Journal of Experimental Psychology: General
2017, Vol. 146, No. 6, 896–909

0096-3445/17/$12.00 http://dx.doi.org/10.1037/xge0000310

Seeing "What" Through "Why": Evidence From Probing the Causal Structure of Hierarchical Motion

Haokui Xu, Ning Tang, Jifan Zhou, and Mowei Shen
Zhejiang University

Tao Gao
Massachusetts Institute of Technology

Although our world is hierarchically organized, the perception, attention, and memory of hierarchical structures remain largely unknown. The current study shows how a hierarchical motion representation enhances the inference of an object's position in a dynamic display. The motion hierarchy is formed as an acyclic tree in which each node represents a distinctive motion component. Each individual object is instantiated as a node in the tree. In a position inference task, participants were asked to infer the position of a target object, given how it moved jointly with other objects. The results showed that the inference is supported by the context formed by nontarget objects. More importantly, this contextual effect is (a) *structured*, with stronger support from objects forming a hierarchical tree than from those moving independently; (b) *degreed*, with stronger support from objects closer to the target in the motion tree; and (c) *directed*, with stronger support from the target's ancestor nodes than from its descendent nodes. Computational modeling results further indicated that the contextual effect cannot be explained by correlated and contingent movements without an explicit causal representation of the motion hierarchy. Together, these studies suggest that human vision is a type of intelligence, which sees what are in the dynamic displays by recovering why and how they are generated.

Keywords: hierarchical representation, Bayesian modeling, causality, motion perception, visual working memory

One of the most important discoveries made over the past several decades of vision research is the strikingly limited capacity of attention and other higher-level visual cognition, including evidence from attentional blink (e.g., Chun & Potter, 1995; Raymond, Shapiro, & Arnell, 1992), change blindness (e.g., Simons & Levin, 1997; Rensink, 2002), visual working memory (e.g., Luck & Vogel, 1997), and multiobject tracking (e.g., Pylyshyn & Storm, 1988; Flombaum, Scholl, & Pylyshyn, 2008). Interestingly, such resource limitations do not necessarily imply the failure of human vision. Instead, they collectively raise a research question that is even more challenging: how can humans navigate through the real world by processing a stream of complex visual inputs with such a small amount of attention and working memory resources? Addressing this challenge is probably beyond the scope of human vision alone and may require interdisciplinary perspectives and research paradigms, including those from psychophysics, computer vision, and artificial intelligence. Consistent with this proposal, recent state-of-the-art computer vision algorithms have started to implement attention and memory as core components (e.g., Caicedo & Lazebnik, 2015; Karpathy & Li, 2015; Xu et al., 2015). Although these vision models are cognitively motivated to some degree, their implementation of attention and memory are primarily for computational purposes: it turns out that even with a massive amount of computational resources, a computer vision model still performs better when it can focus its resources on a selected region of the visual image and then updates its selection sequentially. This architecture naturally leads to implementations of selective attention over space and working memory over time in these computer vision studies.

From the perspective of human vision research, addressing the resource limitation challenge requires not only revealing the properties of the capacity limitations but also demonstrating the types of visual representation that can be efficiently constructed, maintained, and operated with these limited resources. Numerous studies have investigated "object" as the underlying visual representation (e.g., Duncan, 1984; Scholl, 2001). Constructing coherent object representations is certainly indispensable. However, in the current study, we do not emphasize the individualization of certain visual entities. Instead, we treat the entire visual scene as a hierarchically composed structure that can be recursively parsed into substructures.

This article was published Online First April 27, 2017.

Haokui Xu, Ning Tang, Jifan Zhou, and Mowei Shen, Department of Psychology, Zhejiang University; Tao Gao, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

This research was supported by the National Natural Science Foundation of China (31571119, 31600881) and Project of Ministry of Science and Technology of the People's Republic of China (2016YFE0130400). Tao Gao is supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF STC award CCF-1231216. A preliminary version of this work was presented at the 31st International Congress of Psychology. Abstract of this conference can be found in International Journal of Psychology—Special Issue: 31st International Congress of Psychology (http://onlinelibrary.wiley.com/doi/10.1002/ijop.2016.51.issue-S1/issuetoc?winzoom = 1).

Correspondence concerning this article should be addressed to Tao Gao, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, or to Mowei Shen and Jifan Zhou, Department of Psychology, Zhejiang University, Hangzhou, China, 310028. E-mail: taogao@mit.edu or mwshen@zju.edu.cn or jifanzhou@zju.edu.cn

We investigate the hierarchical representation in the context of perceiving and storing the movements of multiple objects for the following reasons. First, tracking and understanding the movements of multiple objects are among the core capabilities of human vision. They serve as the foundation for representing not only objects (e.g., Scholl & Pylyshyn, 1999) but also events (e.g., Zacks & Tversky, 2001; Zacks & Swallow, 2007) and animacy (Castelli, Happé, Frith, & Frith, 2000; Gao, Newman, & Scholl, 2009; Heider & Simmel, 1944). Second, previous studies (e.g., Papenmeier, Huff, & Schwan, 2012; Sun et al., 2015; Yantis, 1992; Zhao et al., 2014) have established contextual effects in tracking and storing the movements of multiple objects, showing that the processing of an object's movements is placed into a global context of configuration formed by all the objects' joint movements. Context in these studies is loosely defined. However, these studies collectively suggests that defining and manipulating a hierarchical motion structure can produce novel results that have not been revealed by object-based paradigms. Third, a novel approach for modeling hierarchical motions has recently been discovered (Gershman, Tenenbaum, & Jäkel, 2016). This approach provides us with a useful opportunity to compare our psychophysical results against model predictions. According to the three levels of analysis (Marr, 1982), we view this approach as a good practice of synthesizing the computational level (the first level) and the algorithmic level (the second level).

In the following subsections, we briefly review the history of exploring hierarchical representation in cognitive science, the unique advantages of hierarchical representation as a type of mental representation, and the specific hierarchical model explored in the current study.

Why Hierarchical Representation?

It is necessary to discuss what hierarchical representation can offer before delving into its details. First, compared with a shallow structure, a deep hierarchical structure is computationally more efficient. This is mainly because in a hierarchical structure, the properties of a parent node will be shared and reused by all its descendent nodes.[1] This advantage partly explains the rise of deep neural networks in recent years (e.g., LeCun, Bengio, & Hinton, 2015). Specifically, in a visual scene represented by a tree structure, a feature associated with a node will be shared by all its descendent nodes, causing it to appear multiple times. Nevertheless, it needs to be stored only once in a hierarchy if vision can correctly parse the scene as a tree structure. The computational efficiency of feature sharing is especially important, given the limited resources of visual cognition. Second, a hierarchical structure enables flexible abstraction, which can further reduce the computational demands of representing a visual scene. The physical world itself is hierarchically organized, and it is causally understood by different levels of abstraction. As an example specific to vision, when viewing the coordinated movements of several people, an observer can adeptly set the abstraction at the group level, individual level, or subbody level (including gaze and body parts). All of the three levels of abstraction in the human mind can be provided by a unified hierarchical representation that mirrors the structure of the real world. Third, a hierarchical representation can support two-way interaction between language and vision. Humans can easily translate vision into language by describing a dynamic visual display as a story (e.g., Heider & Simmel, 1944) or by grounding nouns and verbs in a sentence to objects, spatial properties, and events in the visual scene (e.g., Gorniak & Roy, 2004; Jackendoff, 1996; Talmy, 1988). It is well known that vision starts by extracting low-level features, such as color and orientation (e.g., Treisman & Gelade, 1980), and those defined by Gabor-filters (e.g., Julesz, 1981). However, the properties of the outputs of vision that allow such seamless interactions with language remain largely unknown. This challenge is known as the semantic gap of vision. A hierarchical visual representation can potentially address this challenge. In linguistics, understanding the meaning of a sentence starts from parsing it as a hierarchical grammar tree (e.g., Chomsky, 1964). The vision-language interaction can be more efficient if a similar parsing process also exists for understanding the meaning of a visual scene, although the grammar of vision can be quite different.

Hierarchical Representation in Language and Vision

Hierarchical representation in the human mind has long been explored, mainly in the domains of language (e.g., Chomsky, 1964) and semantic representation (e.g., Smith, Shoben, & Rips, 1974). A thorough review of hierarchical language representation is outside the scope of the current study. In short, it has been shown that the power of a hierarchical structure lies in the "infinite use of finite means" (Humboldt, 1836): by composing limited primitives with limited rules in a hierarchical tree, a sentence can nevertheless express infinite meanings. Recent studies have further investigated hierarchical representation in the domain of concept learning, showing that young children learn and organize new concepts in a hierarchical structure (e.g., Kiley Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Johnson & Keil, 2014).

