bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Structure in motion: visual motion perception as online hierarchical inference

Johannes Bill[1,2,*], Samuel J. Gershman[2,3,4,†], and Jan Drugowitsch[1,3,†]

[1]Department of Neurobiology, Harvard Medical School, United States
[2]Department of Psychology, Harvard University, United States
[3]Center for Brain Science, Harvard University, United States
[4]Center for Brains, Minds, and Machines, MIT, United States

[*]Corresponding author: johannes_bill@hms.harvard.edu
[†]Co-senior authors

Abstract

Identifying the structure of motion relations in the environment is critical for navigation, tracking, prediction, and pursuit. Yet, little is known about the mental and neural computations that allow the visual system to infer this structure online from a volatile stream of visual information. We propose online hierarchical Bayesian inference as a principled solution for how the brain might solve this complex perceptual task. We derive an online Expectation-Maximization algorithm that explains human percepts qualitatively and quantitatively for a diverse set of stimuli, covering classical psychophysics experiments, ambiguous motion scenes, and illusory motion displays. We thereby identify normative explanations for the origin of human motion structure perception and make testable predictions for new psychophysics experiments. The algorithm furthermore affords a neural network implementation which shares properties with motion-sensitive cortical areas and motivates a novel class of experiments to reveal the neural representations of latent structure.

Introduction

Efficient behavior requires identification of structure in a continuous stream of volatile and often ambiguous visual information. To identify this structure, the brain exploits statistical relations in velocities of observable features, such as the coherent motion of features composing an object (**Fig. 1a**). *Motion structure* thus carries essential information about the spatial and temporal evolution of the environment, and aids behaviors such as navigation, tracking, prediction, and pursuit [1–8]. It remains, however, unclear how the visual system identifies a scene's underlying motion structure and exploits it to turn noisy, unstructured, sensory impressions into meaningful motion percepts.

In recent years, Bayesian inference has provided a successful normative perspective on many aspects of visual motion perception [9–17]. Human perception of motion stimuli spatially constrained by an aperture is well-explained by Bayesian statistical inference [9–11, 14], and neural circuits that integrate local retinal input into neural representations of motion have been identified [18–23]. For the perception of structured motion spanning multiple objects and larger areas of the visual field, however, a comprehensive understanding is only beginning to emerge [15, 24–27]. While "common fate", that is, the use of motion coherence for grouping visual features into percepts of rigid objects, received some experimental support [24, 28], the perception of natural scenes requires more flexible structure representations (e.g., nested motion relations and non-rigid deformations) than common fate alone. Recent theoretical work [15] has introduced a representation of tree structures for the mental organization of observed velocities into nested hierarchies. Theory-driven experiments subsequently demonstrated that the human visual system indeed makes use of hierarchical structure when solving visual tasks [16], and that salient aspects of human motion structure perception can be explained by normative models of Bayesian inference over tree structures [17]. Because these studies were restricted to modeling motion integration only with regard to the perceptual outcome—they analyzed presented visual scenes offline using ideal Bayesian observer models—it remained unclear how the visual system solves the chicken-and-egg problem of parsing (in real time) instantaneous motion in a scene while simultaneously inferring the scene's underlying structure.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We address this question by formulating visual motion perception as online hierarchical inference in a generative model of structured motion. The resulting algorithm is able to explain human perception of motion stimuli covering classical psychophysics experiments, ambiguous motion scenes, and illusory motion displays. The algorithm, which relies on online Expectation-Maximization [29–31], separates inference of instantaneous motion from identifying a scene's underlying structure by exploiting the fact that these evolve on different time-scales. The resulting set of interconnect differential equations decomposes a scene's velocities with the goal of minimizing prediction errors for subsequent observations. Beyond capturing human percepts in many psychophysics experiments qualitatively, the online algorithm explains human motion structure classification quantitatively with higher fidelity than a previous ideal observer-based model [17]. Furthermore, the algorithm provides a normative explanation for the putative origin of human illusory motion perception, and yields testable predictions for new psychophysics experiments.

Finally, we address how motion structure discovery could be supported by neural circuits in the brain. Studying the neural representations underlying motion structure perception is challenging, as the perceived structure often has no direct physical counterpart in the environment (e.g., the concept of a flock velocity in **Fig. 1a**). We derive a recurrent neural network model that not only implements the proposed online algorithm, but shares many properties with motion-sensitive middle temporal area (MT) [21] and dorsal medial superior temporal area (MSTd) [19, 32]. This model in turn allows us to propose a new class of stimuli for neuroscientific experiments that make concrete predictions for neural recordings.

Results

In what follows, we first present the online algorithm for simultaneous hierarchical inference of instantaneous motion and of the scene's underlying structure. Next, we demonstrate the algorithm's ability to explain human motion perception across a set of psychophysics experiments and discuss testable predictions for novel studies. Finally, we propose a biologically realistic neural implementation of online hierarchical inference and identify targeted experiments to reveal neural representations of latent structure.

Online hierarchical inference in a generative model of structured motion

A structural understanding of the scene in **Fig. 1a** requires the observer to decompose observed velocities of objects or their features into latent *motion sources*, s, that, together, compose the scene (**Fig. 1b**). These latent sources might or might not have a direct counterpart in the physical world. In **Fig. 1b**, for instance, each bird's velocity on the observer's retina can be decomposed into the observer's self-motion, s_{self}, the flock's motion, s_{shared}, plus a smaller, animal-specific component, s_{ind}. Here, "flock motion" is an abstract mental concept that is introduced to organize perception, but doesn't have an immediate physical correlate. A correct decomposition leads to motion sources that aid interpretation of the visual scene, and thus supports behaviors such as navigation, tracking, prediction and pursuit. Such decomposition requires knowledge of the scene's structure, like the presence of a flock and which birds it encompasses (**Fig. 1c**). Wrong structural assumptions might lead to faulty inference of motion sources, like wrongly attributing the flock's motion in the sky to self-motion. Thus, the challenge for an observer is to simultaneously infer motion sources and structure online from a stream of noisy and ambiguous visual information.

We formalized the intuition of structured motion in the generative model shown in **Fig. 1d-g**. The stochastic model, first introduced in [16], accommodates fundamental principles of physics (isotropy and inertia) and psychophysics (continuity of trajectories [33] and slow-velocity priors [9]), without making assumptions on specific object trajectories. For example, the motion of three flocking birds viewed by a stationary observer (motion tree in **Fig. 1d**) can be decomposed into four independent motion sources—one shared (magenta) and three individual (green, one per bird)—that evolve according to Ornstein-Uhlenbeck processes [34], generating smooth motion with changes typically occurring at time scale τ_s (**Fig. 1e**). The resulting speed (absolute velocity) distribution of each motion source is governed by an associated *motion strength*, λ, such that the expected speed is proportional to λ. The observable velocities, $\boldsymbol{v}_t$, are in turn noise-perturbed (noise magnitude σ_{obs}; **Fig. 1g**) sums of the individual motion sources (collected in vector $\boldsymbol{s}_t$), with the contribution of each individual motion source specified by a different column of the component matrix $\boldsymbol{C}$ (see **Fig. 1f**). This formalizes the intuition that observable velocities are the sum of their ancestral motion sources in the tree.

