

Hierarchical structure is employed by humans during visual motion perception

Johannes Bill^{a,b,1}, Hrag Pailian^b, Samuel J. Gershman^{b,c,2}, and Jan Drugowitsch^{a,c,1,2}

^aDepartment of Neurobiology, Harvard Medical School, Boston, MA 02115; ^bDepartment of Psychology, Harvard University, Cambridge, MA 02138; and ^cCenter for Brain Science, Harvard University, Cambridge, MA 02138

Edited by Alan A. Stocker, University of Pennsylvania, Philadelphia, PA, and accepted by Editorial Board Member Renée Baillargeon August 12, 2020 (received for review May 13, 2020)

In the real world, complex dynamic scenes often arise from the composition of simpler parts. The visual system exploits this structure by hierarchically decomposing dynamic scenes: When we see a person walking on a train or an animal running in a herd, we recognize the individual's movement as nested within a reference frame that is, itself, moving. Despite its ubiquity, surprisingly little is understood about the computations underlying hierarchical motion perception. To address this gap, we developed a class of stimuli that grant tight control over statistical relations among object velocities in dynamic scenes. We first demonstrate that structured motion stimuli benefit human multiple object tracking performance. Computational analysis revealed that the performance gain is best explained by human participants making use of motion relations during tracking. A second experiment, using a motion prediction task, reinforced this conclusion and provided fine-grained information about how the visual system flexibly exploits motion structure.

motion perception | hierarchical structure | multiple object tracking | generative models | Bayesian inference

The visual scenes our brains perceive in everyday life are filled with complex dynamics. Information hitting the retina changes not only with every motion in the scene but also with every head movement and saccade. Noise, occlusions, and ambiguities further make visual information inherently unreliable. In order to maintain stable, coherent percepts in the face of complex and unreliable inputs, our brains exploit the spatially and temporally structured nature of the environment (1).

Motion structure refers to statistical relations of velocities. One form of structure common in natural scenes is motion grouping: When a rigid object is set in motion, all of its visual features move coherently. This structure allows us to infer the existence of objects based on the coherent motion of features—the Gestalt grouping cue known as *common fate* (2). Grouping based on common fate has been shown to influence our ability to track objects (3–6), to search displays (7), and to store information in short-term memory (8).

However, the strict definition of common fate is too brittle to accommodate natural scenes in which visual features do not move together rigidly and yet are still grouped together. In some cases, this is because we perceive objects as deforming nonrigidly. In other cases, we perceive objects that are hierarchically structured (9–11): The parts of an object move rigidly relative to a reference frame (the object), which itself could be a rigidly moving part of another object, and so on. For example, we perceive the motion of hands relative to the motion of arms, and the motion of arms relative to the motion of the torso (Fig. 1 A and B). The entire body may be moving relative to another reference frame (a train or escalator). The perception of hierarchically organized motion suggests a powerful "divide-and-conquer" strategy for parsing complex dynamic scenes.

Recent work has formalized the representation and discovery of hierarchical motion structures (10). However, we still know relatively little about whether and how the visual system exploits

this structure for visual tasks like object tracking. We address this question in two experiments, one testing multiple object tracking (MOT) (12) and one testing multiple object trajectory prediction. A key innovation of these experiments is the use of tightly controlled stimuli that isolate the signatures of hierarchical structure from other forms of motion structure. We show that a Bayesian observer model, when equipped with the appropriate structural representation, can provide insight into the mental computations underlying human behavior across a variety of motion structures. In the MOT task, we find that improvement of tracking performance for structured stimuli cannot be explained without exploiting structure knowledge during inference, indicating that humans make use of motion structure knowledge when perceiving dynamic scenes. In the prediction task, we uncover the employed motion structure knowledge of our participants, finding that humans can flexibly recruit different motion relations for scene parsing, including deep motion hierarchies which nest groups of objects within another moving reference frame.

Results

The study was preregistered (13) prior to data collection. All preregistered analyses are presented in the main text; any additional analyses are labeled as such.

Significance

The structured organization of motion in visual scenes is highly informative for our everyday perception: We recognize people by the way they walk, track objects through occlusion, or predict hazardous situations from the traffic flow. It is, however, unclear how our minds tame the overwhelmingly complex stream of dynamic information received by the retina to form such stable percepts. We argue that an observer can exploit a "divide-and-conquer" strategy where complex motion relations are broken down into compositions of simpler motions. Evidence for hierarchical decomposition comes from multiple object tracking and prediction experiments in which humans are able to exploit motion structure knowledge to improve performance. Our results can guide neuroscience experiments on the neural representation of structure.

Author contributions: J.B., H.P., S.J.G., and J.D. designed the research; J.B., S.J.G., and J.D. developed the theory; J.B. and H.P. conducted the experiments; J.B. analyzed the data; and J.B., H.P., S.J.G., and J.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. A.A.S. is a guest editor invited by the Editorial Board.

Published under the PNAS license.

¹To whom correspondence may be addressed. Email: johannes.bill@hms.harvard.edu or jan_drugowitsch@hms.harvard.edu.

²S.J.G. and J.D. contributed equally to this work.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/ doi:10.1073/pnas.2008961117/-/DCSupplemental.

First published September 16, 2020.

Representation of Motion Structure. In this section, we introduce a representation of hierarchical motion structure. To motivate this representation, consider a flock of birds. The velocities of individual animals in the flock are naturally described by the sum of a global motion component, which is shared across all birds, and an individual component for each animal relative to the flock's velocity. Only the aggregate motion is observed; the underlying motion components are latent variables that help us organize our perception.

To formalize the idea of shared, latent motion components, we assume all velocities in a visual scene to be driven by motion sources (nodes in the graph in Fig. 1C) that either represent observable objects (filled orange nodes) or are latent (unfilled gray nodes). A motion source inherits velocity from a parent source when connected by an edge, thus supporting tree-like hierarchies. Motion sources can be shared by multiple objects, and the total velocity of an object is the sum of all inherited motions. In addition to the graph connectivity, each motion source has a motion strength λ that determines the source's contribution strength to the speed of dependent observable objects. The simple "global motion" motif in Fig. 1C might, for example, to a first order, describe the motion structure underlying a flock of birds, where larger motion strengths are illustrated by larger vertical distances (curly braces) between the motion sources.

The separation of motion composition (graph connectivity) and motion strength (vertical node location) gives rise to a flexible, modular representation of motion structure, covering many real-world scenes such as independent motion, clustered motion, or deep motion hierarchies. The graph in Fig. 1 D, Top, for example, describes a faster orange and a slower green cluster of otherwise independently moving composed objects (e.g., two noninteracting flocks of birds). A further global motion component, such as an observer moving his head, would introduce another motion source at the root of the tree (Fig. 1 D, Bottom). This illustrates how our motion representation can be used in a modular fashion to describe deep nested motion hierarchies. In fact, any tree-like motion structure built in this way can be represented by a motion structure matrix L, accommodating both composition and strengths (see Materials and Methods), such that we will often refer to specific motion structures by their associated matrices L.