Hierarchical representation has also been explored in vision science. It has been suggested that the entire human visual processing system is hierarchical in nature, (e.g., Biederman, 1987), from generic perceptual features (e.g., Palmer, 1977) to midlevel object representations (e.g., Kahneman, Treisman, & Gibbs, 1992) and high-level category-specific concepts (e.g., Ullman, 2007). Cluster as a specific type of hierarchy has been explored in the context of working memory (Brady & Tenenbaum, 2013; Lew & Vul, 2015) and ensemble perception (e.g., Cain, Dobkins, & Vul, 2016), indicating that vision encodes the clusters formed by individual objects. A recent study has modeled human perceptual grouping with a hierarchical tree structure (Froyen, Feldman, & Singh, 2015). In this model, a primitive element is represented as a child node, whereas an upper-level object is represented as a parent node. By using Bayesian inference, the model can infer which object owns a primitive element. The results show that inference over a hierarchical structure can solve a wide range of classical perceptual grouping problems, including challenging cases such as dot clustering, contour integration, and part decomposition.

In the field of computer vision, various algorithms based on hierarchical representation have been proposed, many of which are actually inspired by models of human language. In a framework proposed by Zhu and his colleagues (e.g., Zhu, 1999; Zhu &

[1] Here we use terms from graph theory to introduce the properties of a hierarchical structure.

Mumford, 2007), an entire visual scene is represented by an and-or graph, which is essentially a type of context-free grammar (e.g., Chi & Geman, 1998). Using this graph, a visual scene is parsed as a tree composed of primitive features, just as a sentence is parsed as a tree composed of words. As suggested by its name, an and-or graph has two types of nodes: an "and" node defines the conjunction of parts that a visual concept must exhibit, whereas an "or" node defines the alternative design choices for instantiating an object. For example, a table must have a top and four legs (and node), but its color may vary (or node).

Most relevant to our current study on dynamic displays is a recently proposed hierarchical model of motion perception (Gershman et al., 2016). According to this model, the visual system performs a vector analysis of moving objects, decomposing them into shared and relative motions that are placed at different levels of a motion tree. Objects sharing the same motion are represented as sibling nodes in the tree; although they own their relative motions, their shared motion is owned by a parent node. The results show that Bayesian inference over this motion tree structure can not only explain classical motion perception phenomena but also confirm new hypotheses (Johansson, 1950; Snowden & Verstraten, 1999). The details of this model will be further described in the *General method* section.

Current Study

A hierarchical representation implies that objects are not processed in isolation but are placed into a global context that is recursively structured. It has been shown that visual processing of a target object is facilitated by its connections with the surrounding nontarget objects (e.g., Gmeindl, Nelson, Wiggin, & Reuter-Lorenz, 2011; Jiang, Olson, & Chun, 2000; Sun et al., 2015; Yantis, 1992). How do we demonstrate that human vision represents such a global context with a hierarchical structure? Studies on object-based representation have been successful in part because researchers have devised many paradigms for manipulating objecthood. Accordingly, in the current study, we propose three approaches for manipulating hierarchy-hood while keeping the number of objects fixed. The first approach is to manipulate the depth of a hierarchical structure as no structure, shallow structure, and deep structure. The prediction is that visual performance should improve when the scene is hierarchically organized owing to feature sharing. The second approach is to manipulate the distance in a hierarchy. In a hierarchical graph, the distance between any two nodes is typically defined as the shortest traversal distance, which can be computed by various algorithms (Bondy & Murty, 1976). The prediction is that the contextual effect is stronger from objects closer to the target in the tree structure. The third approach is to manipulate the direction in a hierarchy. A hierarchy is directed in nature. In a directed tree structure (such as an and-or graph or other causal graphical models), each node has well-defined parent nodes and children nodes. A parent–child relationship is clearly asymmetric. The prediction is that for a pair of parent-child objects, the contextual effect is also asymmetric, depending on whether the target is the parent or the child.

The current study tests the three above-mentioned predictions based on depth, distance, and direction manipulations in dynamic displays by combining psychophysics and computational modeling. (Demonstrations of the visual stimuli and the computational modeling of the motion tree can be found at https://github.com/coreknowledge2016/hierarchical-motion.)

General Method

Here we introduce the general method for all the experiments reported in the current study. It has the following components: the hierarchical motion tree for generating the dynamic display; the position inference task for measuring the contextual effect; the experimental designs for manipulating depth, distance, and direction in the motion hierarchy; and the computational models of the psychophysics.

Hierarchical Motion Tree

Dynamic displays are generated by a hierarchical motion tree model (Gershman et al., 2016). It has two major processes. The first one is to generate a hierarchical tree structure with the nested-Chinese-Restaurant process (Blei, Griffiths, & Jordan, 2010) (See demonstration at https://youtu.be/pBX3joM_JPo.) Initially each object is assigned to the root node of the tree. In each iteration, an object can stop at the current node or move down to the next level of the tree. If it moves down, it can either follow an existing branch or create a new branch of the tree. The second process is to create a motion vector for each node in the tree. The motion directions are updated every 80 ms. At each update, the motion vectors are generated by a Gaussian process (Rasmussen & Williams, 2006) that are independent of previous motions directions. Critically the movement of each object is not determined by only a single motion vector. Instead, its motion direction is computed by composing all the motion vectors along its path to the root node. (See demonstration at https://youtu.be/rGgNBHuT2RY.) A schematic illustration of this motion tree model is shown in Figure 1. Formal descriptions of the nest-Chinese Restaurant process (nCRP) and Gaussian process are provided in the Appendix.

Psychophysics: The Position Inference Task

To measure the effects of the hierarchical representation, the current study adopts a position inference task, which requires an observer to infer the position of a moving target after it disappears for a while (Yin et al., 2016). Across trials, the relationship between the target and nontarget objects is systematically manipulated. The contextual effect is measured by how these objects influence the precision of inferring the target's position. Data collected from this task support both psychophysical analyses and computational modeling.

A schematic illustration of the task is shown in Figure 2 (see demonstration at https://youtu.be/IFgncqst8L8). Each trial starts with a motion display involving multiple objects moving for 4 s (Complete Observation). Subsequently, the motions continue for 2 s, but one randomly selected object (the target) will become invisible (Partial Observation). The target is indistinguishable from nontargets until it becomes invisible. The task is to infer the target's position after it is moved invisibly for 2 s. In response, an observer needs to move the mouse to the inferred position and then click. The inference error is defined as the distance between the inferred position and the target's actual position.

Across the experiments, we manipulated the relationship between the target and nontarget(s) by varying both the tree structure

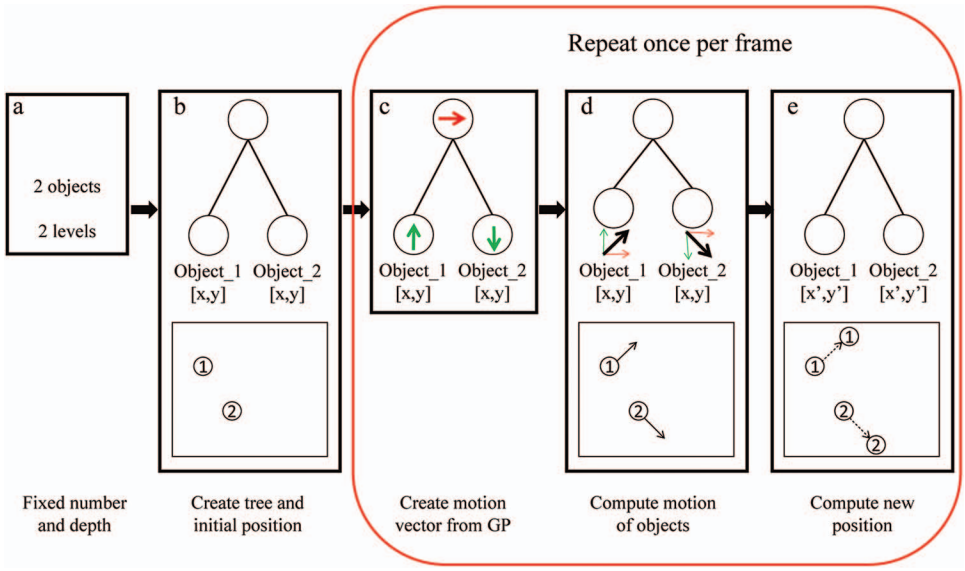

Figure 1. Schematic illustration of the hierarchical motion tree model. (a) The model starts from a fixed number of objects and depths. (b) The objects are assigned to their initial positions and terminal nodes of a tree generated by the nest-Chinese Restaurant process. (c) For each node in the tree, a motion vector is generated by a Gaussian process. (d) The motion of each object is computed by composing motion vectors along its path to the root node. (e) The dynamic displays are created by objects moving with their composed motions. See the online article for the color version of this figure.

