 0.83°; noncritical dots: 5.17° ± 1.75°; $t(14) = -2.00$, $p > .05$—or the nontopological condition—critical dots: 5.02° ± 1.21°; noncritical dots: 5.76° ± 1.60°; $t(14) = -1.30$, $p > .05$.

Paired $t$ tests revealed a significantly higher hit rate for critical dots (74.17% ± 14.38%) than for noncritical dots, 56.25% ± 10.88%; $t(15) = 4.02$, $p < .01$, in the nonprojective condition. The same pattern was found in the nontopological condition—critical dots: 72.92% ± 7.97%; noncritical dots: 60.21% ± 11.83%; $t(15) = 3.43$, $p < .01$.

Taken together, these results could not provide supportive evidence that the varied performances of the critical and noncritical

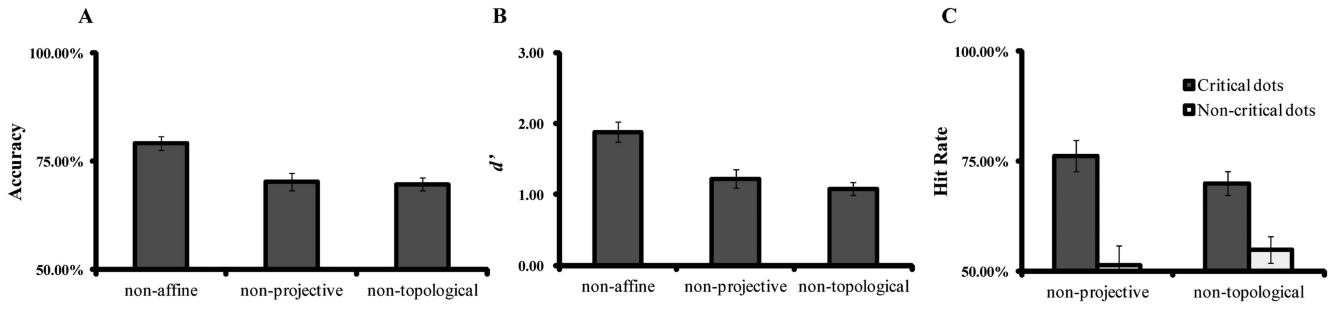


*Figure 4.* Results of Experiment 2. Response accuracies of nonaffine, nonprojective, and nontopological conditions (A). *d′*s of nonaffine, nonprojective, and nontopological conditions (B). Hit rates of critical and noncritical dots in the nonprojective and nontopological conditions, respectively (C).

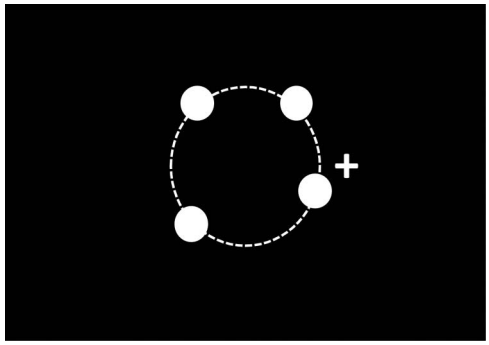

*Figure 5.* An example after the manipulation of fixation in Experiment 3. The dashed circle indicates the location of four dots before moving that did not appear during the experiment.

dots were attributed to their different distances to the fixation. These varied performances were based on their separate roles during the dynamic configuration transformation.

These results showed that the unique effects from breaking projective transformation could be replicated by manipulating the relative position of the memory display and the observer's fixation. Most importantly, these results verified that the superior performance of detecting the critical dots was due to their role in the global geometric transformation.

## Experiment 4: Virtual Polygon Versus Real Polygon

Here, we further explored the scope of our findings by directly comparing the effects of geometric transformation on virtual and real polygons. On one hand, our interpretations of the discoveries so far were based entirely on the virtual polygon formed by the four individual moving dots. These explanations predicted that these effects would still exist with a real polygon by connecting adjacent dots with auxiliary lines. In fact, these effects should be even stronger since the lines can further strengthen the global representation. On the other hand, there was no consensus among early work on the roles of different geometric invariances in real shape constancy (Niall & Macnamara, 1989, 1990; Pizlo, 1994; Todd et al., 1998). It is thus worth exploring how geometric transformations impact the representation of a smoothly changing shape.