This modular representation enabled us to generate motionstructured visual stimuli, which isolated the computational role



Fig. 1. Modular representation of hierarchical motion structure. (*A*) Observed velocity components of a running human. The hands inherit motion from the arms, which inherit motion from the torso. (*B*) Corresponding nested hierarchy of motion relations. Observed velocity is the sum of local motion components. (C) Motion graph describing global motion with a strong (strength λ_{glo}) shared motion source (gray node) and weaker (strength λ_{ind}) individual motion sources (orange nodes). Here, the global motion source is not directly observed (i.e., latent) but introduces correlations in the motion of observable objects (orange). (*D*) two motion clusters (*Top*) can be embedded into a deep hierarchy by adding another latent motion source at the tree's root (*Bottom*). (*E*) Illustration of a stimulus of stochastically rotating dots with global motion structure. These are the class of stimuli used in the experiments.

of motion structure in dynamic visual scenes, and were thus particularly suited for psychophysics experiments. Specifically, we strove to design stimuli in which all object properties and statistical relations among objects were dominated by structure in velocities, while keeping other factors like individual object velocities or spatial structure uninformative. To do so, we generated the velocities of observable objects by random draws from a continuous time stochastic process [namely, a multivariate Ornstein-Uhlenbeck process (14); see Materials and Methods for details] that yielded smooth random trajectories with the desired statistical motion relations imposed by a chosen motion structure L. In this process, motion sources play the role of random forces which accelerate or decelerate any dependent object. For example, the latent global source in Fig. 1C would induce correlations among the velocities of all observed objects. The resulting trajectories are mathematically tractable and feature real-world properties, such as inertia and friction, without making any assumptions about specific trajectory realizations. To further remove any persistent spatial structure among the objects' locations, stimuli were positioned on a circle (illustrated in Fig. 1E; Movies S1–S3). This makes the locations of objects asymptotically independent. The mathematical tractability of the trajectory-generating process was crucial for precise stimulus control. Knowing the exact joint probability distribution across object velocities allowed us to vary velocity correlations induced by motion structure while keeping the motion velocity statistics of individual objects unchanged, thus making structure in velocities the dominant feature in the presented scenes.

Motion Structure Improves Visual Tracking Performance. We first asked whether motion structure knowledge benefits humans in an MOT task (12). Evidence exists that structure among objects, such as grouping (15), symmetry (6), or global translation (4), impacts tracking performance, but hierarchical structure has not been investigated before.

In the MOT task, K = 7 dots rotated about a circle with three dots being initially marked as targets. After a few seconds, all dots changed to identical appearance while dot motion continued for another 6 s. After that, the dot motion stopped, and participants had to reidentify the initially marked targets. We tested 20 participants on four different blocked motion conditions of 30 trials each, with motion graphs shown in Fig. 2A (targets marked by *; Movies S1 and S2). Independent (IND) motion, the standard MOT task, served as a baseline. In the global (GLO) condition, dot motion is composed of a dominating shared stochastic motion source, as well as small individual per-dot motion sources to break emergent spatial patterns. In counterrotating (CNT) motion, the latent motion source affects dot velocities in opposite directions: Similar to gears, the force accelerating the dots k = 1, 2, 3 in one direction accelerates dots k = 4, 5, 6, 7 (three of which are targets) in the reverse direction. The last motion condition is a counterrotating deep hierarchy (CDH) in which two three-dot groups are driven by a counterrotating motion source that shares a global motion source with the seventh dot. This deep nested structure generates distinguishable velocity patterns that cannot be approximated by any shallow (at most one latent source layer) motion structure. In the CDH condition, we tested two different target sets (CDH_1 and CDH₂) to probe set-specific effects in tracking performance. We measured participant performance as the average number of correctly identified targets within each condition. Since the aim of the experiment was to test how humans can exploit motion structure knowledge rather than learn it, at the beginning of each motion block, participants were explicitly presented with a diagram and three demonstration trials of the motion structure that would be presented in the upcoming block. Overall average stimulus speed was titrated on a per-participant basis to



Fig. 2. Use of motion structure knowledge during MOT. (A) Tested motion conditions included IND, GLO, CNT, and CDH motion; *s mark targets. Two different target sets were tested for CDH motion. (B) Average performance (number of correctly identified targets) on different motion conditions by human participants, the Bayesian computational observer model using the correct motion structure, and the Bayesian computational observer model disregarding motion relations (IND prior). Using motion structure during inference is required to explain human performance gains on motion-structured stimuli. (C) The observer model consists of a Kalman filter with motion structure prior L (Left) and a mental assignment of dot identities (Center). Perceptual and neural noise can lead to ambiguous assignments and, ultimately, errors in the reported target set (Right). (D) Fraction of trials with zero, one, two, and three dots correct (red, orange, light green, and dark green, respectively) for human participants, the observer model with the correct motion prior, and the model with an IND prior.

approximately reach a performance of 2.15 [midpoint between chance level $(3 \times 3/7)$ and perfect (3)] correctly identified dots in the IND condition. This per-participant speed level was subsequently maintained for the rest of the experiment. During the following data collection, the marginal motion statistics of individual dots were then held constant for each participant across all conditions L and dots k such that conditions only differed in their dot velocity correlations, as determined by L. A separate IND condition block (see left-most bar in Fig. 2B) that was not used to adjust the stimulus speed confirmed the validity of the adjustment and marks the reference for performance changes in motion-structured stimulus conditions. The performance on the four conditions with motion structure (GLO, CNT, CDH_1 , and CDH_2) is shown in Fig. 2 B, Left next to the IND reference performance. The introduction of structured motion significantly impacted dot tracking performance ($p \approx$ 1.03×10^{-14} , Greenhouse–Geisser corrected repeated measures ANOVA), resulting in a significant performance boost in all motion conditions (one-sided paired t tests, $p \approx 1.2 \times 10^{-8}$, 2.4×10^{-9} , 7.9×10^{-3} , and 2.6×10^{-4} ; see *SI Appendix*, Fig. S1 for additional pairwise comparison of all conditions). In

conclusion, motion structure clearly improved human tracking performance.

Use of Structure Drives MOT Performance Gain. The observed boost in tracking performance for structured stimuli could simply be a byproduct of dot velocity correlations, that is, an intrinsic stimulus property, rather than the result of using motion structure knowledge by the observer. For instance, the paths of dots with positively correlated velocities might cross less frequently during a trial, making their confusion less likely (16). To distinguish the contribution of employed motion structure knowledge from stimulus-intrinsic factors, we extended a Bayesian MOT observer model by Vul et al. (17) to incorporate motion structure priors L. The deliberately simple model (illustrated in Fig. 2C) includes only the core components required to perform the MOT task, making it a minimalistic, normative model of motionstructured MOT (see Materials and Methods for details). Similar to Vul et al. (17), visible dot locations \mathbf{x}_t in individual video frames are subject to perceptual and neural noise σ_{obs} while being integrated with mental estimates of location z_t and velocity \mathbf{v}_t via a multidimensional Kalman filter (Fig. 2 C, Left). To capture structured motion, we extended this Kalman filter to incorporate the motion structure matrix L as a Bayesian prior distribution of how dot velocities are expected to evolve over time. A GLO motion prior, for example, would favor positively correlated dot trajectories. The Kalman filter alone, however, assumes that every observed dot location is labeled by the dot's identity, making dot confusions, and therefore imperfect dot tracking, impossible. To produce such confusion, we needed to additionally model the mental assignment of dot identities to the visually identical looking dots on the screen (Fig. 2 C, Center; cf. ref. 17), which is a manifestation of the correspondence problem (18, 19). Errors in this mental assignment can lead to imperfect reidentification of the target dots at the end of an MOT trial, and arise when dots come close (20) or even cross: Uncertainty in internal location estimates \boldsymbol{z}_t can render multiple mental assignments possible, as exemplified in Fig. 2 C, Right. Since the Bayesian model is derived solely from computational considerations for solving MOT and disregards any psychophysical constraints, for example, velocity-dependent observation noise (21), we refer to it in what follows as the *computational observer* model.