(Experiment 1) and their relative positions in the tree (Experiments 2 and 3). There were five different trajectories for each type of tree structure.[2] Each of the five trajectories were rotated by 0°, 90°, 180°, and 270° about the center of the screen to (a) generate 20 trajectories in total and (b) control potential confounding owing to the objects' absolute spatial positions. The distance from the participants to the monitor was approximately 70 cm. All stimuli were presented on a black screen (size, 36.6° × 27.6°; refreshing rate, 100 Hz). The size of the motion display was 25° × 25°. Each object was a solid disk having a diameter of 1° and RGB values of 255, 255, and 255. All the settings of these experiments were based on a previous study (Yin et al., 2016). The speed of motion was set as 0.12°/frame, and the directions were changed by 7.05°/frame on average. There was no nonreported pilot experiment. The participants in all the experiments had normal or corrected-to-normal visual acuity. None of them reported knowing the actual aim of the experiment in a postexperiment questionnaire. The procedures were approved by the Research Ethics Board of Zhejiang University and the granting agency. The psychophysical experiments were written in MATLAB using Psychtoolbox 3.0 (Brainard, 1997).

Computational Model: Hierarchy Versus Contingence

The current study also constructs computational models to reverse engineer the human mind. The position inference task can be naturally formalized by a learning-inference framework. During the Complete Observation stage, a model (M) learns the values of latent variables (θ) underlying the objects' movements (D), indicated by $p(\theta \mid D, M)$. For example, the latent variables of the motion tree model are the tree structure and the motion vector associated with each node in the tree. The model and its learned values together form a representation of the observed movements. During the Partial Observation stage, the model infers the movements of the target (D_{target}), based on both learned latent values and the nontargets' motions ($D_{nontargets}$), indicated by $p(D_{target} \mid D_{nontargets}, \theta, M)$. This becomes a classical modeling problem of inferring the value of missing data (the target's motion) given a learned model and partially observed data (nontargets' motions). In each frame, the model infers the movement of the target and integrates the movements over time to obtain the target's position. From the perspective of human cognition, this inference process engages both a visual representation in memory (consolidated during the Complete Observation stage) and online perceptual inputs (from the Partial Observation stage).

The current study compares the Hierarchy model against a Contingence model that simply exploits the correlations among objects' motions. The Hierarchy model is the same model as that used for generating the motion display, except that it is inversed by Bayesian inference: Given a dynamic display, the model infers which tree structure can best explain the observed display (see demonstration at https://youtu.be/nwfx8gMMxNY). It then inferred the target position given inputs from the Partial Observation stage. The inference process followed the maximum likelihood principle, which did not require additional free parameters. In contrast, the Contingence model captures the correlations among objects' motions, but it does not assume any hierarchical structure. It treats objects' joint movements as samples from a multivariate Gaussian, whose mean and covariance matrix are estimated by

[2] Straightforward processes are applied to transfer each trajectory in the same part of the display around the fixation (25° × 25° square region) with smooth motion at the same average speed (0.12°/frame).

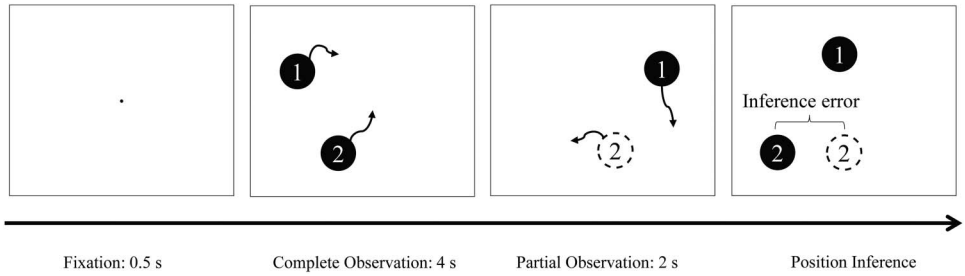

Figure 2. Schematic illustration of the position inference task. After fixation for 0.5 s, the target and nontarget(s) move for 4 s. The motions continue for 2 s while the target becomes invisible. The inference error is measured by the distance between an observer's inferred position and the target's actual position.

data from the Complete Observation stage. The motion contingences among objects are formalized by the covariance matrix of the Gaussian. The inference process also followed the maximum likelihood principle. Detailed inference algorithms of these two models are provided in the Appendix.

It is worth noting that the current study does not focus on the overall performance of the two above-mentioned models. Because the dynamic displays are generated as a hierarchical tree, presumably, the Hierarchy model will outperform the Contingence model in terms of performance. Therefore, the current study focuses on whether the patterns of a model's performance match the patterns of the human results across the experiments.

Experiment 1: Hierarchical Effects Across Tree Structures

A variety of structures can be derived from a display with only two objects. Here we explored the effects of depth, distance, and direction of a hierarchy in a display with two objects. The hierarchy depth was manipulated by assigning the two objects either to a single node in a tree with depth-1 (the Depth-1 condition, Figure 3a, see demonstration at https://youtu.be/0LpF4LNqQfk) or to two nodes in a tree with depth-2 (Depth-2 condition, Figures 3b and c). The hierarchy distance was manipulated by further varying the structure of a depth-2 tree. In the Single-branch condition (Figure 3b, see demonstration at https://youtu.be/I-Dhc1GSJBI), both objects were assigned to a two-level tree with a single branch. In the Diff-branch condition (Figure 3c, see demonstration at https://youtu.be/W6R6iplEPn8), the two objects were assigned to two different branches that share a parent. As a result, the two objects had Distance-1 in the Single-branch condition and Distance-2 in the Diff-branch condition. The hierarchy direction was manipulated by zooming into the Same-branch condition and controlling which object, that is, the parent (Object 1) or the child (Object 2), was the target. An Independent (Figure 3d, see demonstration at https://youtu.be/XvB3E3CYsmY) condition was considered a baseline, in which the two objects were assigned to two independent nodes that did not form a tree structure.

Participants, Stimuli, and Materials

Sixteen college students (seven males, 21–25 years of age) from Zhejiang University participated in Experiment 1 for a financial reward. The stimuli and materials were identical to those introduced in the *General Method* section; each participant performed 40 trials per condition, resulting in a total of 160 trials. Trials from all the conditions were presented in a randomized order. The experiment was divided into four blocks at intervals of 2 min.

Human Results

The inference errors from three different depth conditions are shown in Figure 4a. The main effect of depth is significant, $F(2, 30) = 107.82$, $p < .001$, $\eta_p^2 = 0.878$. Further direct

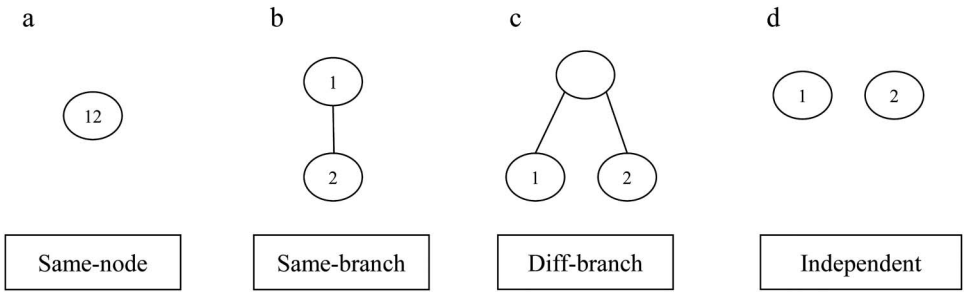

Figure 3. Illustration of trees with different depths, distances, and directions in Experiment 1. (a) Same-node condition, in which two objects share the same motion direction but have different initial positions and additive noise. (b) Same-branch condition, in which Object 1 is the parent of Object 2 in the same branch of the tree. (c) Diff-branch condition, in which the two objects share the same parent but are assigned to two different branches. (d) Independent condition, in which the two objects move independently without forming a tree structure.