### Method

A new group of 16 participants (11 females, 19–28 years of age) from Zhejiang University were enrolled.

A 2 (polygon type: virtual vs. real) × 3 (type of transformation: nonaffine, nonprojective, or nontopological) within-subject design was adopted. In the real shape condition, the envelope lines of four dots were displayed as solid lines (RGB 255, 255, 255, width of 0.2°) during each trial.

Participants were required to complete all six conditions. Two different polygon types were blocked, while different transformations were randomized within each block. The orders of these blocks were counterbalanced across participants. Within each block, each participant performed 80 trials per type-of-transformation condition, resulting in 240 trials for each block and 480 trials for the entire experiment. Each block was further divided into four sections, with a 2-min break between each two of them. The other aspects of Experiment 4 were identical to those of Experiment 2.

### Results and Discussion

Figures 7A and 7B depict the accuracy and $d'$ as a function of polygon type and type of transformation. The most important result was that, in each type of geometric manipulation, the effect on the real polygon was the same as that on the virtual polygon. Two-way ANOVA for accuracy revealed a significant main effect of type of transformation, $F(2, 30) = 82.46$, $p < .001$, $\eta_p^2 = 0.85$, and polygon type, $F(1, 30) = 5.35$, $p < .05$, $\eta_p^2 = 0.26$, as well as a significant interaction between polygon type and type of transformation, $F(2, 30) = 5.35$, $p < .05$, $\eta_p^2 = 0.26$.

We then conducted two one-way repeated-measures ANOVAs for virtual and real conditions, respectively. In the virtual polygon condition, the main effect of type of transformation was significant, $F(2, 30) = 37.37$, $p < .01$, $\eta_p^2 = 0.71$. Post hoc contrasts revealed that participants' performance in the nonaffine condition (78.13% ± 5.53%) was significantly better than that in the nonprojective condition (68.98% ± 5.46%; $p < .01$) or the nontopological condition (69.22% ± 6.42%; $p < .01$), while no difference was found between the latter two conditions ($p > .05$). For the performance in the real polygon condition, the main effect of type of transformation was significant, $F(2, 30) = 54.37$, $p < .01$, $\eta_p^2 = 0.78$. Post hoc contrasts revealed that participants' performance in the nonaffine condition (78.91% ± 6.27%) was significantly better than that in the nonprojective condition (64.22% ± 3.65%; $p <$

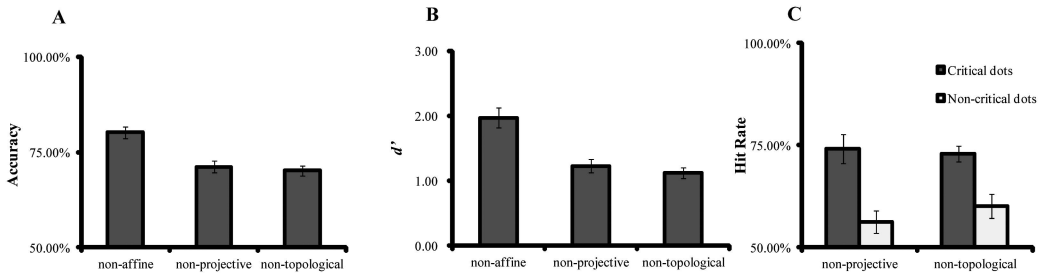


*Figure 6.* Results of Experiment 3. Response accuracies in the nonaffine, nonprojective, and nontopological conditions (A). $d'$s in the nonaffine, nonprojective, and nontopological conditions (B). Hit rates for critical and noncritical dots in the nonprojective and nontopological conditions, respectively (C).