We presented the computational observer model with the exact same trials that had been shown to human participants. The single free parameter of the model, σ_{obs} , was adjusted to match average human performance in the baseline IND motion condition, and was subsequently fixed to predict performance in all other conditions. This resulted in a qualitative match between human performance and the observer model's predictions (Fig. 2 *B*, *Left* vs. *Center*). Besides the motion structure dependence, the observer also correctly predicted the increased performance for the nonsplit target set (CDH₂) over the split set (CDH₁). For a discussion of the quantitative mismatch in the GLO condition, see *Probing Components of Human Object Tracking* and *SI Appendix*.

More importantly, the computational observer allowed us to distinguish whether the performance gain for structured motion stimuli was simply due to stimulus-intrinsic factors: If some motion structures were generally easier to track than others, the computational observer should be able to take advantage of these differences even without explicitly exploiting the structural knowledge in its prior. To test this, we repeated the above procedure with an observer that assumed an IND motion prior, which respects the correct marginal dot velocities, but disregards any knowledge on velocity correlations, even when motion structure was present in the stimuli. As Fig. 2 *B*, *Right* vs. *Left* reveals, this modification caused all motion structure dependent performance gains to vanish. Hence the observed

performance gain of human participants cannot be explained by stimulus-intrinsic factors alone. Instead, humans appear to make use of motion structure knowledge when performing the task.

The finding that the computational observer with correct motion prior explains large parts of human responses, while the motion structure-agnostic observer (IND prior) does not, is further supported by comparing the fraction of MOT trials with zero, one, two, or all three dots identified correctly (Fig. 2D, additional analysis). This visual impression is corroborated by preregistered statistical measures: While the cosine similarity between per-participant human and correct-prior observer model performance gains ("structured condition minus IND condition") is highly significant (angle between the performance gains $\triangleleft 42.0^\circ$; $p < 10^{-5}$, one-tailed test against H0: Positive cosine similarity is a random effect), the one between perparticipant human and IND-prior observer model performance gains is not ($\triangleleft 105.3^\circ$; $p \approx 0.99$).

Probing Components of Human Object Tracking. The modularity of the Bayesian observer affords insights into the differential role of different psychophysical constraints impacting MOT performance. Specifically, we examined two alternative observer models which add/remove components to/from the computational observer (Fig. 3*A*; additional analysis; for details, see *Materials and Methods*).

1) The momentum-free observer lacks the concept of inertia, that is, that objects usually change their trajectory smoothly. This model mirrors the perceptual assumptions from established work (10) where motion relations give rise to only correlated location displacements of objects between stimulus frames. We further removed any observation noise, $\sigma_{obs} = 0$, to isolate the effect of inertia for MOT.

2) For the *Weber's law observer*, in contrast, we augmented the computational observer with two key psychophysical constraints modeled in ref. 17, namely, speed-dependent observation noise (21) and decision noise (22, 23) in dot assignments.

The MOT performance of both models is shown in Fig. 3Bwhen either the correct prior or an IND prior are assumed. We find that the momentum-free observer's performance drops to chance level in the IND condition, confirming that the use of inertia is crucial for performing MOT in the circular task design: Whenever two dots come close in one frame, they can only be tracked to the next frame by exploiting the fact that trajectories continue smoothly. This insight complements previous studies of two-dimensional motion tracking (17, 24) for which very close dot proximity was less likely and hence the use of trajectory extrapolation less critical. In fact, this might have been the reason why previous studies found little evidence of use of inertia in MOT. However, when structure is present, as seen in the CNT and CDH conditions, even a momentum-free observer can benefit from motion structure knowledge for partial performance recovery. Notably, such motion prior-dependent performance gain is not present when tracking GLO stimuli, pointing to a peculiarity of this stimulus condition (see SI Appendix, Supplementary Text and Fig. S4 for a computational discussion based on the dot proximity autocorrelation function). The Weber's law observer, on the other hand, benefits from structured priors in all conditions and predicts human responses better than the computational observer ($\triangleleft 35.5^{\circ}$; see also SI Appendix, Fig. S2 for trial speed-dependent performance, and see SI Appendix, Fig. S3 for exact choice sets), underscoring the contribution of known psychophysical constraints to human visual tracking.

A second direction supported by the Bayesian observer is a detailed analysis of how well the different models predict the exact set of dots chosen by human participants (note that these are additional analyses; *SI Appendix*). For probing the puta-



Fig. 3. Functional components underlying human MOT. (A) Alternative observer models with components added to or removed from the computational observer. The "momentum-free observer" lacks the concept of inertia. The "Weber's law observer" adds two known psychophysical constraints: velocity-dependent observation noise and stochastic decision noise. (B) Average MOT performance on different motion conditions for the alternative observer models when either the correct motion structure or an IND prior are assumed (like Fig. 2B). Employing trajectory extrapolation is computationally indispensable on the circle, as highlighted by the momentum-free observer's performance dropping to chance in the IND stimulus condition. Adding psychophysical constraints, in contrast, leads to an improved match with human performance. (C) Bayesian model comparison of employed motion structure priors based on the exact pertrial choice sets. Shown are log-likelihood ratios of human choice sets under different putative motion priors L relative to the true structure underlying the stimulus. Negative values indicate that the participant's behavior was better explained by the correct motion prior. Human participants make use of motion structure knowledge during tracking, presumably employing approximately correctly structured priors. Yet, the limited information provided by the discrete response sets prevented insight into the exact structural features used. Each dot represents one participant per stimulus condition and putative motion prior. Horizontal lines show mean log-likelihood ratios across participants (values in parentheses). Asterisks indicate significance of paired t tests (p < 0.05, 0.01, 10^{-3} , and 10^{-4} for one, two, three, and four asterisks, respectively; ns, not significant).

tive use of motion prior L, we used simulations to estimate the likelihood of choosing a particular set of dots in each trial of the experiment for observer models with different motion priors. We then asked how well these simulated likelihoods match

the participants' choices. Results of this analysis are shown in Fig. 3C for the computational observer, and in *SI Appendix*, Fig. S3 for the momentum-free and Weber's law observers. This analysis confirms the above result that some motion structure knowledge is employed by human observers, in that an IND prior explains human responses significantly worse than the respective correct motion priors. Besides the structures used in the stimulus conditions, the comparison includes two additional priors: clustered motion (CLU) and clusters-plus-independent (CLI) motion. These priors approximate the motion relations of the CDH condition without considering its nested nature. This analysis suggests that human participants typically employ the correct or close-to-correct motion priors, even for nested structures—yet, the effect size is often too small to make decisive statements.