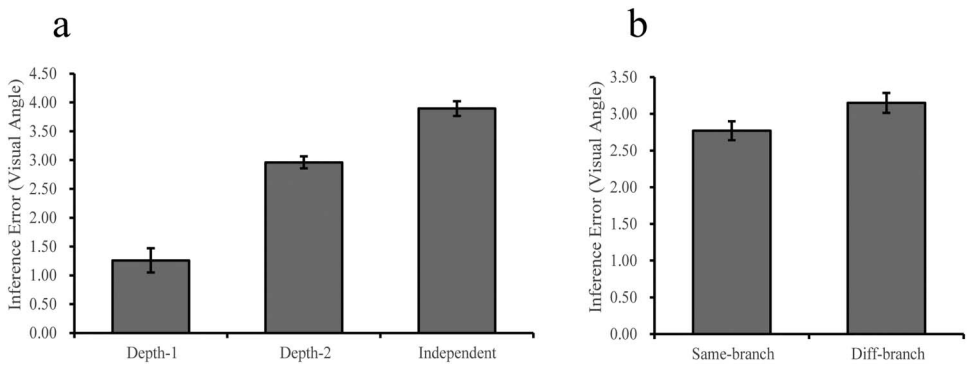

Figure 4. Results of Experiment 1. (a) Inference error as a function of depth in the tree. (b) Inference error as a function of distance in the tree.

comparisons (with Bonferroni correction) show that the differences between any two conditions are all significant (see Table 1).

To explore the effect of distance, we compared the results from the Same-branch (Distance-1) and Diff-branch (Distance-2) conditions. A paired *t* test showed a significantly lower inference error in the Same-branch condition (2.77°) than in the Diff-branch condition (3.15°), $t(15) = 2.325$, $p = .035$, $d = 0.609$. The results are shown in Figure 4b.

To explore the direction effect, we further split the Same-branch condition according to whether the target was a parent or a child in the tree (see Figure 5). A direct *t* test showed a higher error for inferring the parent given the child (2.92°) compared with inferring the child given the parent (2.62°), $t(15) = 2.357$, $p = .032$, $d = 0.589$. Moreover, 13 of 16 participants showed this result pattern (Figure 5b). This effect is not due to low-level motion properties, such as variance of speed, $t(119) = 0.723$, $p > .250$, and acceleration, $t(119) = 0.670$, $p > .250$. This parent–child asymmetric effect was replicated in an additional experiment with a new group of participants.[3]

Modeling Results

The results of the Hierarchy and Contingence models are shown in Figure 6. The results of the Hierarchy model were consistent with the human results, showing the effects of depth, $F(2, 30) = 2094.862$, $p < .001$, $\eta_p^2 = 0.993$, distance (Same-branch vs. Diff branch: $t(15) = 16.035$, $p < .001$, $d = 4.009$), and direction (parent vs. child: $t(15) = 6.447$, $p < .001$, $d = 1.612$). In contrast, the results of the Contingence model showed only the effect of depth, $F(2, 30) = 212.546$, $p < .001$, $\eta_p^2 = 0.934$; the distance effect was opposite to that of the human results, $t(15) = 5.322$, $p < .001$, $d = 1.331$, whereas there was no direction effect, $t(15) = 1.099$, $p = .289$, $d = 0.275$. These results collectively demonstrate that the Hierarchy model explains human results better than the Contingence model.

Table 1
Detailed Comparisons of Three Depths in Experiment 1

Tree types	Same-node	Same-branch	Independent
Same-node	—	$t = 9.133$ $p < .001$ $d = 2.283$	$t = 12.267$ $p < .001$ $d = 3.067$
Same-branch		—	$t = 6.868$ $p < .001$ $d = 1.717$
Independent			—

Discussion

The psychophysical and modeling results of this experiment showed that the visual system encodes the overall structure of the visual display, even when the explicit task is to infer the movements of individual objects. The structure can be efficiently represented by a hierarchical tree, varying from overlapping nodes to independent nodes. The distance between nodes within a structure is also important, with stronger contextual supports from objects closer to the target in the tree. The structure effect is also directed because it is easier to infer a child given a parent, compared with inferring a parent given a child.

Experiment 2: Hierarchical Effects Within a Single-Tree Structure

This experiment investigated the hierarchical nature of the contextual effect by using a tree structure with three levels. On the one hand, a deep tree structure may yield more robust hierarchical effects; on the other hand, a complex display is more difficult to control, with a much larger trial-by-trial variance that can make the hierarchical effects insignificant.

In contrast to Experiment 1, here the tree structure was fixed and only the identity of the target node within the tree was manipulated. As shown in Figure 7a (see demonstration at https://youtu.be/JPn5W5WcVXk), Objects 1 and 2 were assigned to two sibling nodes that not only shared a parent node but also were the grandchildren of the root node. Object 3 was assigned to a child of the root node on a different branch; it was an aunt or uncle of Objects 1 and 2. This three-level tree allowed us to examine the effects of hierarchy in more complex displays. Across the trials, the identity of the target (Object 1, 2, or 3) was manipulated. An Independent

[3] The replication results showed a similar pattern: the inference error of the parent was larger than that of the child, $t(15) = 2.719$, $p = .016$, $d = 0.679$.

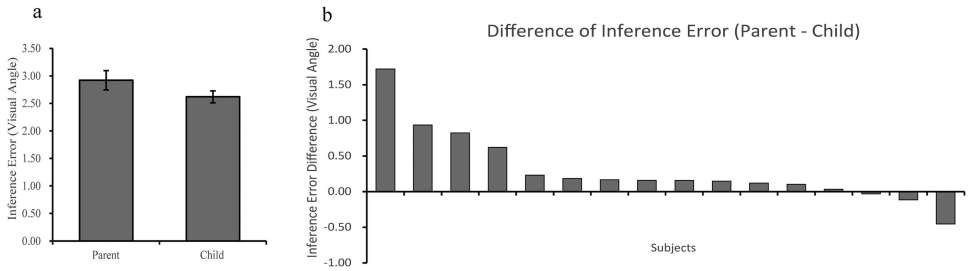

Figure 5. (a) Asymmetric effect from the Same-branch condition; the inference error of the parent (Object 1) is higher than that of the child (Object 2). (b) Individual subject data of the asymmetric effect. The Y-axis represents inference error of the parent–Inference Error of the child.

condition was introduced as a baseline, in which the three objects moved independently without forming a tree structure.

According to the Hierarchy model, when Object 3 is the target (Target-3 condition), the performance should be worse than that in the Target-1 and Target-2 conditions because Object 3 is (a) farther away from Objects 1 and 2 in the tree (the distance effect) and (b) an ancestor node of Objects 1 and 2 (the direction effect). There should be no significant difference between the Target-1 and Target-2 conditions because their nodes are interchangeable in the tree. The prediction of the Contingence model was less clear, owing to the complex correlations that could be generated by this tree structure.

Participants, Stimuli, and Materials

Sixteen other college students (five males, 20–26 years of age) from Zhejiang University participated in Experiment 2. The identity (1, 2, or 3) of the target object was systematically manipulated. There were 40 trials for each of Targets 1, 2, and 3 as well as the Independent condition, and the trial orders were randomized.