In this model, the structure of a scene is fully characterized by the vector of motion strengths, $\boldsymbol{\lambda} = (\lambda_1, .., \lambda_m, .., \lambda_M)$, which describe the presence ($\lambda_m > 0$) or absence ($\lambda_m = 0$) of motion components, as well as their typical speed. In other words, given a reservoir of components, $\boldsymbol{C}$, which might have been learned to occur in visual scenes in general, knowing $\boldsymbol{\lambda}$ is equivalent to knowing the motion structure of the scene. Inferring this structure in turn becomes equivalent to inferring the corresponding motion strengths.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

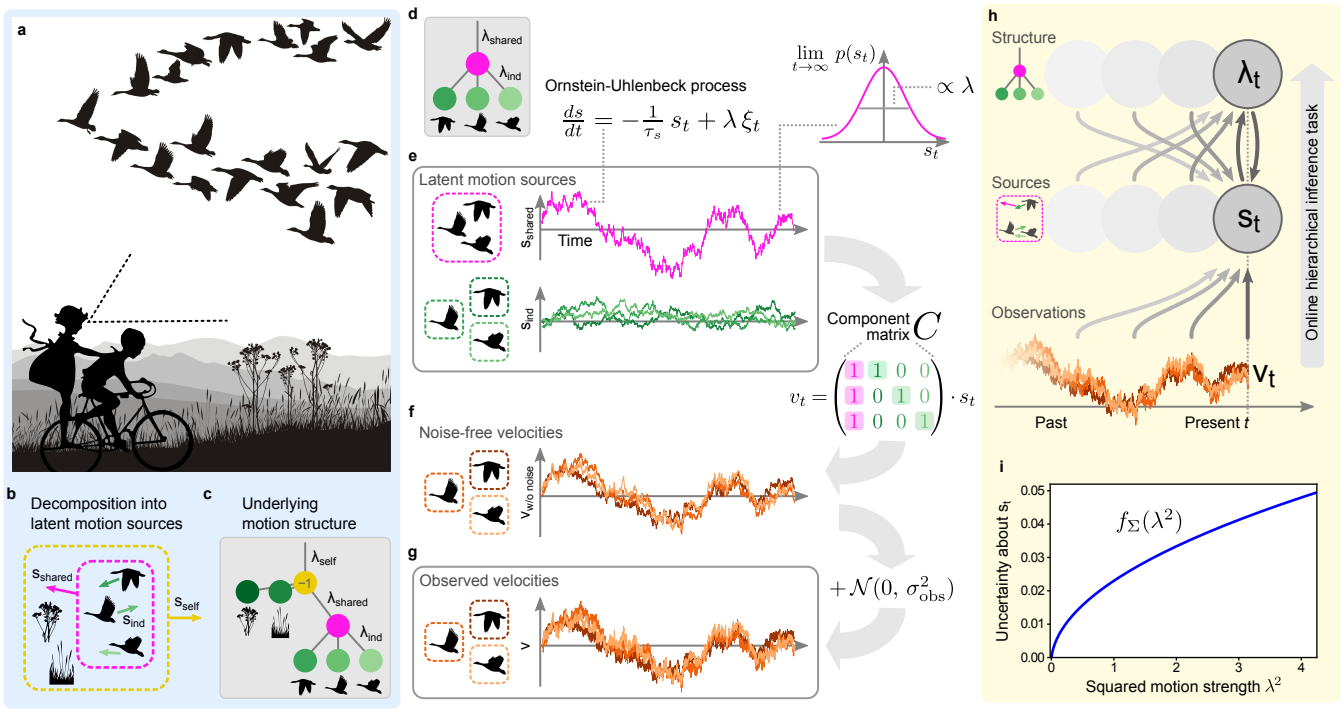

Figure 1. Visual motion perception as an online hierarchical inference task. (a) Scene with nested motion relations. Observed velocities reaching the observer's retina are perceived as a combination of self-motion, flock motion and every bird's individual motion relative to the flock. **(b)** Formal decomposition of the scene's motion into latent motion sources. **(c)** Tree-structured graph representation of the underlying motion structure with nodes corresponding to latent motion sources. Self-motion contributes in the opposite direction to retinal velocity (–1). Vertical distances between nodes, termed motion strengths, λ, describe the long-term average speed of the source. Vanishing motion strength indicates that the corresponding motion source is not present the scene. **(d)–(g)** Generative model of structured motion. **(d)** Graph for a simpler motion scene with three flocking birds and a stationary observer. **(e)** Latent motion sources follow independent Ornstein-Uhlenbeck processes. **(f)** The component matrix, C, composes noise-free velocities from the motion sources, such that each velocity is the sum of all its ancestral sources. **(g)** Observed velocities are noisy versions of the noise-free velocities. **(h)** Inverting the generative model according to Bayes' rule poses an online hierarchical inference task characterized by interdependent updates of motion sources and structure. **(i)** Using an adiabatic approximation, the motion sources' posterior variances reduce to a function of the motion strengths.

An agent faces two challenges when performing inference in this generative model (**Fig. 1h**). First, inference needs to be performed on the fly (i.e., online) while sensory information arrives as an ongoing stream of noisy velocity observations. Second, how observed motion is separated into latent motion sources, $\boldsymbol{s}$, and motion structure, $\boldsymbol{\lambda}$, is inherently ambiguous, such that inference needs to resolve the hierarchical inter-dependence between these two factors. We address both challenges by recognizing that motion structure, $\boldsymbol{\lambda}$, typically changes more slowly than the often volatile values of motion sources, $\boldsymbol{s}$, facilitating the use of an online Expectation-Maximization (EM) algorithm to infer both. This separation of time scales yields a system of mutually dependent equations for updating $\boldsymbol{\lambda}$ and $\boldsymbol{s}$ and furthermore affords a memory-efficient, continuous-time online formulation that is amenable to a neural implementation (see *Methods* for an outline of the derivation, and *Supplemental Information, Section 3* for the full derivation). While the algorithm is approximate, it nonetheless performs adequate online hierarchical inference and closely resembles more accurate solutions, even for deeply nested motion structures (see *Supplemental Figure S1*).

Our online algorithm computes, at any time, a posterior belief over the latent motion sources, $\boldsymbol{s}_t$, which is Gaussian with mean vector $\boldsymbol{\mu}_t$ and covariance matrix $\boldsymbol{\Sigma}_t$, as well as an estimate, $\boldsymbol{\lambda}_t$, of the underlying structure. The dynamics of $\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_t$, and $\boldsymbol{\lambda}_t^2$ (the algorithm is more elegantly formulated on the squared values) read:

$$\partial_t \boldsymbol{\lambda}_t^2 = -\tfrac{1}{\tau_\lambda}\,\boldsymbol{\lambda}_t^2 + \boldsymbol{\alpha} \odot \left(\boldsymbol{\mu}_t^2 + f_\Sigma(\boldsymbol{\lambda}_t^2)\right) + \boldsymbol{\beta}\ , \tag{1}$$

$$\partial_t \boldsymbol{\mu}_t = -\tfrac{1}{\tau_s}\,\boldsymbol{\mu}_t + f_\Sigma(\boldsymbol{\lambda}_t^2) \odot \boldsymbol{C}^\mathsf{T}\,\boldsymbol{\epsilon}_t \qquad \text{with} \qquad \boldsymbol{\epsilon}_t = \frac{v_t}{\sigma_{\text{obs}}^2} - \frac{\boldsymbol{C}\boldsymbol{\mu}_t}{\sigma_{\text{obs}}^2}\ , \tag{2}$$

$$\text{and} \quad \boldsymbol{\Sigma}_t = \operatorname{diag}\left[f_\Sigma(\boldsymbol{\lambda}_t^2)\right]\ . \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