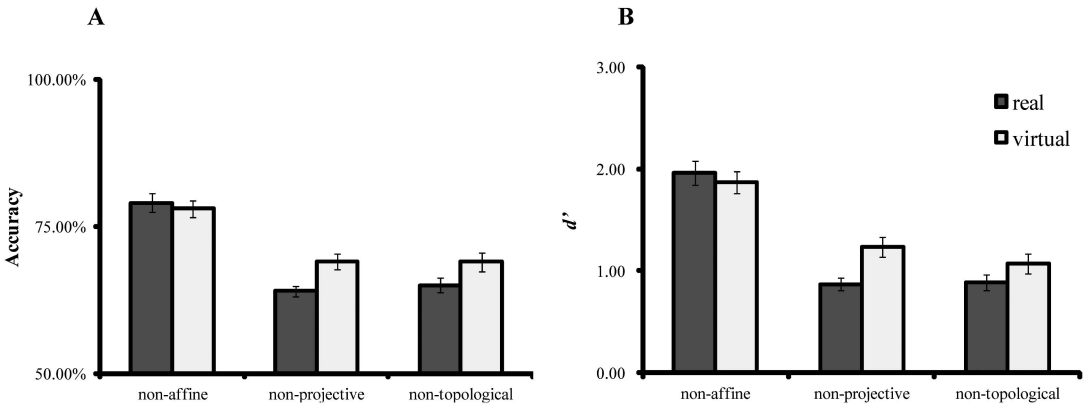


*Figure 7.* Results of Experiment 4. Response accuracies in the nonaffine, nonprojective, and nontopological conditions (A). $d'$s in the nonaffine, nonprojective, and nontopological conditions (B).

.01) or the nontopological condition (64.92% ± 5.05%; $p < .01$), while no difference was found between the latter two conditions ($p > .05$).

Concerning the comparison between the real and virtual polygon conditions, paired $t$ tests for both the nonprojective and nontopological conditions revealed worse memory accuracies in the real condition than those in the virtual condition—nonprojective: $t(15) = 3.22$, $p < .01$; nontopological: $t(15) = 2.20$, $p < .05$. However, in the nonaffine condition, no significant difference was found between the real and virtual polygons, $t(15) = -0.58$, $p > .05$.

In order to exclude any carryover effect in Experiment 4, the participants were divided into two subgroups according to their polygon type order; that is, one group first did the virtual polygon block, while the other group first did the real polygon block. We conducted 2 (block order: virtual first vs. real first) × 3 (type of transformation: nonaffine, nonprojective, or nontopological) mixed-design ANOVAs for both the accuracy and $d'$. ANOVA for accuracy revealed a significant main effect of type of transformation, $F(2, 28) = 36.33$, $p < .001$, $\eta_p^2 = 0.72$; however, there was no interaction between the order and the type of transformation, $F(2, 28) = 0.43$, $p > .05$, $\eta_p^2 = 0.03$. Analyzing sensitivity ($d'$ value) showed exactly the same pattern of results, which are not included here because of space limitations.

The results demonstrated that representations of virtual polygons and real polygons are governed by the same set of geometric rules, in which projective invariance plays a particularly important role. Memory performance was improved when projective invariance was kept during movement. On the contrary, the performance was impaired when projective invariance was violated.

## General Discussion

In the current study, we explored the geometric rules governing the representation of dynamic configurations in VWM. We disrupted different types of geometric invariances one at a time and observed how that particular geometric rule influenced the performance of a change-detection task. This approach yielded a rich set of results: (a) Memory performance dropped abruptly when the projective invariance of the memory display was disrupted, (b) breaking other types of invariance had no effect (e.g., affine) or had no additive effect (e.g., topology), (c) memory was biased toward processing the object(s) responsible for breaking projective invariance, and (d) all of the above effects also occurred when the virtual polygon was turned into a real polygon by connecting each individual object with solid lines.

These results collectively demonstrated how the representation of a configuration is maintained and collapsed in a dynamic scene with a smooth transformation. The implications of these results are discussed in detail below.

### Configuration as a Hierarchical Representation

Configuration is an important component of human vision. It should be noted that in psychophysical studies on VWM, configuration is typically objectively manipulated but not explicitly defined (Jiang et al., 2000; Zimmer & Lehnert, 2006). Here, we start our discussion by explicitly defining what we mean (and do not mean) by configuration by connecting human psychophysics with information and probability theories.