Human Results

The results of Experiment 2 are shown in Figure 7b. One-way ANOVA showed a significant effect of object identity, $F(3, 45) =$

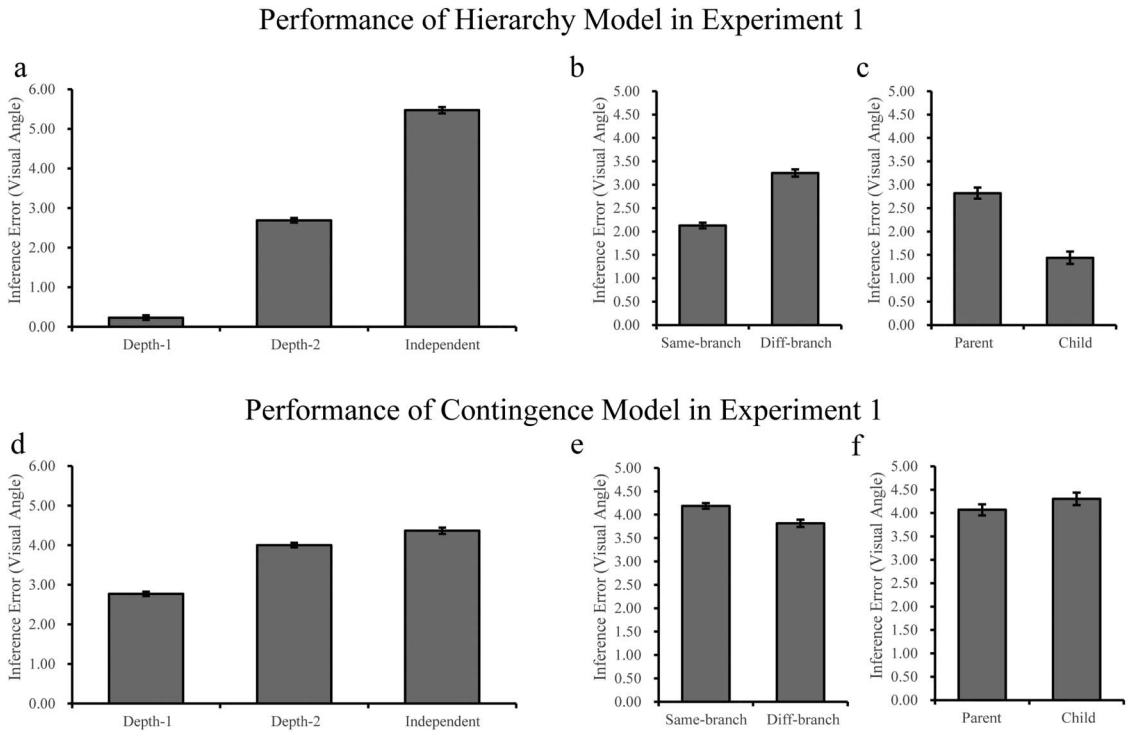

Figure 6. Performance of the two types of computational models in Experiment 1. The results of the Hierarchy model were consistent with the human results in terms of depth (a), distance (b), and direction (c). The Contingence model showed different results (d–f).

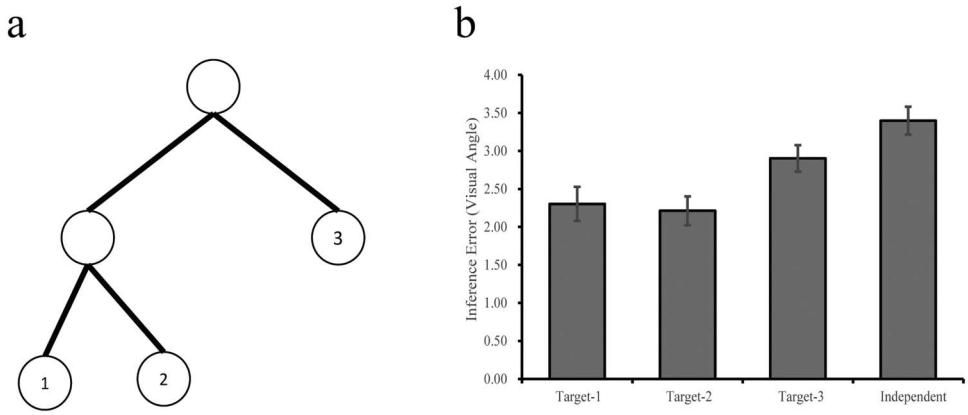

Figure 7. Tree structure and results of Experiment 2. (a) A three-level motion tree formed by three objects. (b) The performance was different in Experiment 2. The inference error showed no significant difference between Target-1 and Target-2, and it was lower than that in the other two conditions. The inference error in Target-3 was lower than that in independent-node.

40.45, $p < .001$, $\eta_p^2 = 0.729$. Further comparison (Bonferroni correction) indicated that the differences of all the comparisons were significant, except Target-1 versus Target-2 (see Table 2). These results collectively demonstrated that the contextual effect was modulated by the hierarchical tree structure.

Modeling Results

The results of the Hierarchy and Contingence models are shown in Figure 8. The results of the Hierarchy model were consistent with the human results; the inference of Target-3 was significantly worse than that of Target-1, $t(15) = 2.653$, $p = .050$, $d = 0.663$, and Target-2 $t(15) = 2.735$, $p = .046$, $d = 0.684$, whereas the difference between the Target-1 and Target-2 conditions was not significant, $t(15) = 0.116$, $p > .250$, $d = 0.029$.

In contrast, the results of the Contingence model diverged from the human results drastically; the inference of Target-2 was significantly worse than that of Target-1, $t(15) = 3.284$, $p = .015$, $d = 0.821$, and Target-3, $t(15) = 3.992$, $p = .004$, $d = 0.998$, whereas the difference between the Target-1 and Target-3 conditions was not significant, $t(15) = 0.389$, $p > .250$, $d = 0.097$. This pattern of results is unexpected because Object-1 and Object-2 are sibling nodes in the tree, whose identities are interchangeable—there should be no systematic difference between these two conditions at all. One interpretation is that the correlations among objects are different in these two conditions owing to the random process of sampling the motion trajectories. This interpretation was confirmed by analyzing the correlations between Object-1 and Object-2 in the trajectories of these two conditions. It was found that in the Target-1 and Target-2 conditions, the average correlation was 0.727 and 0.507, respectively. Thus, this detailed analysis explained why the Contingence model produced a higher inference error in the Target-2 condition.

Table 2
Detailed Comparisons (With Bonferroni Correction) of Experiment 2

Conditions	Target-1	Target-2	Target-3	Independent
Target-1	—	$t = 1.237$ $p > .250$ $d = .309$	$t = 3.563$ $p = .017$ $d = .891$	$t = 8.864$ $p < .001$ $d = 2.216$
Target-2		—	$t = 5.164$ $p = .001$ $d = 1.291$	$t = 10.994$ $p < .001$ $d = 2.749$
Target-3			—	$t = 4.362$ $p = .003$ $d = 1.091$
Independent				—

Discussion

This experiment explored the contextual effect with three objects that form a more complex tree structure. The results showed that the Hierarchy model can explain the patterns of the human results despite larger variance of the trajectories owing to the motion tree's complex stochastic process. In contrast, the Contingence model failed to capture the patterns of the human results. In addition, it is not robust, that is, it is sensitive to artifacts of the

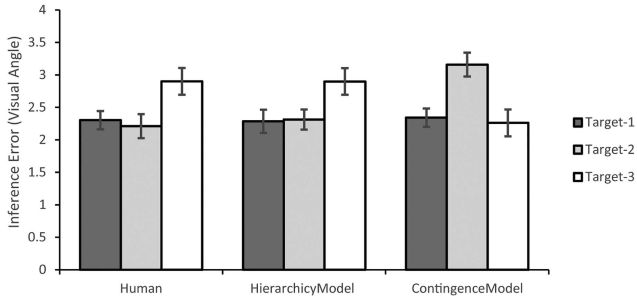

Figure 8. Comparison between human results and model results of Experiment 2. The results of the Hierarchy model were consistent with the human results. In contrast, the results of the Contingence model diverged from the human results.

trajectories, whereas humans and the Hierarchy model are immune to such artifacts.

Experiment 3: Correlated Motion Without Composition

The hierarchical tree structure composes motion vectors along a tree branch to generate the motion of each object. The goal of this experiment was to demonstrate the importance of hierarchically composed motions by removing them from the displays while maintaining the correlations among the objects' movements. To achieve this goal, we adopted the Mirror-Image operation introduced for distinguishing Chasing from correlated motion (Gao, Newman, & Scholl, 2009). In this operation, an object is replaced with its mirror image over the center of the display (see Figure 9a). Here we reused the trajectories from Experiment 2 and applied the Mirror-Image operation on one object in each trial. A motion tree is defined by a structure of shared motion. The Mirror-Image operation can isolate the mirrored object from such a structure because it now does not share any motion with other objects. Furthermore, because the mirrored motion is completely determined by the original motion, the strength of correlations among the three objects is intact. Across the trials, the target was fixed as Object-1, whereas the identity of the mirrored object was systematically manipulated, yielding different trajectories for Mirror-1, Mirror-2, and Mirror-3 conditions. A No-Mirror condition was introduced as a baseline, in which no object was mirrored.

According to the Contingence hypothesis, human performance should not be influenced by the mirror-image operation; the performances of all the mirror conditions should be similar to those of Experiment 2 and the No-Mirror condition herein. However, according to the Hierarchy hypothesis, the identity of the mirrored object is critical. Mirror-1 should cause the largest drop in performance because the target object is mirrored and completely isolated from the tree; Mirror-3 should cause the smallest drop in performance because Object-3 is far away from the target in the tree and contributes little to the inference task.