In conclusion, the MOT task revealed that humans make use of motion structure knowledge during visual tasks. However, the limited information provided by the participants' responses (three chosen dots per trial) did not allow us to identify which exact motion features were employed.

Revealing the Structure of Human Motion Priors. We therefore developed a second experiment, multiple object prediction (pre-registered), that grants a more fine-scaled insight into the cognitive machinery underlying motion perception.

Multiple object prediction uses partially observed dynamic visual scenes to test human perception in face of uncertainty. As before, in each trial, seven dots rotated stochastically about a circle (Fig. 4A). After 5 s, however, two target dots (the larger green and red dots) became invisible, while the motion of the other dots remained visible. After another 1.5 s, the scene froze, at which point participants had to predict the location of both target dots with a computer mouse. We used highly volatile motion-structured stimuli with quickly changing dot velocities, such that the use of motion structure information was indispensable for making predictions. We further colored the dots according to their role in the motion structure, to prevent dot confusion. In contrast to previous single-dot prediction tasks (11), the rationale for simultaneously predicting two dot locations is that the covariance pattern of errors conveys additional information about the motion structure assumptions of our participants. For example, if global motion was assumed, we would expect the participants to jointly overestimate or underestimate the red and green targets' final locations.

We tested 12 participants on three motion conditions each: GLO, CLU, and CDH (motion graphs in Fig. 4*E*; Movie S3), with 100 trials per condition. As before, participants were briefly trained on all motion structures and were informed about the specific structure underlying each trial. Participants reported the task to be challenging, but performed reasonably well (see Fig. 4*B* for the GLO condition, and *SI Appendix*, Fig. S8 for CLU and CDH; additional analysis).

In order to identify the structure of the motion prior employed by each participant in each stimulus condition, we formulated a Bayesian observer model of human responses in the prediction task. This model was the same as our MOT model, but without the dot confusion component, and with location observations assumed to be practically noise-free ($\sigma_{obs} = 0$, except for the invisible dots for which no further observations were provided). The reduced computational complexity of the prediction task rendered exact inference possible: at the end of each trial, motion-structured Kalman filters with different putative motion structure priors \boldsymbol{L} predicted the statistically optimal, most likely location of the target dots based on the motion of the other dots, and the uncertainty in this prediction. Building on ref. 22, we linked this prediction to the locations reported by the participants by assuming additional response variability that scaled with prediction uncertainty (correlated across the two target dots), motor noise (uncorrelated across dots), and the possibility to confuse the red and green dots when reporting their locations (see *Materials and Methods* for details). The first two variability sources were modeled as Gaussians, owing to the Gaussianity of observed human response error distributions (*SI Appendix*, Fig. S6; additional analysis), whereas the third variability source was implemented by a small swapping probability. Overall, this led to three model parameters that we fit by maximum-likelihood separately for each participant, condition, and putative motion structure prior. Applied to simulated behavior with the same trials, we found that this procedure was able to correctly recover the motion structure underlying simulated responses (*SI Appendix*, Fig. S7).

To identify the most likely motion structure prior employed by the participants, we compared, across seven structurally different putative motion priors, how well the model can explain the participants' trial-by-trial predictions. For the GLO motion condition, the model predictions are shown in Fig. 4 C and D when assuming a GLO prior (Fig. 4C) or an IND prior (Fig. 4D) for all 100 trials of one representative participant (see SI Appendix, Fig. S9 for all participants and stimulus conditions). The visually apparent better match of the GLO over the IND prior is statistically quantified by the log-likelihood ratio of the human responses under the two motion priors (Fig. 4E, highlighted data point in the top left matrix cell). The resulting log-likelihood ratios for all motion conditions and priors are shown in Fig. 4E, with putative motion priors sorted from simpler to more complex structures (preregistered; see SI Appendix, Fig. S10 for additional comparison across a larger range of motion structures, yielding similar results). Note that all compared models have the same number of free parameters. For the GLO motion condition (Fig. 4 E, Main Panel, top row) human responses are best explained either by the correct global motion prior or by more complex priors which would be similarly suited for solving the task (e.g., because they contain a global motion source). Motion priors without a global motion source, such as IND, CNT, and CLI, resulted in significantly worse model fits. For the CLU motion condition (Fig. 4E, second row), all alternative model priors led to worse model fits. Together, these results establish confidence in the applicability of the Bayesian observer model.

Our most complex motion condition, the CDH (Fig. 4 E, Main Panel, bottom row), was designed to require inference over both latent motion sources for correct prediction. In other words, none of the tested shallow (nonnested) motion priors can mimic the error covariance pattern expected under a CDH prior. GLO motion would miss the counterrotating component of the blue and green groups. CNT motion would miss the global motion component. A CLI motion prior (Fig. 4 E, Top) could approximately predict the green dot's location, but lacks the global component for predicting the red dot. The only structures that could approximate a CDH covariance pattern are other deep nested hierarchies, such as a standard deep hierarchy (SDH) that employs a third latent motion source. The result of Bayesian model comparison (Fig. 4 E, Main Panel, bottom row) clearly favors deep hierarchical motion priors for explaining human predictions. Overall, our results strongly suggest that human observers are able to use motion structure, including hierarchically nested motion relations, when perceiving dynamic scenes.

Systematic and Stochastic Errors Underlie Human Suboptimality. Even though participants employed the qualitatively correct motion structure in the prediction task, they featured additional, suboptimal variability when compared to a statistically optimal observer (Fig. 4B). This variability could arise from consistent deterministic bias, indicating a systematic misintegration of the



Fig. 4. Revealing human motion priors in a multiple object prediction task. (*A*) Illustration of the stimuli. The highlighted green and red dots disappeared after 5 s. Participants had to predict their location at the end of the trial. Dots were color-coded to indicate their role in the motion structure. Here, a GLO stimulus condition is illustrated. (*B*) Mean-squared prediction error of the green and red dots for all participants in the GLO condition. Due to task-inherent uncertainty, even perfect inference, as given by the mean values of a Kalman filter with the correct motion structure prior, will exhibit nonzero predictions of an observer model "Bayes opt."). Humans do not reach this optimal accuracy, but perform better than chance. (*C*) Human responses (dots) relative to the predictions of an observer model with correct GLO prior, in all 100 trials for the participant highlighted in *B* and *E*. The fitted observer model (ellipses) predicts human responses well (ellipses indicate 1, 2, 3 SD). (*D*) Same as *C*, but for an observer model assuming an IND motion prior. Neither the predicted locations nor the covariance in human prediction errors is captured by the model. (*E*) (*Left*) Tested motion conditions included GLO, CLU, and CDH motion. (*Top*) Putative motion priors tested for explaining human responses via a Bayesian observer model, ordered by their complexity. (*Main Panel*) Each cell shows per-participant log-likelihood model fit ratio for a particular motion prior, compared to the correct prior underlying the stimulus (indicated by gray background). Negative values indicate that the participant per stimulus condition and putative motion prior. Humans flexibly employed correctly structured motion priors. Each dot represents one participant per stimulus condition and putative motion prior (comparison between *C* and *D* highlighted in orange). Horizontal lines show mean log-likelihood ratios across participants (values in parentheses). Asterisks indicate significance of paired *t* tests (p <