Configuration can be defined as a Markov random field (see Brady & Tenenbaum, 2013; Kindermann & Snell, 1980, for an excellent example of using this model to explore working memory). This type of configuration has the following assumptions: (a) An item's features should correlate with those of its neighbors and (b) being conditional on its neighbors, an item is independent of other items in the scene (i.e., the Markov assumption). Figure 8A illustrates one possible Markov random field model of the memory displays in the current study. In this particular model, Item A's motion is independent of Item D given B and C, as the paths from A to D are all blocked by observing B and C. We should empha-

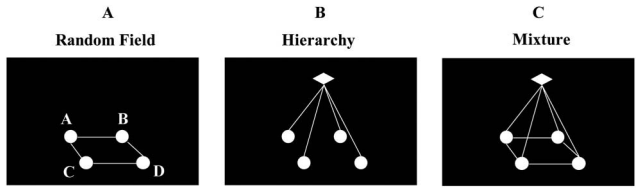


*Figure 8.* Three possible configurations. See details in General Discussion.

size here that this type of configuration does not fit our geometric manipulation of the memory display for the reasons discussed below.

The second type of configuration is defined as a hierarchical structure (see Figure 8B), in which each item is the component of a "super" item defined at a higher level in the hierarchy. Given this definition, each item's motion is the outcome of the smooth transformation of the higher level unit, which is a polygon in our study. For comparison, while the Markov random field is defined as an undirected, noncausal correlation in a (flat) field, this type of configuration, to which our present geometric invariance manipulation belongs, is defined as directed and causal interactions in a hierarchy. These two types of configurations are not mutually exclusive but can be combined to represent complicated real scenes (Figure 8C). However, this is outside the scope of our psychophysical manipulations, so we shall not expand on this point.

One intriguing aspect of the hierarchical structure is that the concept of *object* does not have a special status in it. A *unit* or a *node* in the hierarchy can be as global as the entire scene or as specific as a component of a part of an object (e.g., the texture of a leg of a table). This idea has long been discussed in psychology. For instance, according to Palmer (1977), no fundamental difference should exist between parts and a whole object. An object is composed of parts, while the parts themselves are possibly composed of parts, and each of these levels is referred to as a structural unit. This argument has been supported by studies showing that a similar mechanism underlying object-based attention (Egly, Driver, & Rafal, 1994) also operates at the level of parts (Vecera, Behrmann, & Filapek, 2001) and at the level of the group (Driver & Baylis, 1998). We argue that the distinction between an object and a structured unit is not rhetorical but has significant theoretical implications. An object emphasizes an entity that can be segmented from a background, while a unit or a node emphasizes an entity that exists only in a hierarchical structure, which is the focus of the current study. Highlighting this distinction is especially important for future research on VWM. In contrast to the extensive VWM studies on the storage of individual objects, only a few studies have explored the storage of multiple objects' relationships (e.g., Clevenger & Hummel, 2014; Hummel & Biederman, 1992). How VWM can be allocated to a hierarchical structure is largely unknown.

## Why Projective Invariance?

Our results indicate that projective invariance plays a critical role in defining dynamic configurations in VWM. Since the configuration is defined by the smooth transformation of a virtual polygon, these results shed light on how vision persistently represents 3D rigid shapes with various deformations. Interestingly, studying *deformable shapes* is also a major research topic in the computer vision community (e.g., Siddiqi & Pizer, 2008). The challenge of representing a shape persistently is that the same 3D object's projection onto 2D images can change dramatically due to the motion of the object or the observer. As we have elaborated in the Introduction, the central task of vision is to recognize 3D scenes from their projections onto 2D images. Such a process is governed by projective geometry. Therefore, it is desirable to represent the 2D shape in a way that is invariant to changes caused by the projections, including size, scaling, rotation, and stretching. Shape representation that fails to capture such invariances will make vision incapable of representing the same shape persistently across viewing conditions. On the contrary, tolerating dramatic changes beyond projective transformation (such as concavity and connectivity) will impose risks of representing different shapes as the same one. From this perspective, the current result does reflect a shape representation that is optimal in the context of "inverse projection." In the meantime, we also want to point out that the above interpretation is an ad hoc explanation. As we have argued in the Introduction, there is no strong theoretical account that can precisely predict how global geometric transformation can impact VWM. Future studies are required to examine the validity of such an interpretation.