Participants, Stimuli, and Materials

Sixteen new college students (eight males, 20–25years of age) from Zhejiang University participated in this experiment. There were 40 trials for each of the four mirror conditions.

Results and Discussion

The results of the three mirrored conditions and the No-Mirror condition are shown in Figure 9b and Table 3. The inferences of Mirror-1 and Mirror-2, but not the inference of Mirror-3, were significantly worse than that of the No-Mirror condition (see Table 3). These results collectively demonstrate that the contextual effect is based on a hierarchical representation of composed motions rather than correlated motions.

General Discussion

Visual objects are not encoded in isolation but represented as part of a global context in both static images (Brady & Alvarez, 2015; Gmeindl et al., 2011; Jiang et al., 2000; Olson & Marshuetz, 2005; Woodman, Vecera, & Luck, 2003) and dynamic displays (Sun et al., 2015; Yantis, 1992). Revealing the nature of the visual context is critical to understanding how higher-level vision represents complex scenes, in which objects interact with one another, and each object may have complicated internal structures (e.g., Zhou et al., 2016). Using dynamic displays, the current study explored the nature of the scene context by combining psychophysics and computational modeling. The results collectively demonstrated that the contextual effect is based on a hierarchical representation formed by multiple moving objects.

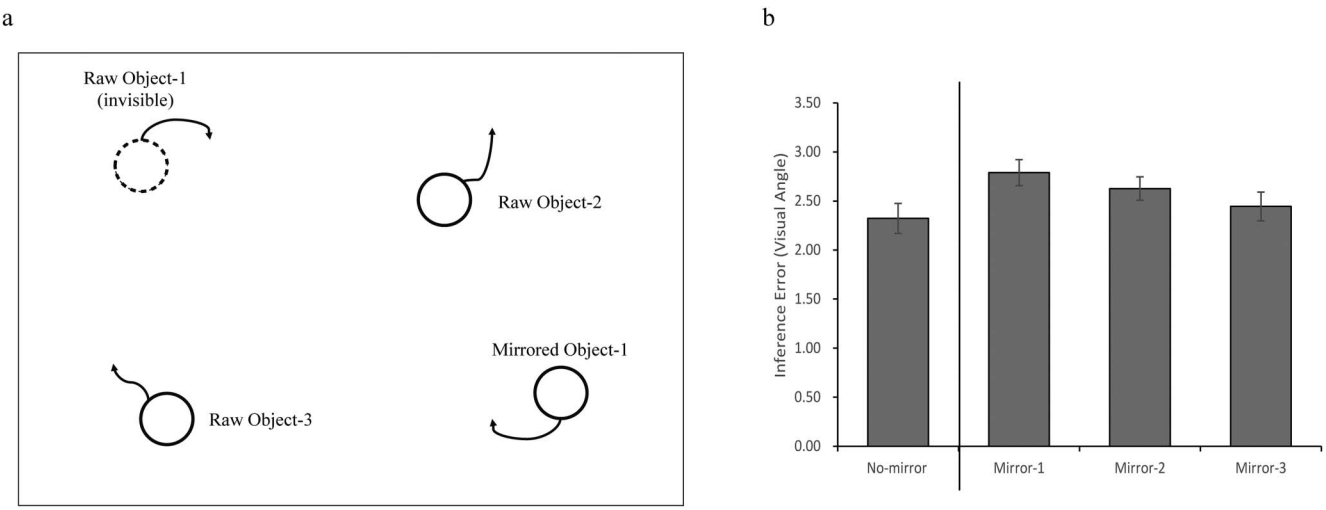

Figure 9. Mirror process and results of Experiment 3. (a) Illustration of the Mirror-Image operation, in which Object-1 was replaced by its mirror image over the center of the display. (b) Performance of position inference as a function of the mirror conditions.

Table 3
Comparing Mirror Conditions With the No-Mirror Condition of Experiment 3

Conditions	Mirror-1	Mirror-2	Mirror-3
No-Mirror	$t = 3.443$ $p = .004$ $d = .861$	$t = 2.126$ $p = .050$ $d = .532$	$t = 1.485$ $p = .158$ $d = .371$

Quantifying the Global Context

A directed acyclic tree is a typical hierarchical structure. Inspired by the key properties that define a tree structure, we systematically introduced experimental manipulations that altered the hierarchical-hood of the display. These manipulations yielded several interesting psychophysical results. First, the depth manipulation showed that the contextual effect is structured. Displays with a hierarchical tree structure form a stronger context than those without a tree structure. In addition, human performance drops as the tree depth increased. Second, the distance manipulation showed that the contextual effect is degreed. Objects closer to the target in the tree provide stronger contextual support. Third, the direction manipulation showed that the contextual effect is asymmetric and directed. The contextual support provided by a parent node to a child node is stronger than that provided by a child node to a parent node.

Based on these psychophysical results, two computational models, namely the Hierarchy model and the Contingence model, were constructed and tested. The Hierarchy model is a Bayesian model that infers the mostly likely hidden tree structure given the display. It replicated all the above-mentioned patterns of human results. The Contingence model is a Gaussian model that exploits only the correlations among objects' motions. This model was unable to capture the degreed and directed patterns of the human results. In addition, it tends to overfit the observed data, influenced by artificial correlations generated by chance.

Hierarchical Representation and Limited Resources

As we have argued in the introduction, in addition to revealing the capacity constraints of working memory, it is equally important to explore the types of representation that can be efficiently operated by the limited resources. As a critical mechanism of visual cognition, working memory manipulates representations to effectively connect the past, present, and future in a dynamic visual world. A hierarchical representation is a promising candidate for these operations owing to the following factors.

First, it is a sparse and structured summary of the past, which improves the efficiency of storying visual information with limited resources. As shown in both Experiments 1 and 2, the inference error was much lower for motions with a latent tree structure than for independent motions. This result is consistent with previous studies that have shown improved memory performance of structured static images (e.g., Brady & Tenenbaum, 2013).

Second, a hierarchical representation enables efficient interaction between the past and the present of a visual display. As we have proposed in a previous study, a major function of working memory is to enable an active interaction between previous information and current perceptual inputs (Gao, Gao, Li, Sun, & Shen, 2011). The position inference task in the current study is consistent with this proposal because it demands a dynamic interaction between stored motion representation (from the Complete Observation stage) and online perception (from the Partial Observation stage) to infer the target's current position. Using a Hierarchy model, such an interaction can be naturally formalized by a learning-inference framework. By using the rules of probability, the Hierarchy model infers the values of current hidden variables given a previously learned tree structure and the latest partial observation.

Third, a hierarchical representation can support prediction (or imagination) of the future. Prediction of an object's future movement was not tested in the current study with human observers. However, the model constructed by explaining human results can be easily extended for such prediction owing to the intrinsic nature of the Hierarchy model: it is a generative model that reveals the causal processes of how motion displays are created. To make a prediction, one needs only to draw a sampled trajectory by running the model forward. Indeed, the ability to synthesize data through forward simulation is a unique advantage of generative/causal models. Predicting the future of a visual scene through forward simulation has been explored by Battaglia, Hamrick, and Tenenbaum (2013). In addition, understanding visual perception by drawing samples with a model has also been explored (e.g., Gao, Baker, & Tenenbaum, 2016; Srivastava & Vul, 2015). In summary, a hierarchical representation provides a promising solution to how limited working memory can store and manipulate information of a complicated visual scene. The causal nature of the Hierarchy model is further discussed in the next section.

Hierarchy as a Type of Causal Structure

One of the most well-known laws of statistics is that correlation does not imply causality. The Contingence model in the current study follows this law strictly by only using objects' correlated motions. On the other hand, inferring causality from correlation has been formalized (e.g., Pearl, 2009), and it plays a major role in modern artificial intelligence (Russell & Norvig, 2003). The Hierarchy model is a type of causal model because it infers the latent hierarchical processes that generate the observed display. The causality is defined by the direction of an edge that is oriented from a parent node (the cause) to a child node (the effect). From this perspective, the perception of hierarchical motion can be treated as a special case of the perception of causality.