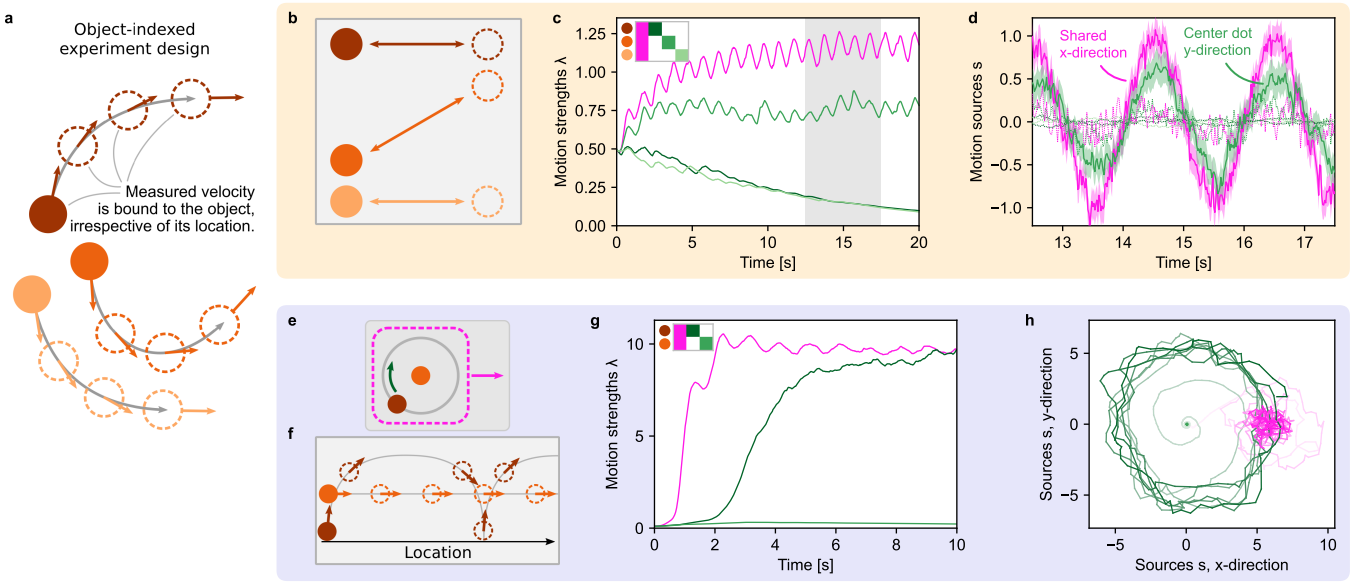

Figure 2. Online hierarchical inference replicates human perception of classical motion displays. **(a)** In object-indexed experiment designs, every observable velocity is bound to an object irrespective of its location. Many psychophysics studies fall into this class of experiment design. **(b)** Johansson's 3-dot motion display. Humans perceive the stimulus as shared horizontal motion with the central dot oscillating vertically between the outer dots. **(c)** The online algorithm's estimate of the motion strengths, $\boldsymbol{\lambda}_t$ (a single motion strength is shared across both spatial dimensions). The component matrix, $\boldsymbol{C}$, is shown in the top-left as a legend for the line colors. Circles next to the matrix show the assignment of the rows in $\boldsymbol{C}$ to the dots in panel b. **(d)** The algorithm's posterior distribution over the motion sources, $\boldsymbol{s}_t$, during the gray-shaded period in panel c. Shown are the mean values, $\boldsymbol{\mu}_t$, as lines along with the algorithm's estimated standard deviation (shaded, only for two components for visual clarity). **(e)** The Duncker wheel resembles a rolling wheel of which only the hub and one dot on the rim are visible. **(f)** Despite its minimalist trajectory pattern, humans perceive a rolling wheel. **(g)** Inferred motion strengths, $\boldsymbol{\lambda}_t$. The algorithm identifies shared motion plus an individual component for the revolving dot. **(h)** Inferred motion sources, $\boldsymbol{\mu}_t$, for the duration in panel g. Color gradients along the lines indicate time (from low to high contrast). For visual clarity, $\boldsymbol{\mu}_t$ has been smoothed with a 50 ms box filter for plotting.

The coupled equations (1)–(3) support the following intuition. Eqn. (1) calculates a running average of the motion strengths $\boldsymbol{\lambda}_t^2$ by use of a low-pass filter with time scale τ_λ. Here, $\odot$ denotes element-wise multiplication and the function $\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)$ (**Fig. 1i**) estimates the variance of the s-posterior distribution according to an adiabatic approximation (cf. eqn. (3), see *Methods*). The constants $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ contribute a sparsity-promoting prior, $p(\boldsymbol{\lambda}^2)$, for typical values of the motion strengths (see *Methods* for the their full expressions). By eqn. (2), the motion source means $\boldsymbol{\mu}_t$ are estimated by a slightly different low-pass filter that relies on a prediction error, $\boldsymbol{\epsilon}_t$, between the model's expected velocities, $\boldsymbol{C}\boldsymbol{\mu}_t$, and those actually encountered in the input, $\boldsymbol{v}_t$ (both normalized by observation noise variance to facilitate the later network implementation). This prediction error on observable velocities is transformed back to the space of latent motion sources via the transposed component matrix $\boldsymbol{C}^\mathsf{T}$ and then, importantly, gated by element-wise multiplication ($\odot$) with the variance estimates $\boldsymbol{f}_\Sigma(\boldsymbol{\lambda}_t^2)$. This gating implements a credit assignment as to which motion source was the likely cause of observed mismatches in $\boldsymbol{\epsilon}_t$, and thus uses the scene's currently inferred motion structure to modulate the observed velocities' decomposition into motion sources. For flocking birds, for example, a simultaneous alignment in multiple birds' velocities would only be attributed to the shared flock velocity if such a flock had been detected in the past (λ_shared large, and λ_ind small). Otherwise it would be assigned to the birds' individual motions, s_ind.

Together, eqns. (1) and (2) implement a coupled process of structure discovery and motion decomposition, which distinguishes them through different time-scales. Notably, the proposed algorithm is not a heuristic, but is derived directly from a normative model of online hierarchical inference. Next, we explored if the algorithm can explain prominent phenomena of human visual motion perception.

Online inference replicates human perception of classical motion displays

To explore if the proposed online algorithm can qualitatively replicate human perception of established motion displays, we simulated two classical experiments from Gunnar Johansson [25] and Karl Duncker [35]. These experiments belong to a class of visual stimuli which we refer to as *object-indexed experiments* (**Fig. 2a**) because the observed velocities, $\boldsymbol{v}_t$,

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

belong to objects irrespective of their spatial locations. (*Location-indexed experiments* will be discussed below.)

In Johansson's experiment, three dots oscillate about the screen with two of the dots moving horizontally and the third dot moving diagonally between them (see **Fig. 2b** and *Supplemental Video S1*). Humans perceive this stimulus as a shared horizontal oscillation of all three dots, plus a nested vertical oscillation of the central dot. Similar to previous offline algorithms [15], our online algorithm identifies the presence of two motion components (**Fig. 2c**): a strong shared motion strength, λ_{shared} (magenta) and weaker individual motion, λ_{ind}, for the central dot (green). The individual strengths of the outer two dots (light and dark green), in contrast, decay to zero. Most motion sources within the structure are inferred to be small (dotted lines in **Fig. 2d**). Only two sources feature pronounced oscillations: the x-direction of the shared motion source, $\mu_{\text{shared},x}$, (magenta, solid line) and the y-direction of the central dot's individual source, $\mu_{\text{ind},y}$, (green, solid line), mirroring human perception. As observed velocities are noisy, they introduce noise in the inferred values of $\boldsymbol{\mu}_t$, which fluctuate around the smooth sine-functions of the original, noise-free stimulus. As expected from well-calibrated Bayesian inference, the magnitude of these fluctuations is correctly mirrored in the algorithm's uncertainty, as illustrated by the posteriors' standard deviation $\sqrt{f_\Sigma(\boldsymbol{\lambda}_t^2)}$ (shaded areas in **Fig. 2d**).

In the second experiment, known as the Duncker wheel, two dots follow the motion of a rolling wheel, one marking the hub, the other marking a point on the rim (**Fig. 2e**). The two dots describe an intricate trajectory pattern (see **Fig. 2f** and *Supplemental Video S2*), that, despite its impoverished nature, creates the impression of a rolling object for human observers, a percept that has been replicated by offline algorithms [15]. Likewise, our online algorithm identifies a shared (magenta in **Fig. 2g**) plus one individual (dark green) component, and decomposes the observed velocities into shared rightward motion plus rotational motion for the dot on the rim (see **Fig. 2h**). Notably, the shared motion component is discovered before the revolving dot's individual motion, leading to a transient oscillation in the inferred shared motion source, μ_{shared} (see light magenta trace in **Fig. 2h**) — an onset effect that could be tested experimentally.