We also want to point out that the shape invariance explored in our work is mostly like the primitives for more sophisticated configurational representation. For instance, in models of the animal body, skeletons of the animal can be represented by composing primitive shapes with joints. The composition of primitive shapes is governed by a set of "visual grammar" (e.g., Zhu & Yuille, 1996). The deformation of these composed shapes can certainly violate projective transformation, which only limits the motion of each primitive.[6]

## From Randomized Displays to Structured Informative Scenes

In the current study, the memory display is not randomly generated by simply combining the motion of each object. Instead, the entire display is generated given certain geometric rules at the scene level. This methodological choice has theoretical implications. We argue that it is necessary to scale up the investigation of VWM from the storage of individual objects to the entire visual scene with a deep hierarchical structure. To establish a connection between VWM and scene understanding in general, we refer to the memory display in the change-detection task as a single scene. In most of the previous studies on VWM (including those on static and dynamic configurations), the scene is randomly generated with little cross-object correlation (e.g., the color of each item is independently sampled from a set of distinctive colors). If we assume that there is a scene distribution (defined explicitly or implicitly) and that a scene used in a single change-detection trial is a sample from that distribution, then the scene distribution is a uniform distribution in most of the previous studies. In other words, every combination of features from a different object has the same probability of appearing in the scene. It is well known that a uniform distribution has the maximized randomness or entropy (without other constraints on the distribution; e.g., Park & Bera, 2009). Such a randomized distribution is quite useful for experimental design, as it allows researchers to eliminate any potential systematic effect that is not manipulated or controlled in an experiment.

Nevertheless, the adoption of maximally randomized scenes also creates two challenges for understanding the nature of VWM. First, it makes the connection between VWM and real scene understanding (e.g., Li, VanRullen, Koch, & Perona, 2002; Oliva

[6] We thank an anonymous reviewer for raising the question of articulated motions.

& Torralba, 2001) rather unclear. It has been demonstrated that the distributions of visual features in real scenes are far from uniform or Gaussian but contain complicated structures (e.g., Mumford & Desolneux, 2010). These structures include either short-range correlations that can be captured by a Markov random field (Figure 8A) or long-range correlations demanding a hierarchical top-down representation (such as the virtual polygon employed in our work; see Figure 8B). Since any structure beyond an individual object is eliminated by the randomized design, given our knowledge of VWM so far, it remains largely unknown how VWM engages in real scene perception. Second, with randomized scenes, information theory becomes largely useless. It is interesting to note that while the phrase *visual information* is used pervasively in studies of human vision, few of them actually make a connection to information theory. One notable exception is the studies on human eye movements, which argued that eye movements maximize information gain (e.g., Bruce & Tsotsos, 2009; Najemnik & Geisler, 2005). Specific to studies of VWM, the usage of maximally randomized scenes may be an important reason for the lack of such a connection. Information theory, at its most fundamental level, relies on the structure of the signal's distribution. Perhaps the most basic application of information theory is to assign shorter codes to signals with higher probabilities. With scenes sampled from a uniform distribution, even this basic principle is not helpful, as every scene has exactly the same low probability. Therefore, no efficient coding scheme exists. This is not surprising, as by using randomized scenes, one essentially excludes any information at the scene level. One cannot apply information theory where information does not exist.

The importance of combining structured scenes and information theory has been well demonstrated in the computer vision community. In fact, the central goal of many modern computer vision algorithms is to model real scene distributions by using principles derived from information theory, such as maximum entropy, maximum information gain, and minimum encoding length (Yuille, Ruiz, Pérez, & Bonev, 2009). By integrating these principles creatively, one can model real scene distributions as tightly as possible by using as few features as possible, so that the model is both parsimonious and generalizable (for an excellent example, see Zhu, Wu, & Mumford, 1997). Nevertheless, it is difficult to bridge the gap between human vision and computer vision, since these types of models do not apply to scenes that are highly randomized.