stimuli, or stochastic fluctuations, indicating potentially noisy computations. To determine the contribution of each alternative to overall variability, we repeated each trial twice within each condition, supporting a bias-variance decomposition (23) of this variability (Fig. 5A). Specifically, if all of the variability was due to deterministic biases, participants should feature the exact same, systematic errors, $\Delta^{(1)} = \Delta^{(2)}$, in both trials. If, in contrast, all variability was purely stochastic, the observed errors should be uncorrelated across paired trials. This idea allowed us to quantify the ratio between bias and variance by one noise factor f_{noise} per participant and motion condition, with $f_{\text{noise}} = 0$ and $f_{\text{noise}} = 1$ denoting purely bias-driven and purely noise-driven observers, respectively (see Materials and Methods for details). We evaluated the noise factor separately for the green and red target dots, with the results shown in Fig. 5B. Note that these results are additional analyses. We found that human suboptimality is a combination of systematic and stochastic errors. The noise factors of both dots were, on average, equally strong (Fig. 5B, black \times). The positive correlation between red and green dot noise factors ($\rho = 0.51$, Pearson correlation) suggests a participant-dependent level of "noisiness." This impression is further supported by consistently higher or lower noise factors across participants (colors and filling in Fig. 5B; $p < 10^{-4}$, one-way Welch's ANOVA).

Discussion

We have shown that humans make use of motion structure knowledge during demanding perceptual tasks, and that they can flexibly employ structured motion priors for the task at hand. Beyond shallow motion motifs, such as global motion or clustered motion, we were able to show that humans can even use deep nested motion hierarchies in the prediction task. Key to revealing the covert motion priors of human participants was the design of analytically tractable stimuli that capitalize on motion relations. Analytical tractability facilitated the development of normative Bayesian observer models, our mathematical scalpel for dissecting the use of motion priors by the participants from stimulus-intrinsic contributions to visual motion perception.

Understanding motion perception as Bayesian inference has a rich scientific history (10, 11, 17, 25–29). Bayesian models of low-level perception (25–27) were extended to explain human MOT of independent dot movement (17, 29, 30). Our work extends this line of research to structured motion, revealing the use of structured motion priors during tracking. The



Fig. 5. Bias-variance decomposition for the prediction task. (A) Human prediction errors are assumed to be the sum of a systematic (bias) and a stochastic (noise) component. The relative contributions of each component can be estimated from repetitions of the same trial, leading to responses $\varphi^{(1)}$ and $\varphi^{(2)}$, and associated errors $\Delta^{(1)}$ and $\Delta^{(2)}$. (B) Noise factors for the green and the red dots, one marker per participant (color and filling) and motion condition (shape). A combination of systematic and stochastic errors underlies human suboptimality. Black cross and ellipsoids indicate mean and iso-density curves (1 to 3 SDs) of a bivariate Gaussian fitted to all noise factors.

modularity of the observer model further enabled us to study the role of specific computational components, such as employing only subtrees of a motion structure, as well as the role of specific perceptual components, such as noise at different stages or exploiting the concept of inertia during tracking. We did not consider further features underlying human performance, such as attention effects (15, 29, 31), tracking of subsets of dots, or not consistently employing all features of a motion structure, as these are more challenging to operationalize within our framework, and were not central to this study. A systematic deviation between human MOT performance and model predictions in the GLO condition finally points to additional temporal integration on intermediate time scales (~ 1 s; *SI Appendix*, Fig. S4), going beyond established explanations based on object distance (20, 32).

Our aim to resolve the fine structure of human motion priors has thus led us to the development of a probabilistic multiple object prediction task that augments single-object prediction (11) with velocity covariances within a tractable protocol. Including posterior covariance matrices in the observer model greatly enhanced the log-likelihood ratios in the prediction task beyond the resolution achievable in single-object prediction (SI Appendix, Fig. S11), an effect potentially related to selfconsistency biases (33). While we had originally developed our modular matrix representation L of hierarchical structure as an extension to the motion trees in ref. 10, the proposed decomposition into graph connectivity and motion strengths bears some resemblance to the semantic structure interpretation of singular value decomposition applied in ref. 34 to categorical data and to Pythagorean tree embeddings used in ref. 35 for describing language structure. Flexible structure representations are expected to foster cognitive science research on how humans infer the motion structure of a visual scene-potentially recruiting from a set of motion features-and how such motion features for modular combination could be learned in the first place.

Biologically, the importance of motion for visual scene perception is reflected in the tuning of cells in primate visual areas. Neurons in the middle temporal (MT) area are frequently tuned to the speed and direction of velocity within their receptive field (36), while downstream areas, like the dorsal medial superior temporal (MSTd) area, encode progressively richer motion primitives such as selective tuning to expansion, rotation, and spiraling (37, 38). This points to a feature repertoire that is tailored to behaviorally relevant stimuli such as radial expansion (the visual pattern on the retina of a forward-moving observer) or rotation (when tilting your head to the left/right). Little, however, is known about how such motion primitives are further recruited in neural circuits for high-level, hierarchical motion processing. Our motion structure representation is compatible with existing neural implementations of Bayesian sensory integration, such as the neural Bayesian filtering model of Beck et al. (39). In combination with physiology-based models of motion integration (40, 41), this could bring forth normative neural models of structured motion perception in higher visual areas, and guide experiments on the neural code along the visual motion pathway.

Materials and Methods

Motion Structure Matrix Representation. We describe the motion of K visible objects that are driven by M motion sources. Usually, M > K, since, besides shared latent sources, each object can feature individual motion. We represent how motion source m affects object k via a composition matrix of motion motifs $\mathbf{C} \in \mathbb{R}^{K \times M}$. If m drives k, we set $C_{km} = 1$; if m does not affect k, we set $C_{km} = 0$; for counterrotating motion, we set $C_{km} = +1$ and -1 for opposing directions. Thus, each column of \mathbf{C} encodes a motion motif (e.g., all "1"s for global motion), which is tied to motion source m. Each motion source has an associated real-valued motion strength $\lambda_m \ge 0$, and we define the motion structure matrix L as the product of composition and strengths, $L = \mathbf{C} \Lambda$ with $\Lambda = \text{diag}(\lambda_1, ..., \lambda_M)$, that is, strength λ_m scales the mth motion motif.

Motion-Structured Stimuli. The stochastic dynamics for location $z_t \in [0, 2\pi)^k$ and velocity $v_t \in \mathbb{R}^k$ are given by

$$\mathrm{d}\boldsymbol{v}_t = -\boldsymbol{v}_t/\tau \,\mathrm{d}t + \boldsymbol{L}\,\mathrm{d}\boldsymbol{W}_t \tag{1}$$

$$d\boldsymbol{z}_t = \boldsymbol{v}_t \, dt, \qquad [2]$$

with friction time constant $\tau \in \mathbb{R}^+$ and M independent Wiener processes in vector W_t . The values of τ and Λ used in each experiment are provided in *SI Appendix*. For simulations, we employed Euler–Maruyama integration to advance the dynamics, and we map $z_t \mapsto (z_t \mod 2\pi)$ after each step to keep locations on the circle. The stationary distributions under above dynamics are (see ref. 14)

$$\lim_{t \to \infty} p(\mathbf{v}) = \mathcal{N}(\mathbf{v}; \mathbf{0}, \Sigma_{\mathbf{v}}), \quad \lim_{t \to \infty} p(\mathbf{z}) = \prod_{k=1}^{K} \text{Uniform}(z_k; \mathbf{0}, 2\pi), \quad [3]$$

where $\Sigma_{\mathbf{v}} = (\tau/2) \mathcal{L}^{\mathsf{T}}$, and the expression for $p(\mathbf{z})$ assumes that each object has a nonvanishing independent motion component. Knowing the stationary velocity covariance $\Sigma_{\mathbf{v}}$ in closed form gives convenient control over the stimulus. For example, the marginal velocity distribution of the *k*th object is $\mathcal{N}\left(v_k; \mathbf{0}, \sigma_{v_k}^2\right)$ with $\sigma_{v_k}^2 = \sum_m \lambda_m^2 \cdot C_{km}^2$, which is simply the sum of all parent motion sources' squared strengths. In all experiments, we initialize the stimuli from a sample of their stationary distribution. Then, velocities follow a multivariate Gaussian at any time, and locations become asymptotically independent.