In summary, the online hierarchical inference algorithm successfully identified the structure underlying the motion displays, provided Bayesian certainty estimates for the inferred motion, and replicated human perception in these classical psychophysics experiments.

Online inference outperforms ideal observers in explaining human structure perception

Having qualitatively replicated motion structure inference in common motion displays, we next asked if our online algorithm could quantitatively explain human motion structure perception. To address this question, we reevaluated behavioral data from Yang et al. [17], where participants had to categorize the latent structure of short motion displays (see **Fig. 3a**). Motion scenes followed one of four structures (**Fig. 3b**) and were generated stochastically from the same generative model underlying our algorithm. Owing to their stochastic generation, scenes often were ambiguous with regard to their latent structure, prompting distinct error patterns in human responses (see confusion matrix in **Fig. 3c**). For instance, independently moving dots were more frequently misclassified as clustered motion (I-C element) than vice versa (C-I element), global motion was highly recognizable, and nested hierarchical motion was more frequently misperceived as clustered than as global.

To test if human responses arise from normative, Bayesian motion structure inference, Yang et al. modeled these responses in two steps (blue branch in **Fig. 3d**): first, an offline Bayesian ideal observer, which was provided with the trajectories of all objects within a trial, calculated the likelihood for each of the four structures. Then, these four probabilities were fed into a choice model with a small set of participant-specific fitting parameters (see *Methods)*. This model captured many aspects of human responses, including task performance, typical error patterns, single-trial responses, and participant-specific differences. Yet, the model arrived at these probabilities by comparing the likelihoods of the full sequences for all four candidate structures, and so had no notion of how a percept of structure could emerge over the course of the trial.

Thus, we next asked if our online inference algorithm, which gradually infers the structure during the stimulus presentation, was better able to account for the observed response pattern. As our algorithm by design inferred real-valued motion strengths $\boldsymbol{\lambda}$ rather than only discriminating between the four structures used in the experiment, we added an additional stage that turned the inferred motion strengths into a likelihood for each of the four structures at trial end (red branch in **Fig. 3d**, see *Methods*). To do so, we computed five hand-designed features from the seven-dimensional vector $\boldsymbol{\lambda}_t$ (besides one global and three individual strengths, there are three possible two-dot clusters), and trained a multinomial logistic regression classifier on the features to obtain likelihood values for each of the structures. The classifier was trained on the true structures of the trials, and thus contained no information about human responses. Finally, we fitted the same choice model as Yang et al. to the participants' responses.

The confusion matrix predicted by our model shows an excellent agreement with human choices, both when

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

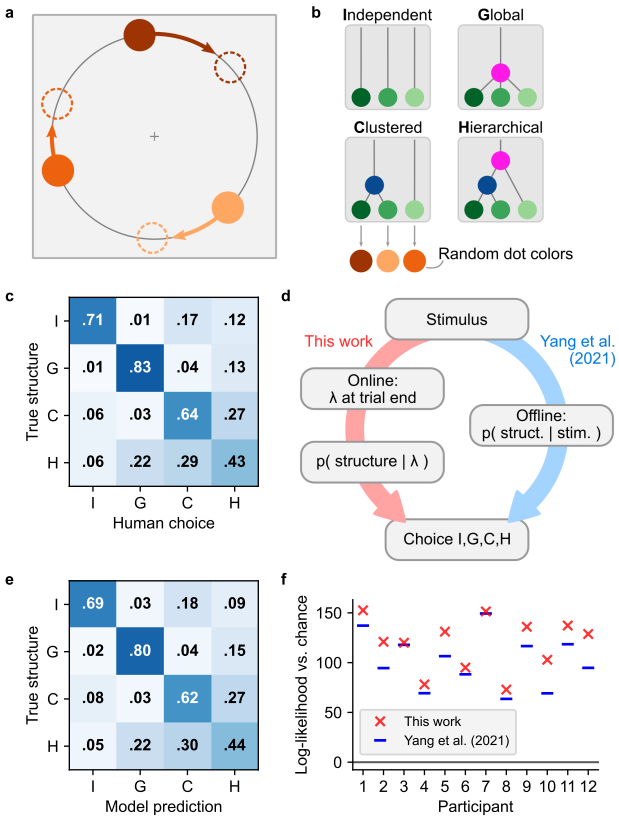

Figure 3. The model quantitatively explains human perception of nested and ambiguous motion scenes. **(a)** Stochastic motion stimulus from Yang et al. [17] consisting of three dots rotating on a circle. **(b)** Each trial followed one of four motion structures. If clustered motion was present (C or H structure), any pair of dots could form the cluster. **(c)** Confusion matrix of human responses, averaged over all 12 participants. **(d)** Models for predicting human responses. Yang et al. employed a Bayesian ideal observer as the basis for fitting a participant-specific choice model. Our algorithmic model, in contrast, calculates the likelihood for each structure from the motion strengths, λ_t, at trial end and then fits the same choice model as Yang et al. for translating probabilities into human responses. **(e)** Confusion matrix of our algorithmic model. **(f)** Log-likelihood of human responses relative to chance level, for both models. The analyses in panels e and f are leave-one-out cross-validated to prevent overfitting.

averaged across participants (**Fig. 3e**), and on a per-participant basis (see *Supplemental Figure S3 and S4*). Indeed, our model beats the original computational model in terms of response log-likelihoods for all of the 12 participants (see **Fig. 3f**; $p < 0.001$, two-sided paired Wilcoxon signed-rank test). Furthermore, the algorithmic model overcomes the systematic under-estimation of global motion (G-G matrix element) that previous, ideal observer-based approaches suffered from [16, 17]. Importantly, in our model, any information connecting the stimulus to the eventual choice is conveyed through the motion strengths, λ_t, as a bottleneck. The fact that the online hierarchical inference-based approach describes human responses better than the ideal observer-based model of Yang et al. indicates that our algorithm may share mechanistic features with the human perceptual apparatus.

Explaining motion illusions that rely on spatial receptive fields

In contrast to the *object-indexed experiments* discussed above, another class of psychophysics experiments employs velocity stimuli that remain at stationary locations (see **Fig. 4a**), typically in the form of apertures of moving dots or drifting Gabors. This class, which we refer to as *location-indexed experiments*, is furthermore popular in neuroscience as it keeps the stimulus' local visual flow within an individual neuron's spatial receptive field throughout the trial [21]. We investigated our model's ability to explain illusory motion perception in two different types of location-indexed experiments: motion direction repulsion in random-dot kinematograms (RDKs) [36, 37, 39, 40], and noise-dependent motion integration of spatially distributed stimuli [38, 41].