In our study, we move one step forward to address this challenge by using a display that is not randomly generated by using highly structured distributions defined by several types of geometric transformations. The psychophysical results confirm the importance of the structure beyond individual objects. However, these distributions are still manipulated and created by the experimenters rather than estimated from a set of real sense. In the future, it would be important to explore how VWM encodes structural displays sampled from real scene distributions. In the meantime, there are also recent studies on ensemble perception and scene statistics (e.g., Albrecht & Scholl, 2010; Alvarez, 2011; Ariely, 2001; Chong & Treisman, 2003, 2005; Marchant, Simons, & de Fockert, 2013), which we hope can also help the studies on working memory move toward more informative scenes with intriguing probability distributions.

## Conclusion

Our study explored the representation of dynamic configurations by synthesizing two perspectives of visual perception. One perspective is that the construction and maintenance of visual representation are partially determined by certain geometric rules, such as those in Klein's Erlangen program. The other perspective is that VWM should be sensitive to scene structure beyond individual objects. Based on these two perspectives, we conducted a series of psychophysical experiments and demonstrated how projective invariance of the global configuration is critical in maintaining a dynamic configuration in VWM.

## References


Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science, 21,* 560–567. http://dx.doi.org/10.1177/0956797610363543

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

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

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision, 7,* 14. http://dx.doi.org/10.1167/7.13.14

Anderson, J. W. (2006). *Hyperbolic geometry*. London, UK: Springer Science & Business Media.

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

Baddeley, A. (1998). Recent developments in working memory. *Current Opinion in Neurobiology, 8,* 234–238. http://dx.doi.org/10.1016/S0959-4388(98)80145-1

Blake, R., Cepeda, N. J., & Hiris, E. (1997). Memory for visual motion. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 353–369. http://dx.doi.org/10.1037/0096-1523.23.2.353

Blaxton, C. L., Fehd, H. M., & Seiffert, A. E. (2011). Center-looking suggests grouping rather than separate attentional foci in multiple object tracking. *Journal of Vision, 11,* 279. http://dx.doi.org/10.1167/11.11.279

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,* 85–109. http://dx.doi.org/10.1037/a0030779

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9,* 5. http://dx.doi.org/10.1167/9.3.5

Chen, L. (2001). Perceptual organization: To reverse back the inverted (upside-down) question of feature binding. *Visual Cognition, 8,* 287–303. http://dx.doi.org/10.1080/13506280143000016

Chen, L. (2005). The topological approach to perceptual organization. *Visual Cognition, 12,* 553–637. http://dx.doi.org/10.1080/13506280444000256

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43,* 393–404. http://dx.doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45,* 891–900. http://dx.doi.org/10.1016/j.visres.2004.10.004

Clevenger, P. E., & Hummel, J. E. (2014). Working memory for relations among objects. *Attention, Perception, & Psychophysics, 76,* 1933–1953. http://dx.doi.org/10.3758/s13414-013-0601-3

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

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General, 123,* 161–177. http://dx.doi.org/10.1037/0096-3445.123.2.161

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. http://dx.doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Feldman, J., Singh, M., Briscoe, E., Froyen, V., Kim, S., & Wilder, J. (2013). An integrated Bayesian approach to shape representation and perceptual organization. In S. J. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision* (pp. 55–70). London, UK: Springer Science & Business Media. http://dx.doi.org/10.1007/978-1-4471-5195-1_4

Findlay, J. M. (1982). Global visual processing for saccadic eye movements. *Vision Research, 22,* 1033–1045. http://dx.doi.org/10.1016/0042-6989(82)90040-2

Franconeri, S. L., Jonathan, S. V., & Scimeca, J. M. (2010). Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity. *Psychological Science, 21,* 920–925. http://dx.doi.org/10.1177/0956797610373935

Gmeindl, L., Nelson, J. K., Wiggin, T., & Reuter-Lorenz, P. A. (2011). Configural representations in spatial working memory: Modulation by perceptual segregation and voluntary attention. *Attention, Perception, & Psychophysics, 73,* 2130–2142. http://dx.doi.org/10.3758/s13414-011-0180-0