Experiment Details: MOT Task. Twenty Harvard University undergraduates (mean age = 20.6 y, SD = 2.01; 14 females) with normal or corrected-tonormal vision participated in the experiment for course credit. All participants provided informed consent at the beginning of the experiment. The experiment lasted approximately 2 h and comprised a thresholding phase and a testing phase. Across both phases, participants performed an MOT task, in which they were presented with a 2,000-ms static display wherein seven colored dots encircled by white outlines appeared around a larger black ring. All dots proceeded to move around the ring for 5,000 ms, following predetermined trajectories. Stimulus motion trajectories were precomputed using custom Python code, and initial values for v and zwere drawn from the stimulus motion structure's stationary distributions [3]. When precomputing the trials, we asserted that the final dot locations were nonoverlapping, to prevent selection ambiguity (trials with overlap were regenerated until they met this criterion). Red square outlines subsequently appeared for 3,000 ms around three of the moving dots, marking them as to-be-tracked targets. The target cues and disc colors then faded to black during the following 700 ms, after which the white outlines continued to move for 6,000 ms until coming to a stop. A random number (1 to 7) was subsequently superimposed on each dot, prompting participants to report the identities of the perceived targets by making the appropriate keyboard presses. Feedback was provided after every trial. The

Bill et al.

MOT tasks implemented across both phases were nearly identical, with the following exceptions. Whereas movement trajectories in the thresholding phase followed IND motion, those implemented in the test phase varied across blocks (Latin-square counterbalanced) based on each motion structure condition (IND, GLO, CNT, CDH1, and CDH2). The thresholding phase was used to titrate motion speed by adjusting a per-participant speed factor f_{speed} (SI Appendix, Eq. S2), such that all participants would achieve a common baseline on IND trials. To this end, participants performed 30 IND trials at an initial speed of $f_{\text{speed}} = 2.0$. Whether performance at this initial speed fell above or below the targeted 2.15 performance threshold, participants completed a subsequent 30 trials of the same task at a faster or slower speed (speed factors varied by 0.25 increments). This staircase thresholding procedure was repeated until each participant's average target identification accuracy approximated 2.15 items per trial (mean number of repetitions = 2.40, SD = 0.68). Once thresholding was complete, participants performed 150 trials of the MOT task at their determined speed. Prior to the onset of each motion structure block, participants were presented with three example trials inherent to the forthcoming block of displays, and were provided with a motion graph that explicitly laid out the motion structure based on the colors of the moving items. Example trials and details of experiment conduction are provided in SI Appendix. No data were excluded from the analysis. The study was approved by the Harvard Institutional Review Board (IRB00000109). The data (42) are available for download.

Bayesian Observer Model: Kalman Filtering (Both Experiments). We assume independent Gaussian observation noise on the locations of each video frame, $p(\mathbf{x}_{t,k} | \mathbf{z}_{t,k}) = \mathcal{N}(\mathbf{x}_{t,k}; \mathbf{z}_{t,k}, \sigma_{obs,k}^2)$, for all dots k = 1..7. Together with the latent stochastic linear dynamics of [1] and [2], this defines a generative model. Kalman filters calculate the joint posterior distribution over locations and velocities by inverting the generative model. The circular support of z, however, renders the normal distribution underlying Kalman filtering into an approximation. To maintain a close approximation to the correct posterior, we ensured that variances of the location posterior $p(\mathbf{z}_t | \mathbf{x}_{1:t})$ remained small at all time. For all observable dots, location estimates are much smaller than π anyway; for the unobserved dots in the prediction task, we designed the stimuli such that the correct posterior's SD never exceeded 40°. Thus, most of the probability mass stayed within the (unwrapped) circle at any time, and errors introduced by the Gaussianity assumption are small. The corresponding matrices of the motion-structured Kalman filter are provided in SI Appendix.

Bayesian Observer Models: MOT Task. The above Kalman filter was used for tracking, assuming either the correct motion prior L^{\star} to underlie the stimulus, or an alternative putative motion prior L with identical marginal velocity distributions $\mathcal{N}\left(v_k; 0, \sigma_{v_k}^2\right)$. Following ref. 17, different possible mental dot assignments γ are scored at each frame by comparing the likelihoods of the current observation x_t under the possible assignments. As in ref. 17, this poses a numerically intractable inference process (per frame, there are K! possible assignments; the total number of assignments per trial thus grows as K!#frames). We use discrete particle variational inference (43) for approximate inference, with a single particle and the set of all pairwise dot permutations as differential proposals per frame. For the computational and momentum-free observers, after each frame (at time t), the highest-scoring candidate assignment γ is maintained with the trajectory likelihood $p(\mathbf{x}_{1:t} | \gamma)$ as score function; for the Weber's law observer, the assignment is sampled from $p(\gamma) \propto p(\mathbf{x}_{1:t} \mid \gamma)^4$. The assignment γ at the end of the trial determined which three input dots were chosen as targets. Note that the observer models of the MOT task are not statistically "ideal" observers, since they rely on approximate inference for dot assignments and subsume the effect of noise at all computational stages within a single fitted parameter $\sigma_{\rm obs}.$ In the numerical evaluation, the Kalman filter was provided with noisy observations, $\mathbf{x}_{t,k} \sim \mathcal{N}\left(\mathbf{z}_{t,k}, \sigma_{obs,k}^2\right) \mod 2\pi$, for the computational and Weber's law observers; the momentum-free observer received noise-free observations. The size of $\sigma_{{\rm obs},k}$ was determined via computer simulations to yield, on average, the target performance (2.15) in the IND condition. Therefore the observer models are not expected to perform generally better than human participants. For the computational observer, $\sigma_{\text{obs},k} = 0.05$ for all dots k = 1..7; for the Weber's law observer, $\sigma_{obs,k} = \sigma_0 (1 + 5 |v_{t,k}|)$ and $\sigma_0 \in [0.008, 0.016]$, depending on the speed level f_{speed} of the participant. Given the separate fitting of the noise parameters for both models, the Weber's law observer is not expected to perform generally better or worse than the computational observer in terms of correctly tracked dots. Each trial presented to a human participant was then simulated 25 times (with changing noise instantiations), and simulated performance was averaged over all repetitions to reduce the variance of the observer model's predictions (shrinking the error bars in Fig. 2*B*, *Center* and *Right* and in Fig. 3*B*). The generative model of the momentum-free observer as well as details on the log-likelihood estimation for the additional analysis in Fig. 3*C*, which is based on the simulated repetitions trials, are provided in *SI* Appendix.