We modeled perception in these experiments by including a self-motion component and added a vestibular input signal to the observables (see **Fig. 4b**, and cf. **Fig. 1a-c**). The vestibular input, which we fixed to have zero mean

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

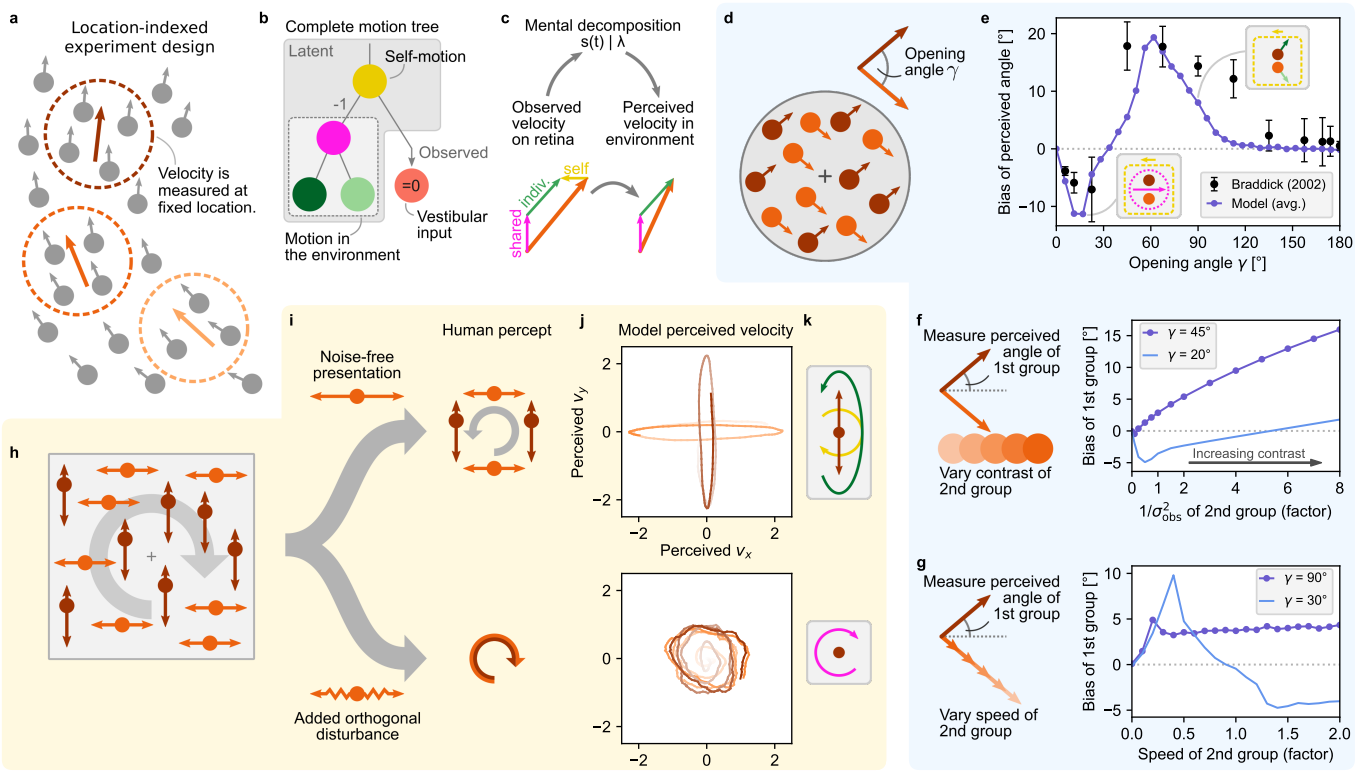

Figure 4. Hierarchical inference explains motion illusions in location-indexed experiments. **(a)** In location-indexed experiments, motion flow is presented at stationary spatial locations. **(b)** Considered latent motion components. Self-motion, which affects all retinal velocities in the opposite direction (-1) integrates both visual input and a vestibular signal (here: zero + noise). **(c)** Perceived object velocities, relative to the environment, are the sum of all inferred motion components excluding self-motion. **(d)** In motion direction repulsion experiments, two groups of dots move at constant velocity with opening angle γ. **(e)** The direction in which human perception of the opening angle is biased depends on the true opening angle. Black dots: human data, reproduced from [36]. Purple line: model precept. Insets: the algorithm's inferred motion decomposition. **(f)** Varying the contrast of one dot group modulates the biased percept of the angle of the other group. Purple: model percept for $\gamma = 45°$, qualitatively matching data from [37]. Blue: predicted inversion of the bias for smaller opening angles. **(g)** Same as panel f, but for varying the speed of the second group. Purple: model prediction for $\gamma = 90°$, qualitatively matching data from [36]. Blue: predicted biphasic function for smaller opening angles. **(h)** In the motion illusion from Lorenceau [38], a vertically and a horizontally oscillating group of dots maintain a $90°$-phase shift consistent with global clockwise rotation (indicated as gray arrow). **(i)** The noise-free stimulus (top) evokes transparent motion with an additional counter-clockwise rotating percept in human observers. Adding motion noise by disturbing dot trajectories orthogonal to their group's oscillation axis (bottom; modeled by increased observation noise σ^2_{obs}) flips the percept to a single coherent rotation of all dots in clockwise direction. **(j)** The algorithm's perceived velocities in both stimulus conditions (time = color gradient from low to high contrast; $t \leq 2$ s in noise-free condition; $t \leq 5$ s in noisy condition). For visual clarity, perceived velocities have been smoothed with a 200 ms box filter for plotting. **(k)** Illustration of the algorithm's inferred motion decomposition. For noise-free stimuli, clockwise rotating self-motion is compensated by counter-clockwise rotating group motion (sketched here for the vertical group). With motion noise, only a single, clockwise rotating shared motion component is inferred for all dots.

plus observation noise, complemented the visual input, which is ambiguous with regard to self-motion and globally shared object motion and can induce illusory self-motion ("vection") [42, 43]. In turn, we model the subjectively perceived velocity of objects, relative to the stationary environment, as the sum of all inferred motion sources excluding self-motion (see **Fig. 4c** and *Methods*).

In the RDK experiment, a participant fixates the center of an aperture in which two groups of randomly positioned dots move linearly with opening angle γ (see **Fig. 4d**) and subsequently reports the perceived opening angle. Motion direction repulsion occurs if the perceived angle is systematically biased relative to the true opening angle.

As previously reported, the repulsion bias can change from an under-estimation of the opening angle for small angles to an overestimation for large angles (data from [36] reprinted as black dots in **Fig. 4e**). We replicated this effect by simulating two constant dot velocities with opening angles that varied across trials. Our algorithm decomposed the stimulus into self-motion, shared motion and individual (group) motion. Across opening angles, it featured a

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

triphasic psychometric function with angles smaller than ~40° being under-estimated, angles between ~40° and ~110° being over-estimated, and even larger angels being unbiased (purple curve in **Fig. 4e**). The match with human biases arose without systematic tuning of simulation parameters (the simulations presented in this manuscript were mostly performed with a set of default parameters, see *Methods*). Inspecting the algorithm's inferred motion components revealed that, for small γ, the negative bias arose from integrating all dots into a single, coherent motion component while disregarding individual dot motions (left inset in **Fig. 4e**). Intermediate γ, in contrast, caused the shared component to be correctly broken up into two individual components—plus a small illusory self-motion component (right inset in **Fig. 4e**). This self-motion, which is ignored in the perceived velocities, widened the perceived opening angle between the two groups of dots. For even larger γ, the illusory self-motion vanished yielding unbiased percepts.

For fixed opening angles, motion direction repulsion is furthermore modulated by relative contrast and speed difference between the two motion components. Specifically, for an opening angle of γ=45°, Chen et al. [37] has shown that increasing the contrast of one dot group inflates the perceived opening angle—here measured relative to horizontal to separate cause and effect—of the other, constant-contrast group (**Fig. 4f**, left). We replicated this effect in simulations that operationalized visual contrast as an (inverse) multiplicative factor on the observation noise variance, σ^2_{obs}. For an opening angle of γ=45°, our algorithm featured a positive and monotonically increasing repulsion bias as the second group's contrast increases (purple line in **Fig. 4f**, right), similar to what has been previously reported. For smaller opening angles, in contrast, our algorithm predicts an inversion of the repulsion bias, which first decreases at low contrast and then increases again for higher contrast (blue line in **Fig. 4f**, right)—a prediction that remains to be tested. Increasing the speed of one motion component for large opening angles also introduces a positive bias in the perceived opening angle of the other component in human participants [36]. We replicated this effect by increasing the second group's speed, which, for a γ=90°opening angle, yielded a relatively stable bias of ~5° across different motion speeds (purple line in **Fig. 4g**), in line with the aforementioned experimental data from Braddick et al. [36]. Furthermore, our algorithm predicts that the speed-dependent bias changes to a biphasic curve for smaller opening angles (blue line), providing another testable prediction.