Grossberg, S. (2014). How visual illusions illuminate complementary brain processes: Illusory depth from brightness and apparent motion of illusory contours. *Frontiers in Human Neuroscience, 8,* 854. http://dx.doi.org/10.3389/fnhum.2014.00854

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review, 92,* 173–211. http://dx.doi.org/10.1037/0033-295X.92.2.173

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99,* 480–517. http://dx.doi.org/10.1037/0033-295X.99.3.480

Jiang, Y., Chun, M. M., & Olson, I. R. (2004). Perceptual grouping in change detection. *Perception & Psychophysics, 66,* 446–453. http://dx.doi.org/10.3758/BF03194892

Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 683–702. http://dx.doi.org/10.1037/0278-7393.26.3.683

Kanizsa, G. (1974). Contours without gradients or cognitive contours? *Giornale Italiano di Psicologia, 1,* 93–113.

Kindermann, R., & Snell, J. L. (1980). *Markov random fields and their applications*. Providence, RI: American Mathematical Society. http://dx.doi.org/10.1090/conm/001

Kulkarni, T. D., Yildirim, I., Kohli, P., Freiwald, W. A., & Tenenbaum, J. B. (2014). *Deep generative vision as approximate Bayesian computation*. Approximate Bayesian Computation Workshop, Neural Information Processing Systems, Montreal, Canada.

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 of the United States of America, 99,* 9596–9601. http://dx.doi.org/10.1073/pnas.092277599

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

Ma, Z., McCloskey, M., & Flombaum, J. (2014). Differentiating between object-dependent and transient-dependent motion percepts through crowding. *Journal of Vision, 14,* 278. http://dx.doi.org/10.1167/14.10.278

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica, 142,* 245–250. http://dx.doi.org/10.1016/j.actpsy.2012.11.002

McKeefry, D. J., Burton, M. P., & Vakrou, C. (2007). Speed selectivity in visual short term memory for motion. *Vision Research, 47,* 2418–2425. http://dx.doi.org/10.1016/j.visres.2007.05.011

Mumford, D., & Desolneux, A. (2010). *Pattern theory: The stochastic analysis of real-world signals*. London, UK: A K Peters/CRC Press.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434,* 387–391. http://dx.doi.org/10.1038/nature03390

Narasimhan, S., Tripathy, S. P., & Barrett, B. T. (2009). Loss of positional information when tracking multiple moving dots: The role of visual memory. *Vision Research, 49,* 10–27. http://dx.doi.org/10.1016/j.visres.2008.09.023

Niall, K. K., & MacNamara, J. (1989). Projective invariance and visual shape constancy. *Acta Psychologica, 72,* 65–79. http://dx.doi.org/10.1016/0001-6918(89)90051-6

Niall, K. K., & Macnamara, J. (1990). Projective invariance and picture perception. *Perception, 19,* 637–660. http://dx.doi.org/10.1068/p190637

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145–175. http://dx.doi.org/10.1023/A:1011139631724

Olson, I. R., & Marshuetz, C. (2005). Remembering "what" brings along "where" in visual working memory. *Perception & Psychophysics, 67,* 185–194. http://dx.doi.org/10.3758/BF03206483

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology, 9,* 441–474. http://dx.doi.org/10.1016/0010-0285(77)90016-0

Papenmeier, F., Huff, M., & Schwan, S. (2012). Representation of dynamic spatial configurations in visual short-term memory. *Attention, Perception, & Psychophysics, 74,* 397–415. http://dx.doi.org/10.3758/s13414-011-0242-3

Park, S. Y., & Bera, A. K. (2009). Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics, 150,* 219–230. http://dx.doi.org/10.1016/j.jeconom.2008.12.014

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

Pizlo, Z. (1994). A theory of shape constancy based on perspective invariants. *Vision Research, 34,* 1637–1658. http://dx.doi.org/10.1016/0042-6989(94)90123-6

Pollatsek, A., Lesch, M., Morris, R. K., & Rayner, K. (1992). Phonological codes are used in integrating information across saccades in word identification and reading. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 148–162. http://dx.doi.org/10.1037/0096-1523.18.1.148

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32,* 3–25. http://dx.doi.org/10.1080/00335558008248231