Experiment Details: Prediction Task. Twelve adult individuals (mean age = 31 y, SD = 9 y; 10 males) participated in exchange for financial compensation (\$10 per hour plus a performance-dependent bonus). All participants reported normal or corrected-to-normal vision, and provided informed consent at the beginning of the study. On average, the experiment lasted approximately 80 min. Each participant completed 100 trials per motion structure (GLO, CLU, and CDH), presented in blocks (50 unique trials per block, each presented twice in randomized order). Block orders were balanced across participants. Participants were informed about the stimulus condition and performed a variable number of training trials prior to each block until they decided to start the experiment block. Dots were color coded as shown in Fig. 4E. Trials were composed as follows. Initial values for v and z were drawn from the stimulus motion structure's stationary distributions [3]. After a 1,000 ms still period, all dots started moving stochastically for 5,000 ms according to [1] and [2]. During the end of the 5,000-ms period, the red and green target dots faded out, and only the remaining five dots were visible for another 1,500-ms period, after which the scene froze. The green and red dots' locations had to be predicted by directing a green/red mouse cursor to the predicted location on the circle. After each trial, the true dot locations were revealed, and participants received points (0 to 20) based on the accuracy of their prediction. The points only served as task engagement and for payment, and played no role for the analysis. Example trials and details of experiment conduction are provided in SI Appendix. No data were excluded from the analysis. The experiment was approved by the Harvard Institutional Review Board (IRB15-2048). The data (44) are available for download.

Bayesian Observer Model: Prediction Task. Kalman filters with different candidate motion priors *L* were presented with the same trials that had been shown to human participants. For the observer model, we assume correct dot assignment ($\gamma = I$) and set $\sigma_{obs} = 0$ for observed dots, since dot confusion and observation noise-induced errors are expected to play a negligible role in the prediction task. While invisible, we set $\sigma_{obs} \rightarrow \infty$ for the green and red dots. For candidate prior *L*, human responses $\varphi^{(i)} = (\varphi^{(i)}_{green}, \varphi^{(i)}_{red})$ in trial i = 1, ..., 100 are then modeled as

$$p(\boldsymbol{\varphi}^{(i)} \mid \boldsymbol{L}, \theta) = (1 - \pi_{s}) \cdot \mathcal{N} \left(\boldsymbol{\varphi}^{(i)}; \ \boldsymbol{\mu}_{kal}^{(i)}, \ \boldsymbol{a} \Sigma_{kal}^{(i)} + \boldsymbol{b} \boldsymbol{I} \right)$$

$$+ \pi_{s} \cdot \mathcal{N} \left(\boldsymbol{S} \boldsymbol{\varphi}^{(i)}; \ \boldsymbol{\mu}_{kal}^{(i)}, \ \boldsymbol{a} \Sigma_{kal}^{(i)} + \boldsymbol{b} \boldsymbol{I} \right),$$
[4]

where $\mu_{kal}^{(0)}$ and $\Sigma_{kal}^{(0)}$ are the (2D) mean and (2 × 2) covariance matrix of the Kalman filter with motion prior *L* for the green and red dots at the end of the trial. The matrix $S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ models dot swapping in the response with probability π_s , and $a, b \in \mathbb{R}^+$ scale structured inference noise and unstructured motor noise, respectively. Eq. 4 describes a stochastic posterior observer model (the winning model identified in the systematic comparison of ref. 22) with lapses and motor noise. The three free parameters $\theta = (\pi_s, a, b)$ were fitted via maximum likelihood for each candidate prior *L*, stimulus condition L^* , and participant. In Fig. 4 *C* and *D*, only the nonswapped model component is illustrated. The log-likelihood ratios to the model with correct motion prior,

$$LLR_{L^{\star},L} = \sum_{\text{Trials } i} \log \frac{p\left(\varphi^{(i)} \mid L, \theta_{L}^{\text{ML}}\right)}{p\left(\varphi^{(i)} \mid L^{\star}, \theta_{L^{\star}}^{\text{ML}}\right)},$$
[5]

are plotted in Fig. 4E. One-tailed paired t tests identify whether the correct prior L^* explained human responses better than alternative priors L.

Distinguishing Systematic and Stochastic Errors in the Prediction Task. We evaluated the noise factor f_{noise} separately for the green and red dots per stimulus condition and participant; that is, each noise factor is based on the repetitions r = 1, 2 of the unique trials j = 1, ..., 50 in one stimulus block. We define the average noise variance,

where $\langle \cdot \rangle_j$ denotes the average over unique trials, Var $[\cdot]_r$ is the unbiased estimator of the variance over trial repetitions, ${\rm d}(\cdot, \cdot) \in (-\pi,\pi]$ is the (signed) distance on the circle, and $\overline{\varphi}_j$ is the circular mean of the two responses $\varphi_j^{(1)}$ and $\varphi_j^{(2)}$. We further define the total prediction variance

$$\overline{\sigma}_{\text{tot}}^2 = \left\langle \mathsf{d}^2(\varphi^{(i)}, \varphi^{(j)}_{\text{opt}}) \right\rangle_i$$
[7]

as the mean-squared error made compared to the optimal prediction $\varphi_{\text{lopt}}^{(l)} = \mu_{\text{kal}}^{(l)}(L^{\star})$. The estimators $\overline{\sigma}_{\text{noise}}^2$ and $\overline{\sigma}_{\text{tot}}^2 - \overline{\sigma}_{\text{noise}}^2$ were verified to recover the stochastic and systematic error components of synthetic data. The noise factor is defined as the ratio $f_{\text{noise}} = \overline{\sigma}_{\text{noise}}^2 / \overline{\sigma}_{\text{tot}}^2$. Based on synthetic data, the SD of the noise factors in Fig. 5B is estimated to be 0.10.