Turning to noise-dependent motion integration of spatially distributed stimuli, we investigated a motion illusion by Lorenceau [38] which has received little attention in the literature. Two groups of dots oscillate in vertical and horizontal orientation, respectively (see **Fig. 4h** and *Supplemental Video S3*). Both groups follow sine-waves with identical amplitude and frequency, but maintain a relative phase shift of $\pi/2$ that is consistent with an imaginary global clockwise (CW) rotation (indicated by a gray arrow in **Fig. 4h**). This stimulus can be considered *location-indexed*, as the small oscillation amplitude of less than 1 degree of visual angle caused the stimulus to conveniently fit into the receptive fields of individual neurons of the human homologue of area MT [44]. Interestingly, the stimulus' percept changes once disturbances orthogonal to the axes of oscillation are added (called "motion noise" in [38], see **Fig. 4i**). Without motion noise, participants perceive transparent motion, that is, the dots within either group are combined to a rigidly moving object according to "common fate", and both groups are perceived as moving separately. Their movement, however, is not perceived as strictly vertically and horizontally, but rather the stimulus induces an impression of slight counter-clockwise (CCW) rotation, that is, "*opposite to veridical*" [38]. With motion noise, in contrast, the percept switches in two ways: all dots appear to move coherently along a circle, and the perceived direction of movement becomes CW. These percepts are illustrated in **Fig. 4i**.

Applied to this stimulus, our algorithm replicates the perceived rotation direction reversal with increased motion noise, which we simulated through an increase in the observation noise σ^2_{obs}. Specifically, the algorithm's perceived velocities for both groups of dots featured a slight global CCW rotation on top of two generally separated groups for the noise-free stimulus, and a single global CW rotation once observation noise is increased (**Fig. 4j**). Inspecting the algorithm's motion decomposition provides a possible answer to how this flip in perceived rotation emerges, which is illustrated in **Fig. 4k** by the example of the vertical group. On noise-free presentation, dot motion was decomposed into clockwise rotating self-motion (golden arrow) plus a vertically elongated, yet slightly CCW rotating group motion (green arrow), leading to the transparent CCW motion percept. Once observation noise increased, the inferred motion structure discarded the separated groups in favor of a single global motion component (magenta), leading to the percept of coherent CW rotation for all dots (see *Supplemental Figure S5* for trajectories of the motion strengths and sources under both conditions).

Experimental predictions from a biological network model of hierarchical inference

Finally, we asked whether and how a biologically plausible neural network could implement our online hierarchical inference algorithm. To this end, we devised a recurrent neural network model of rate-based neurons. Naturally, such modeling attempt relies on many assumptions. Nonetheless, we were able to identify several experimentally testable predictions that could help guide future neuroscientific experiments.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

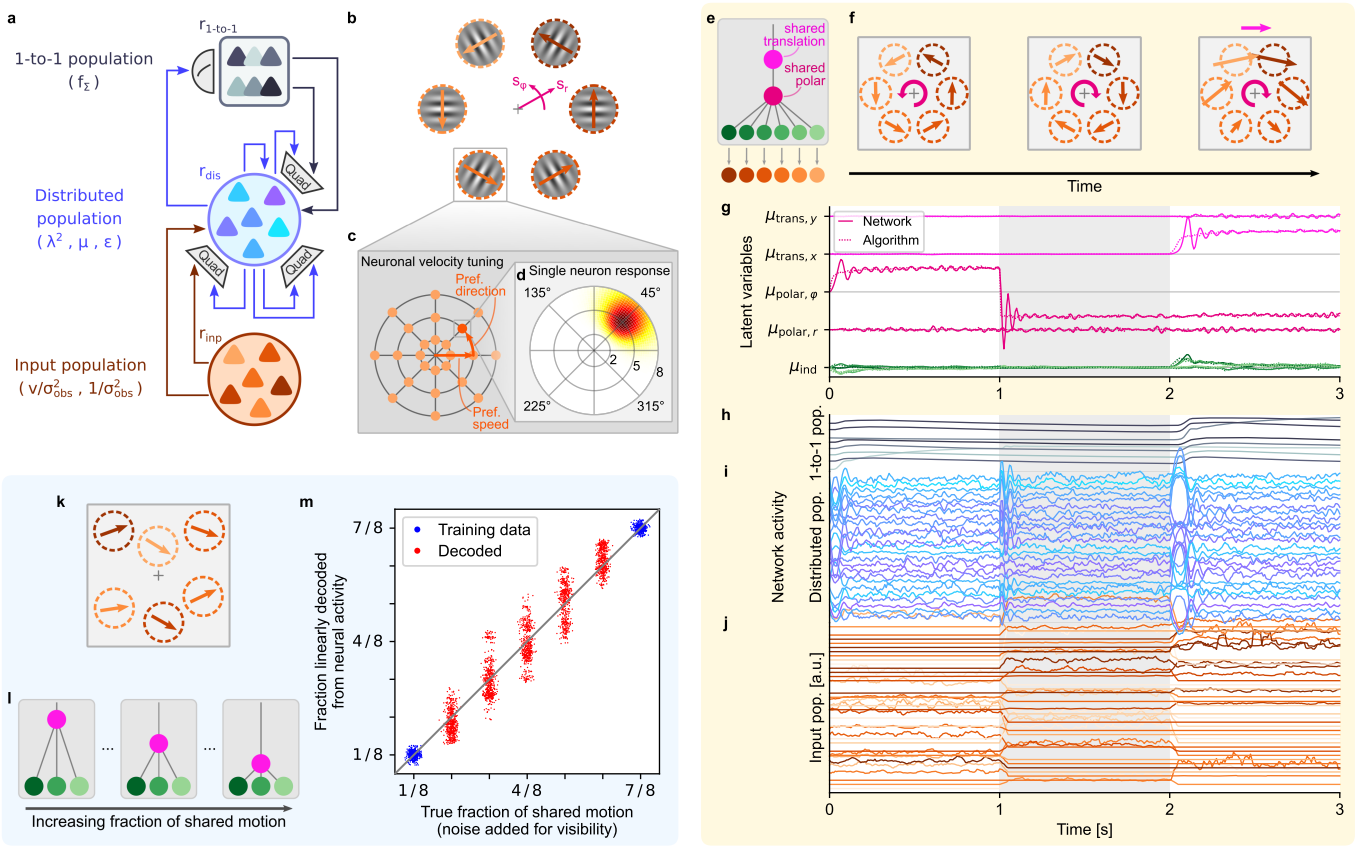

Figure 5. Hierarchical inference can be performed by a biologically realistic network model. (a) Network model implementing the online algorithm. Linear and quadratic interactions are indicated by direct arrows and "Quad" boxes, respectively. In parentheses, the variables represented by each population. **(b)** Rotational stimulus in a location-indexed experiment. Besides translational (Cartesian) motion, the inference algorithm also supports rotational, s_φ, and radial motion, s_r. **(c)** Tuning centers in a model of area MT. A local population of neurons, which share the spatial receptive field highlighted in panel b, cover all directions and speeds with their velocity tuning centers. **(d)** Response function for the neuron highlighted in panel c. The neuron responds strongly to local velocities into the upper-right direction with a speed of ~5 °/sec. Max. rate = 29.5 spikes / s. **(e)** Motion structure used for the network simulation in panels f–j, including simultaneous translational, rotational and radial motion sources. **(f)** Illustration of the stimulus. After $1s$ of counter-clockwise rotation around the fixation cross, the rotation switches to clockwise. At $t=2$ s, rightward translation is superimposed on the rotation. **(g)** Motion sources inferred by the network (solid lines: distributed population read-out; dotted lines: online algorithm solution). Shown is μ_t for translational, rotational, radial and individual motion. Only 4 individual components (2 x- and 2 y-directions) are shown for visual clarity. **(h)** Firing rates of the 1-to-1 population. Rates are in arbitrary units (a.u.) because the theory supports scaling of firing rates with arbitrary factors. **(i)** Same as panel h, but for a random subset of 25 neurons of the distributed population. **(j)** Same as panel h, but for a random subset of 40 neurons of the input population, and smoothed with a 50 ms box filter for plotting. **(k)** Stimulus of a proposed neuroscience experiment. Velocities in distributed apertures follow the generative model from **Fig. 1** using shared motion and individual motion. **(l)** Different trials feature different relative strengths of shared and individual motion, ranging from close-to-independent motion (left) to highly correlated motion (right). **(m)** Linear readout of the fraction of shared motion from neural activity. Seven different fractions of shared motion were presented (x-axis; noise in x-direction added for plotting, only). A linear regression model was trained on the outermost conditions (blue dots). Intermediate conditions were decoded from the network using the trained readout (red dots). Only a subset of 7 × 500 = 3500 points is shown for visual clarity.