Noncausal, undirected models also play important roles in modern artificial intelligence, and they can have a deep structure (LeCun et al., 2015). To demonstrate the causal nature of human motion perception, it is critical to discover a directed effect, in which the parent and the child are not processed symmetrically. The current study revealed such a directed effect based on psychophysics: it is easier to infer a child's positions by observing its parent than vice versa. The modeling results further showed that such an effect is not simply a bias or preference but reflects the computational nature of the inference process. Using a hierarchical causal model, it is computationally easier to infer the child's position, given its parent's movements. Of equal importance is the fact that the Contingence model did not show any directed effect—this is expected because it is an undirected model based on corre-

lation that is always symmetric. In summary, these results collectively demonstrate that, when representing the movements of multiple objects, human vision does represent a causal structure beyond correlated movements.

A hierarchical causal representation of the visual context is consistent with previous findings that indicate that the human mind is adapted to perceive causality, including the classical launching effect (e.g., Michotte, 1963), force (Wolff, 2007), intuitive physics in a more general sense (e.g., Battaglia, Hamrick, & Tenenbaum, 2013), and even the causal history of an image (Chen & Scholl, 2016).

Psychophysics of Hierarchy-Hood: From Surface Features to Latent Causes

Over the past decades, studies on object-based vision have been remarkably successful, partly because of a range of psychophysical paradigms for manipulating object-hood, including object previewing (Kahneman, Treisman, & Gibbs, 1992), object versus space (Egly, Driver, & Rafal, 1994), object-based memory (Luck & Vogel, 1997), and object persistence (e.g., Flombaum & Scholl, 2006). These are excellent examples of psychophysics, which show how the representation in the human mind changes as a function of the visual display's physical properties, including closed contour, connectivity, and spatial and temporal continuity.

To reveal the hierarchical nature of visual representation, there should be a set of psychophysical manipulations as well. The current study introduces three such manipulations: depth, distance, and direction. Interestingly, none of these manipulations directly maps to the physical properties of the objects' motions. In Experiment 1, there were always just two objects whose physical properties, such as speed and acceleration, were held constant. All the manipulations targeted the properties of a latent tree structure. They caused the objects to move in a certain way but could not themselves be reduced to certain features of the objects' motions. This was expected from a modeling perspective because they were the latent variables of the hierarchical model. In other words, they were unobservable by definition.

These manipulations of latent variables are certainly within the realm of psychophysics. However, they require a broader understanding of what physics means in vision science. Physics in the real world is not always visible but involves various latent factors, including forces, fields, energy, and even dark matter. As summarized in a TED talk, the history of physics shows that ". . . the closer we look at anything, the more it disappears" (Lloyd, 2009). Psychophysics may follow the same trend because vision has its own dark matter (Xie, Todorovic, & Zhu, 2013). When modeling visual perception with a deep hierarchy, visual surface features are mostly likely mapped to the bottom of the hierarchy, serving as the observed data. Moving upward along the hierarchy, the variables will gradually diverge from the surface features and become latent causes of the visual scene. We suggest that explicitly recognizing vision as a deep hierarchical process with massive latent structures is important. It can guide researchers to further expand the scope of psychophysics by creating novel manipulations targeting those latent structures beyond visible features.

Conclusion

The current study demonstrated how a global contextual effect can be quantitatively manipulated and modeled by a hierarchical representation. The results showed that the contextual effect is based on a latent tree representation that explains the objects' interactions. In addition to the tree representation investigated herein, many recent studies have explored various latent structures and causes of vision, including events (Strickland & Scholl, 2015; Zacks & Swallow, 2007), forces (Wolff, 2007), intuitive physics (Battaglia, Hamrick, & Tenenbaum, 2013), and animacy (Gao, Newman, & Scholl, 2009). Together these studies suggest that human vision is a type of intelligence, which sees what are in the displays by recovering why and how they are generated.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences, USA, 110,* 18327–18332. http://dx.doi.org/10.1073/pnas.1306572110

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115–147. http://dx.doi.org/10.1037/0033-295X.94.2.115

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the Association for Computing Machinery, 57,* 1–30. http://dx.doi.org/10.1145/1667053.1667056

Bondy, J. A., & Murty, U. S. R. (1976). *Graph theory with applications* (Vol. 290). London, UK: Macmillan. http://dx.doi.org/10.1007/978-1-349-03521-2

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision, 15,* 6. http://dx.doi.org/10.1167/15.15.6

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

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

Caicedo, J. C., & Lazebnik, S. (2015). Active object localization with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile*, 2488–2496.

Cain, S., Dobkins, K., & Vul, E. (2016). Texture properties bias ensemble size judgments. *Journal of Vision, 16,* 54. http://dx.doi.org/10.1167/16.12.54

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage, 12,* 314–325. http://dx.doi.org/10.1006/nimg.2000.0612

Chen, Y. C., & Scholl, B. J. (2016). The perception of history: Seeing causal history in static shapes induces illusory motion perception. *Psychological Science, 27,* 923–930. http://dx.doi.org/10.1177/0956797616628525

Chi, Z., & Geman, S. (1998). Estimation of probabilistic context-free grammars. *Computational Linguistics, 24,* 299–305.

Chomsky, N. (1964). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 109–127. http://dx.doi.org/10.1037/0096-1523.21.1.109

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113,* 501–517. http://dx.doi.org/10.1037/0096-3445.113.4.501

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

Flombaum, J. I., & Scholl, B. J. (2006). A temporal same-object advantage in the tunnel effect: Facilitated change detection for persisting objects. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 840–853. http://dx.doi.org/10.1037/0096-1523.32.4.840

Flombaum, J. I., Scholl, B. J., & Pylyshyn, Z. W. (2008). Attentional resources in visual tracking through occlusion: The high-beams effect. *Cognition, 107,* 904–931. http://dx.doi.org/10.1016/j.cognition.2007.12.015

Froyen, V., Feldman, J., & Singh, M. (2015). Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological Review, 122,* 575–597. http://dx.doi.org/10.1037/a0039540

Gao, T., Baker, C., & Tenenbaum, J. (2016). *Cognitive architecture of intentionality perception: Animacy, attention and memory*. Manuscript under review.

Gao, T., Gao, Z., Li, J., Sun, Z., & Shen, M. (2011). The perceptual root of object-based storage: An interactive model of perception and visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 1803–1823. http://dx.doi.org/10.1037/a0025637

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology, 59,* 154–179. http://dx.doi.org/10.1016/j.cogpsych.2009.03.001

Gershman, S. J., Tenenbaum, J. B., & Jäkel, F. (2016). Discovering hierarchical motion structure. *Vision Research, 126,* 232–241. http://dx.doi.org/10.1016/j.visres.2015.03.004

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

Gorniak, P., & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research, 21,* 429–470.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57,* 243–259. http://dx.doi.org/10.2307/1416950

Humboldt, W. von. (1836). *On language: The diversity of human language-structure and its influence on the mental development of mankind*. Cambridge, England, UK: Cambridge University Press.

Jackendoff, R. (1996). The architecture of the linguistic-spatial interface. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space. Language, speech, and communication* (pp. 1–30). Cambridge, MA: MIT Press.

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

Johansson, G. (1950). *Configurations in event perception*. Uppsala, Sweden: Almqvist & Wiksell.

Johnson, S. G., & Keil, F. C. (2014). Causal inference and the hierarchical structure of experience. *Journal of Experimental Psychology: General, 143,* 2223–2241. http://dx.doi.org/10.1037/a0038192

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature, 290,* 91–97. http://dx.doi.org/10.1038/290091a0

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24,* 175–219. http://dx.doi.org/10.1016/0010-0285(92)90007-O

Karpathy, A., & Li, F. F. (2015). Deep visual-semantic alignments for generating image descriptions. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 3128–3137.

Kiley Hamlin, J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science, 16,* 209–226. http://dx.doi.org/10.1111/desc.12017

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521,* 436–444. http://dx.doi.org/10.1038/nature14539

Lew, T. F., & Vul, E. (2015). Ensemble clustering in visual working memory biases location memories and reduces the Weber noise of relative positions. *Journal of Vision, 15,* 10. http://dx.doi.org/10.1167/15.4.10

Lloyd, J. (2009, September). *An inventory of the invisible* [Video file]. Retrieved from http://www.ted.com/talks/john_lloyd_inventories_the_invisible

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA, Chicago: Henry Holt & Company.

Michotte, A. (1946/English transl. 1963). *The perception of causality*. London, UK: Basic Books.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

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

Pearl, J. (2009). *Causality*. Cambridge, England, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511803161

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

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance, 18,* 849–860. http://dx.doi.org/10.1037/0096-1523.18.3.849

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53,* 245–277. http://dx.doi.org/10.1146/annurev.psych.53.100901.135125

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach*. Upper Saddle River, NJ: Prentice Hall.