- D. Kaiser, G. L. Quek, R. M. Cichy, M. V. Peelen, Object vision in a structured world. Trends Cognit. Sci. 23, 672–685 (2019).
- M. Wertheimer, Untersuchungen zur Lehre von der Gestalt. Psychol. Res. 4, 301–350 (1923).
- 3. S. Yantis, Multielement visual tracking: Attention and perceptual organization. *Cognit. Psychol.* 24, 295–340 (1992).
- G. Liu et al., Multiple-object tracking is based on scene, not retinal, coordinates. J. Exp. Psychol. Hum. Percept. Perform. 31, 235–247 (2005).
- M. Suganuma, K. Yokosawa, Grouping and trajectory storage in multiple object tracking: Impairments due to common item motions. *Perception* 35, 483–495 (2006).
- C. Wang, X. Zhang, Y. Li, C. Lyu, Additivity of feature-based and symmetry-based grouping effects in multiple object tracking. *Front. Psychol.* 7, 657 (2016).
- J. Driver, P. McLeod, Z. Dienes, Motion coherence and conjunction search: Implications for guided search theory. *Percept. Psychophys.* 51, 79–85 (1992).
- R. Luria, E. K. Vogel, Come together, right now: Dynamic overwriting of an object's history through common fate. J. Cognit. Neurosci. 26, 1819–1828 (2014).
- G. Johansson, Configurations in Event Perception: An Experimental Study (Almqvist & Wiksell, 1950).
- S. J. Gershman, J. B. Tenenbaum, F. Jäkel, Discovering hierarchical motion structure. Vis. Res. 126, 232–241 (2016).
- H. Xu, N. Tang, J. Zhou, M. Shen, T. Gao, Seeing "what" through "why": Evidence from probing the causal structure of hierarchical motion. *J. Exp. Psychol. Gen.* 146, 896–909 (2017).
- Z. W. Pylyshyn, R. W. Storm, Tracking multiple independent targets: Evidence for a parallel tracking mechanism. Spatial Vis. 3, 179–197 (1988).
- J. Bill, H. Pailian, S. J. Gershman, J. Drugowitsch, Do humans employ hierarchical motion structure during visual perception? Open Science Framework:osf.io/v8254 (20 March2019).
- 14. C. Gardiner, Stochastic Methods (Springer, 2009), Vol. 4.
- B. J. Scholl, "What have we learned about attention from multiple object tracking (and vice versa)" in *Computation, Cognition, and Pylyshyn*, D. Dedrick, L. Trick, Eds. (MIT Press, 2009), pp. 49–78.
- S. Franconeri, S. Jonathan, J. Scimeca, Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity. *Psychol. Sci.* 21, 920–925 (2010).
- E. Vul, M. C. Frank, G. Alvarez, J. B. Tenenbaum, "Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model" in Advances in Neural Information Processing Systems, M. I. Jordan, Y. LeCun, S. A. Solla, Eds., (MIT Press, 2009), pp. 1955–1963.
- S. Ullman, The Interpretation of Visual Motion, Artificial Intelligence Series (MIT Press, 1979).
- M. R. Dawson, The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem. *Psychol. Rev.* 98, 569–603 (1991).
- G. Y. Bae, J. I. Flombaum, Close encounters of the distracting kind: Identifying the cause of visual tracking errors. *Atten. Percept. Psychophys.* 74, 703–715 (2012).
- R. J. Snowden, O. J. Braddick, The temporal integration and resolution of velocity signals. Vis. Res. 31, 907–914 (1991).
- L. Acerbi, S. Vijayakumar, D. M. Wolpert, On the origins of suboptimality in human probabilistic inference. *PLoS Comput. Biol.* 10, e1003661 (2014).

Data and Code Availability. Computer simulations and data analysis were performed with custom Python code which is available on https://github.com/DrugowitschLab/motion-structure-used-in-perception. All human response data have been deposited in Figshare (MOT experiment: 10.6084/m9.figshare.9856271.v1; prediction experiment: 10.6084/m9.figshare.9856274.v1).

ACKNOWLEDGMENTS. We are very thankful to Rick Born, Till Hartmann, Luke Rast, Ariana Sherdil, Valentin Wyart, and three anonymous reviewers for helpful feedback and discussions, as well as to George Alvarez for allowing us to conduct the MOT experiment in his laboratory. This research was supported by a seed grant from the Harvard Brain Initiative, an Alice and Joseph Brooks Fund Fellowship award (J.B.), a research fellowship from the Alfred P. Sloan Foundation (S.J.G.), and a James S. McDonnell Foundation scholar award (Grant 220020462, J.D.). Parts of this research were conducted using the O2 High Performance Compute Cluster at Harvard Medical School.

- J. Drugowitsch, V. Wyart, A. D. Devauchelle, E. Koechlin, Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* 92, 1398– 1411 (2016).
- B. P. Keane, Z. W. Pylyshyn, Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function. *Cognit. Psychol.* 52, 346–368 (2006).
- Y. Weiss, E. P. Simoncelli, E. H. Adelson, Motion illusions as optimal percepts. Nat. Neurosci. 5, 598–604 (2002).
- A. A. Stocker, E. P. Simoncelli, "Sensory adaptation within a Bayesian framework for perception" in Advances in Neural Information Processing Systems, M. I. Jordan, Y. LeCun, S. A. Solla, Eds. (MIT Press, 2006), pp. 1289–1296.
- A. A. Stocker, E. P. Simoncelli, Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* 9, 578–585 (2006).
- J. H. Hedges, A. A. Stocker, E. P. Simoncelli, Optimal inference explains the perceptual coherence of visual motion stimuli. J. Vis. 11, 14–14 (2011).

PSYCHOLOGICAL AND COGNITIVE SCIENCES

- 29. N. Srivastava, E. Vul, Attention modulates spatial precision in multiple-object tracking. *Topics Cognit. Sci.* **8**, 335–348 (2016).
- W. J. Ma, W. Huang, No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. J. Vis. 9, 1–30 (2009).
- P. Cavanagh, G. A. Alvarez, Tracking multiple targets with multifocal attention. *Trends Cognit. Sci.* 9, 349–354 (2005).
- W. M. Shim, G. A. Alvarez, Y. V. Jiang, Spatial separation between targets constrains maintenance of attention on multiple objects. *Psychon. Bull. Rev.* 15, 390–397 (2008).
- L. Luu, A. A. Stocker, Post-decision biases reveal a self-consistency principle in perceptual inference. *Elife* 7, e33334 (2018).
- A. M. Saxe, J. L. McClelland, S. Ganguli, A mathematical theory of semantic development in deep neural networks. *Proc. Natl. Acad. Sci. U.S.A.* 116, 11537–11546 (2019).
- E. Reif et al., "Visualizing and measuring the geometry of BERT" in Advances in Neural Information Processing Systems 32, H. Wallach et al., Eds. (Curran Associates, 2019), Vol. 32, pp. 8594–8603.
- R. T. Born, DC. Bradley, Structure and function of visual area MT. Annu. Rev. Neurosci. 28, 157–189 (2005).
- M. S. Graziano, R. A. Andersen, R. J. Snowden, Tuning of MST neurons to spiral motions. J. Neurosci. 14, 54–67 (1994).
- P. J. Mineault, F. A. Khawaja, D. A. Butts, C. C. Pack, Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proc. Natl. Acad. Sci. U.S.A.* 109, E972–E980 (2012).
- J. M. Beck, P. E. Latham, A. Pouget, Marginalization in neural circuits with divisive normalization. J. Neurosci. 31, 15310–15319 (2011).
- O. W. Layton, N. A. Browning, A unified model of heading and path perception in primate MSTd. PLoS Comput. Biol. 10, e1003476 (2014).
- O. W. Layton, B. R. Fajen, A neural model of MST and MT explains perceived object motion during self-motion. J. Neurosci. 36, 8093–8102 (2016).
- J. Bill, H. Pailian, S. J. Gershman, J. Drugowitsch, Data of the MOT experiment. Figshare. https://doi.org/10.6084/m9.figshare.9856271.v1. Deposited 16 September 2019.
- A. Saeedi, T. D. Kulkarni, V. K. Mansinghka, S. J. Gershman, Variational particle approximations. J. Mach. Learn. Res. 18, 2328–2356 (2017).
- J. Bill, H. Pailian, S. J. Gershman, J. Drugowitsch, Data of the prediction experiment. Figshare. https://doi.org/10.6084/m9.figshare.9856274.v1. Deposited 16 September 2019.