Following Beck et al. [45], we assumed that task-relevant variables can be decoded linearly from neural activity ("linear population code") to support brain-internal readouts for further processing, actions and decision making. Furthermore, we employed a standard model for the dynamics of firing rates, $r_i(t)$, and assumed that neurons can perform linear and quadratic integration [45–47]:

$$\tau_i \partial_t r_i = -r_i + f_i(\boldsymbol{w}_i^\mathsf{T} \boldsymbol{r} + \boldsymbol{r}^\mathsf{T} \boldsymbol{Q}^{(i)} \boldsymbol{r}) \ , \tag{4}$$

with time constant τ_i, activation function $f_i(\cdot)$, weight vector $\boldsymbol{w}_i$ and matrix $\boldsymbol{Q}^{(i)}$ for linear and quadratic integration,

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

respectively. The rate vector, $\boldsymbol{r}(t)$, here comprises all presynaptic firing rates, including both input and recurrent populations. With these assumptions, we derived a network model with the architecture shown in **Fig. 5a**, which implements the online algorithm via its recurrent interactions and supports linear readout of all task-relevant variables. That is, for every task-relevant variable, x, there exists a vector, $\boldsymbol{a}_x$, such that $x = \boldsymbol{a}_x^\mathsf{T}\,\boldsymbol{r}$ (see *Supplemental Information, Section 5* for the derivation).

The network consists of three populations. The *input population* (bottom in **Fig. 5a**) encodes the observed velocities, $v_t/\sigma^2_{\text{obs}}$, and observation precision, $1/\sigma^2_{\text{obs}}$, in a distributed code. While any code that supports linear readout of these variables could serve as valid neural input, we chose a specific model that, as shown below, captures many properties of motion-sensitive area MT. The *distributed population* (center in **Fig. 5a**) simultaneously represents the squared motion strengths, $\boldsymbol{\lambda}_t^2$, mean of the sources, $\boldsymbol{\mu}_t$, and prediction errors, $\boldsymbol{\epsilon}_t$, in a distributed code with linear readout. For those, almost arbitrary readouts suffice, such that we chose randomly generated readout vectors, $\boldsymbol{a}$. Notably, we propose the prediction errors, $\boldsymbol{\epsilon}_t$, to be linearly decodable, which allowed eqn. (2) to be implemented with the neuron model in eqn. (4) (see *Supplemental Information, Section 5.3 and 5.4*). All neurons in the distributed population have simple activation functions, $f_i(\cdot)$, that are linear around some baseline activity. The linear decodability of $\boldsymbol{\lambda}_t^2$, $\boldsymbol{\mu}_t$, and $\boldsymbol{\epsilon}_t$ are testable predictions. Finally, the *1-to-1 population* (top in **Fig. 5a**) represents the uncertainty, $\boldsymbol{\Sigma} = \boldsymbol{f}_\Sigma(\boldsymbol{\lambda}^2)$, in a one-to-one mapping, $r_m \propto f_\Sigma(\lambda_m^2)$, with r_m being the firing rate of either a single cell or, more likely, a small population. The theoretical motivation behind this representation is two-fold: on the one hand, the non-linear form of $f_\Sigma(\cdot)$ prevents a distributed, linearly decodable representation (see *Supplemental Information, Section 5.5*); on the other hand, the particular shape of $f_\Sigma(\lambda_m^2)$, shown in **Fig. 1i**, mirrors the typical activation function of Type-I neurons [48], such that the proposed representation emerges naturally for the activation function, $f_\Sigma(\boldsymbol{a}_{\lambda_m^2}^\mathsf{T}\,\boldsymbol{r})$, in the 1-to-1 population (using the fact that λ_m^2 can be read out neurally with weights $\boldsymbol{w} = \boldsymbol{a}_{\lambda_m^2}$). Overall, the network structure predicts $\boldsymbol{\lambda}_t^2$, $\boldsymbol{\mu}_t$, and $\boldsymbol{\epsilon}_t$ to be linear decodable, and the components of $\boldsymbol{f}_\Sigma$ to be independently encoded in single neurons or small neural populations.

Even though the network model supports both the object-indexed and location-indexed experiments from **Fig. 2–4**, the retinotopic organization of the early visual system [21, 49] brings a location-indexed perspective closer in line with our understanding of how the cortex encodes visual information. Furthermore, as we show in *Supplemental Information, Section 2.5*, our model can be extended to support motion sources in polar coordinates (see **Fig. 5b**), such that it supports salient real-world retinal input motifs, such as rotation and radial expansion/contraction about the fovea. Representations of angular motion, s_φ, and radial motion, s_r, can also coexist with translational motion (i.e., linear motion in Cartesian coordinates) within the same population. Selective neural response to rotation, expansion/contraction and translation, as well as combinations thereof, such as spiraling, has been frequently reported in the dorsal medial superior temporal area (MSTd) [19, 50].

Before demonstrating this capability in simulations, let us provide further information about the model's input population, and how it relates to known properties of area MT. To do so, consider the location-indexed stimulus in **Fig. 5b**. During fixation, each aperture stimulates a population in retinotopically organized, motion sensitive area MT [21]. Neurons in MT are tuned to respond preferentially to a certain direction and speed (**Fig. 5c**), such that the full population jointly covers all velocities in a polar grid [51, 52]. The response of individual neurons to velocities within their spatial receptive field is commonly modeled by a log-normal function for speed [52] and a von Mises function for direction [53], leading to the bump-like response function shown in **Fig. 5d**. As a third factor, higher visual contrast (smaller σ^2_{obs}) leads to higher firing rates [54]. As we derive in *Supplemental Information, Section 5.6*, a neural population with these response functions supports linear readout of input velocities, $v_t/\sigma^2_{\text{obs}}$, and precision, $1/\sigma^2_{\text{obs}}$, in Cartesian coordinates. This provided us with a biologically realistic and, at the same time, theoretically grounded input population model which we used in the following network simulations.

We tested the network's ability to perform online hierarchical inference in the simulation shown in **Fig. 5e–j**. To challenge the network, we employed a stimulus that combined shared rotation and shared translation (motion tree in **Fig. 5e**). Six input populations with receptive fields shown in **Fig. 5f** projected to a distributed population of 100 neurons and a 1-to-1 population of size 8 (one per motion strength). After one second of retinal velocities of counter-clockwise rotation (**Fig. 5f**, left), these velocities switched to clockwise rotation (center), followed by a superposition of clockwise rotation and rightward translation (right). As the network response for the three populations to this stimulus shows (**Fig. 5h–j**), input neurons fired sparsely and were only active if the stimulus matched their preferred direction or speed. Neurons in the distributed population, in contrast, showed fluctuating activity with little apparent structure, and exhibited population-wide transients upon changes of the input. Finally, the 1-to-1 population responded more graded and with a short delay, suggesting that every rate, r_m, describes a small cortical population rather than individual neurons. Knowledge of the (randomly drawn) vectors, $\boldsymbol{a}_x$, of the simulated network, allowed us to read out the network's latent motion decomposition at each time point (solid lines in **Fig. 5g**). This revealed that the network

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

correctly decomposed the input, including the overlaid rotational and translational motion, and closely matched the online algorithm (dotted lines).