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

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive psychology, 38,* 259–290.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1,* 261–267. http://dx.doi.org/10.1016/S1364-6613(97)01080-2

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81,* 214–241. http://dx.doi.org/10.1037/h0036351

Snowden, R. J., & Verstraten, F. A. J. (1999). Motion transparency: Making models of motion perception transparent. *Trends in Cognitive Sciences, 3,* 369–377. http://dx.doi.org/10.1016/S1364-6613(99)01381-9

Srivastava, N., & Vul, E. (2015). Perceptual and cognitive limitations interact in multiple object tracking. *Journal of Vision, 15,* 882. http://dx.doi.org/10.1167/15.12.882

Strickland, B., & Scholl, B. J. (2015). Visual perception involves event-type representations: The case of containment versus occlusion. *Journal*

of Experimental Psychology: General, 144, 570–580. http://dx.doi.org/10.1037/a0037750

Sun, Z., Huang, Y., Yu, W., Zhang, M., Shui, R., & Gao, T. (2015). How to break the configuration of moving objects? Geometric invariance in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 41,* 1247–1259. http://dx.doi.org/10.1037/xhp0000086

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12,* 49–100. http://dx.doi.org/10.1207/s15516709cog1201_2

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97–136. http://dx.doi.org/10.1016/0010-0285(80)90005-5

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences, 11,* 58–64. http://dx.doi.org/10.1016/j.tics.2006.11.009

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136,* 82–111. http://dx.doi.org/10.1037/0096-3445.136.1.82

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

Xie, D., Todorovic, S., & Zhu, S. C. (2013). Inferring 'Dark Matter' and 'Dark Energy' from Videos. *Proceedings - IEEE International Conference on Computer Vision, 2013,* 2224–2231.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning,* 2048–2057.

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

Yin, J., Xu, H., Ding, X., Liang, J., Shui, R., & Shen, M. (2016). Social constraints from an observer's perspective: Coordinated actions make an agent's position more predictable. *Cognition, 151,* 10–17. http://dx.doi.org/10.1016/j.cognition.2016.02.009

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science, 16,* 80–84. http://dx.doi.org/10.1111/j.1467-8721.2007.00480.x

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127,* 3–21. http://dx.doi.org/10.1037/0033-2909.127.1.3

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

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2016). *Semantic understanding of scenes through the ADE20K dataset*. Retrieved from https://arxiv.org/abs/1608.05442

Zhu, S. C. (1999). Embedding gestalt laws in Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21,* 1170–1187. http://dx.doi.org/10.1109/34.809110

Zhu, S. C., & Mumford, D. (2007). *A stochastic grammar of images*. Hanover, MA: Now Publishers Inc.

Appendix

Correlation and Causal Models of Dynamic Displays

Contingence Model

This model assumes the joint movements of all objects are sampled from a multivariance Gaussian distribution, which is parameterized by its mean (μ) and covariance (Σ). Contingence of objects' movements is defined as the covariance matrix of the Gaussian. During the Complete Observation stage, the parameters of the model is estimated by the maximum likelihood principle:

$$\mu, \sum = \text{argmax}_{\mu,\Sigma}\, p\left(D_{1:t} \mid u, \sum\right)$$

During the Partial Observation stage, the model computes the distribution of the target's motion direction, conditioning on the movements of Nontargets. It is well known that this conditional distribution is Gaussian as well, whose mean and variance can be analytically solved (Murphy, 2012). The mean of the joint distribution u can be partitioned into two parts ($\mu 1$, $\mu 2$), which represent the mean of the target and nontargets respectively. Similarly, the covariance matrix is partitioned as well. To simplify the formulation of the conditional Gaussian, we define Λ as the precision matrix, which is the inverse of the covariance matrix.

$$\mu = [\mu_1, \mu_2],$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},\ \Lambda = \Sigma^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

Let x_1 and x_2 be the motion directions of the target and nontargets, respectively. The conditional distribution of the target's motion is given by:

$$p(x_1 \mid x_2) = N\left(x_1 \mid \mu_{1|2}, \sum_{1|2}\right),$$

where $\Sigma_{1|2} = {\Lambda_{11}}^{-1}$

$$\mu_{1|2} = \sum_{1|2} \times (\Lambda_{11}\mu_1 - \Lambda_{12} \times (x_2 - \mu_2))$$

Hierarchy Model

This model assumes that the trajectories are generated from a hierarchy tree. The critical parameters are the ones defining the structure of tree. According to Baye's rule, it can be inferred given observed trajectories. (Gershman et al., 2016).

$$\mathrm{P}(\boldsymbol{c}, \boldsymbol{d}|\boldsymbol{s}) \propto \mathrm{P}(\boldsymbol{c})\mathrm{P}(\boldsymbol{d})\mathrm{P}(\boldsymbol{s}|\boldsymbol{c}, \boldsymbol{d})$$

Here the parameter c denotes the nodes associated all the objects, and parameter d denotes the depth of all the objects in the tree. The parameter s denotes the trajectories of all the objects. In the following subsections, we describe the computations of P(c), P(d), and P(s|c, d), respectively.

(*Appendix continues*)

Computing P(*c*)

Objects are assigned to a tree structure by the nest Chinese Restaurant Process, which defines the probability of assigning object n to node j at depth d. j can be an existing node at depth d, or a new node, which will open a new branch in the tree. Node assignment is a sequential process, with the assignment of node j depends on all the previous node assignments. The parameter M_j denotes the number of objects that have already been assigned at node j at the current depth. The assignment of the next object to node j is proportional to M_j. Intuitively, this process encourages an object to join a popular node. The parameter γ controls the bias of creating a new node. Here we set $\gamma = 1$. D is the maximal tree depth. Node assignment for each object proceeds at each depth recursively until arriving at the maximal depth D. The result of this process is that each object owns a node at each level of the tree.

The nodes associated with each object is represented as a vector $\boldsymbol{C}_n = [C_{n1}, \ldots, C_{nD}]$, in which n represents the object id and d represents the depth. Cn forms a path from the root to a leaf.

$$P(C_{nd} = j|\boldsymbol{C}_{1:n-1}) = \begin{cases} \dfrac{M_j}{n-1+\gamma}, j \leq J \\ \dfrac{\gamma}{n-1+\gamma}, j = J+1 \end{cases}$$

Computing P(*d*)

In the tree created above, each object has a corresponding node in each depth. The current process truncates this tree by removing nodes below certain depths so that objects can terminate in the tree at different depth levels. Unlike the node assignment, this depth process is not a sequential process. Instead, the depth levels of all objects are sampled simultaneously from a distribution, defined as a Markov random field. d_n denotes the depth level of object n. It encourages (a) an object to have the same depth as the other objects and (b) an object to have shallow, instead of deep depth. The probability is defined as bellow:

$$P(\boldsymbol{d}) \propto exp\left\{\alpha \sum_{m=1}^{N} \sum_{n>m}^{N} \Pi[d_m = d_n] - \rho \sum_{n=1}^{N} d_n\right\}$$

where $\Pi[]$ is an indicator function returns 1 if its argument is true and 0 otherwise. The parameter α controls the penalty for assigning objects to different depths. The parameter ρ controls the penalty for deeper assignments. Here we set $\alpha = 1$ and $\rho = 0.1$ following previous research.

Computing P(*s*|*c*,*d*)

This section describes how to generate smooth trajectories given *c, d*. Unlike most object-based processes, the focus here is not object but node: It computes the motion direction of each node. The motion direction of an object is then composed by adding all the motion directions of the nodes assigned to that object.

In fact, at each frame, each node does not get a single motion direction, but a field of motion direction (f), which is sampled from a Gaussian process (Rasmussen & Williams, 2006). In this field, nearby spatial locations have similar motion directions. $f(s, t)$ returns the motion direction at spatial location s at time t. However, by adjusting the parameters of this Gaussian process, a field can have very little variance, forcing spatial locations of the entire field to have the same motion direction in practice. After sampling a motion field for each node, an object's motion direction is computed by extracting and adding motions directions from nodes associated with that object at different depth levels. At time $t + 1$, the new positon position (s_{t+1}) of object n is computed by adding the motion direction to the old positons (s_t).

$$s_{t+1} = s_t + \sum_{d=1}^{d_n} f_{c_n,d}(s_t, t)$$

Received December 4, 2016
Revision received February 1, 2017
Accepted March 13, 2017 ■