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80,* 127–158. http://dx.doi.org/10.1016/S0010-0277(00)00156-6

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3,* 179–197. http://dx.doi.org/10.1163/156856888X00122

Rayner, K., McConkie, G. W., & Zola, D. (1980). Integrating information across eye movements. *Cognitive Psychology, 12,* 206–226. http://dx.doi.org/10.1016/0010-0285(80)90009-2

Rayner, K., & Morris, R. K. (1992). Eye movement control in reading: Evidence against semantic preprocessing. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 163–172. http://dx.doi.org/10.1037/0096-1523.18.1.163

Scholl, B. J. (2005). Innateness and (Bayesian) visual perception. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and contents* (pp. 34–52). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1093/acprof:oso/9780195179675.003.0003

Shen, M., Huang, X., & Gao, Z. (2015). Object-based attention underlies the rehearsal of feature binding in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 41,* 479–493. http://dx.doi.org/10.1037/xhp0000018

Shooner, C., Tripathy, S. P., Bedell, H. E., & Öğmen, H. (2010). High-capacity, transient retention of direction-of-motion information for multiple moving objects. *Journal of Vision, 10,* 8. http://dx.doi.org/10.1167/10.6.8

Siddiqi, K., & Pizer, S. M. (Eds.). (2008). *Medial representations: Mathematics, algorithms and applications*. London, UK: Springer Science & Business Media. http://dx.doi.org/10.1007/978-1-4020-8658-8

Tittle, J. S., Todd, J. T., Perotti, V. J., & Norman, J. F. (1995). Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 663–678. http://dx.doi.org/10.1037/0096-1523.21.3.663

Todd, J. T., & Bressan, P. (1990). The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Perception & Psychophysics, 48,* 419–430. http://dx.doi.org/10.3758/BF03211585

Todd, J. T., Chen, L., & Norman, J. F. (1998). On the relative salience of Euclidean, affine, and topological structure for 3-D form discrimination. *Perception, 27,* 273–282. http://dx.doi.org/10.1068/p270273

Vecera, S. P., Behrmann, M., & Filapek, J. C. (2001). Attending to the parts of a single object: Part-based selection limitations. *Perception & Psychophysics, 63,* 308–321. http://dx.doi.org/10.3758/BF03194471

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 1436–1451. http://dx.doi.org/10.1037/0096-1523.32.6.1436

von Helmholtz, H. (1867). *Treatise on physiological optics* (Vol. III; J. P C. Southall, Trans.). Toronto, Canada: General Publishing Company, Ltd.

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

Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology, 24,* 295–340. http://dx.doi.org/10.1016/0010-0285(92)90010-Y

Yuille, A. L., Ruiz, F. E., Pérez, P. S., & Bonev, B. I. (2009). *Information theory in computer vision and pattern recognition*. London, UK: Springer Science & Business Media.

Zhao, L., Gao, Q., Ye, Y., Zhou, J., Shui, R., & Shen, M. (2014). The role of spatial configuration in multiple identity tracking. *PLoS ONE, 9,* e93835. http://dx.doi.org/10.1371/journal.pone.0093835

Zhu, S. C., Wu, Y. N., & Mumford, D. (1997). Minimax entropy principle and its application to texture modeling. *Neural Computation, 9,* 1627–1660.

Zhu, S. C., & Yuille, A. L. (1996). FORMS: A flexible object recognition and modelling system. *International Journal of Computer Vision, 20,* 187–212. http://dx.doi.org/10.1007/BF00208719

Zimmer, H. D., & Lehnert, G. (2006). The spatial mismatch effect is based on global configuration and not on perceptual records within the visual cache. *Psychological Research, 70,* 1–12. http://dx.doi.org/10.1007/s00426-004-0186-5

Zokaei, N., Gorgoraptis, N., Bahrami, B., Bays, P. M., & Husain, M. (2011). Precision of working memory for visual motion sequences and transparent motion surfaces. *Journal of Vision, 11,* 2. http://dx.doi.org/10.1167/11.14.2

Received January 24, 2015
Revision received May 5, 2015
Accepted May 11, 2015 ■