In experiments with humans and animals, we have no access to these readout vectors, $\boldsymbol{a}_x$. We therefore simulated a possible experiment that tests our model and doesn't require this knowledge (see **Fig. 5k–m**), while benefiting from precise stimulus control. Several apertures, located at the receptive fields of recorded neurons in motion sensitive areas (e.g., area MT or MSTd), present a motion stimulus according to the generative model from **Fig. 1**. Velocities across the apertures are positively correlated owing to a shared motion source, but also maintain some individual motion (see **Fig. 5k** and *Supplemental Video S4*). A series of trials varies the fraction of shared motion in the stimulus, $q := \lambda^2_{\text{shared}}/(\lambda^2_{\text{shared}}+\lambda^2_{\text{ind}})$, ranging from almost independent motion (**Fig. 5l**, left) to almost perfect correlation (right). According to the network model, $\boldsymbol{\lambda}^2$ can be read out linearly. For the simulation in **Fig. 5m**, we presented the network with trials of seven values of q. We then trained a linear regression model to predict q from the neural activity for the two most extreme structures (blue dots in **Fig. 5m**), and decoded q for the intermediate structures using this regression model (red dots in **Fig. 5m**). Owing to the stochastic stimulus generation, the network's motion structure estimates, $\boldsymbol{\lambda}_t$, fluctuate around the true strength—yet, on average, the trained linear readout correctly identified the fraction of global motion in the stimulus. This is a strong prediction of the network model, which could be tested in a targeted neuroscientific experiment.

Discussion

We have proposed a comprehensive theory of online hierarchical inference for structured visual motion perception. The derived algorithm decomposes an incoming stream of retinal velocities into latent motion components which in turn are organized in a nested, tree-like structure. A scene's inferred structure provides the visual system with a temporally robust scaffold to organize its percepts and to resolve momentary ambiguities in the input stream. Applying the theory to human visual motion perception, we replicated diverse phenomena from psychophysics in both object-indexed and location-indexed experiment designs. Furthermore, inspection of the algorithm's internal variables provided normative explanations for putative origins of human percepts and spawned concrete predictions for new psychophysics experiments. Finally, the algorithm afforded a recurrent neural network model with visual inputs reminiscent of cortical area MT and latent structure representations reminiscent of area MSTd.

In the main text, we have for the sake of clarity limited the presentation of the theory to a basic version that nonetheless covers all essential concepts. In *Supplemental Information, Section 4*, we present several extensions that are naturally covered by the algorithm: (i) observation noise, σ_{obs}, can be time- and object-dependent, which is relevant for modeling temporary occlusion of a subset of stimuli; (ii) observation noise can be non-isotropic (different values in x- and y-direction), which is relevant for angle-dependent edge velocities in apertures [55]; (iii) for optimal inference, different motion components can feature different time constants, since velocity is expected to change more slowly for heavy objects due to higher inertia; (iv) different motion components may tend to co-occur or exclude one another in real-world scenes, which can be modeled by an interaction prior of pairwise component compatibility; and (v) when motion components are not present for a long time, they will decay to zero, preventing their rediscovery, which can be mitigated by a prior on motion strengths.

The current theory is limited to velocities as input, thereby ignoring the well-documented influence of spatial arrangement on visual motion perception, such as center-surround modulation [56, 57], adjacency [26] or motion assimilation [58]. Furthermore, the algorithm does not solve the correspondence problem in object-indexed experiments, but simply assumes that velocities are correctly assigned to the input vector as objects move about the visual field. Our work focuses on the simultaneous inference of motion sources, $\boldsymbol{s}_t$, and motion strengths, $\boldsymbol{\lambda}_t$. Other quantities, such as time constants and, probably more importantly, the motion components, $\boldsymbol{C}$, have been assumed to be given. It is worth noting, however, that gradient-based learning of $\boldsymbol{C}$ is, in principle, supported by the theory on long time scales (see *Supplemental Information, Section 4.5*). Finally, limited experimental evidence of the neural correlates of motion structure perception required the neural network model to rely on many modeling assumptions. The model's predictions should act as a starting point for further scientific inquiry of these neural correlates.

Even though the sensory processes underlying object-indexed motion perception necessarily differ from those of location-index perception, the online inference algorithm describes human perception for both types of experiments. Thus, both types might share the same underlying neural mechanisms for structure inference. This raises the intriguing question whether there exist stable, object-bound neural representations of velocity. Furthermore, our work points towards a tight link between neural representations of latent structure and representations of uncertainty in that the estimated motion strengths, $\boldsymbol{\lambda}_t$, determine the credit assignment of prediction errors through the gating function, $f_\Sigma(\boldsymbol{\lambda}_t^2)$—a function that also computes the variance of motion components, e.g., the brain's uncertainty about flock velocity. Behaviorally, sensory noise directly impacts the perceived structure of a scene as demonstrated experimentally

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

by the perceptual reversal in the motion illusion of Lorenceau [38] (cf. **Fig. 4h–k**). More generally, our theory predicts that the visual system will organize its percepts into simpler structures when sensory reliability decreases. Moreover, the reliability of visual cues plays a role in multisensory integration [59], with area MSTd [60, 61], but not area MT [62], exhibiting tuning to vestibular signals. Thus, MSTd may be a candidate area for multisensory motion structure inference. Overall, we expect our theoretical results to guide targeted experiments in order to understand structured visual motion perception under a normative account of statistical information processing.

Methods

In what follows, we provide an overview of the generative model, the online inference algorithm, the computer simulations, and the data analysis. A more detailed presentation is found in the *Supplemental Information*.

Generative model of structured motion. We consider K observable velocities, $v_{k,d}(t)$, in D spatial dimensions. For notational clarity, we will consider in this *Online Methods* section only the case $D=1$ and use the vector notation, $\boldsymbol{v}_t=(v_1(t),..,v_K(t))^\mathsf{T}$. The extension to $D>1$ is covered in *Supplemental Information, Section 2.4*. Observable velocities, $\boldsymbol{v}_t$, are generated by M latent motion sources, $s_{m,d}(t)$, abbreviated (for $D=1$) by the vector $\boldsymbol{s}_t=(s_1(t),..,s_M(t))^\mathsf{T}$. Velocities are noisy instantiations of their combined ancestral motion sources, $\boldsymbol{v}_t \sim \mathcal{N}\left(\boldsymbol{C}\,\boldsymbol{s}_t, \sigma^2_{\text{obs}}/\delta t\,\boldsymbol{I}\right)$, where $C_{km}=+1, -1$, and 0 in $K\times M$ component matrix, $\boldsymbol{C}$, denote positive, negative and absent influence, respectively. For the formal definition, observations, $\boldsymbol{v}_t$, remain stable within a short time interval $[t, t+\delta t)$, and the observation noise variance, $\sigma^2_{\text{obs}}/\delta t$, ensures a δt-independent information content of the input stream. In the online inference algorithm, below, we will draw the continuous-time limit, which will become independent of δt. In computer simulations, δt is the inverse frame rate of the motion display (default value: $1/\delta t=60\,\text{Hz}$). Each motion source (in each spatial dimension) follows an Ornstein-Uhlenbeck process, $\mathrm{d}s_m = -s_m/\tau_s\,\mathrm{d}t + \lambda_m\,\mathrm{d}W_m$, with time constant τ_s, motion strength λ_m (shared across dimensions), and Wiener process W_m. The resulting marginal stationary velocity distribution of v_k is $v_k \sim \mathcal{N}\left(0, \sigma^2_{\text{obs}}/\delta t + \frac{\tau_s}{2}\sum_{m=1}^{M} C^2_{km}\,\lambda^2_m\right)$.

Radial and rotational motion sources. In location-indexed experiments, the input's location (e.g., a neuron's receptive field) remains fixed. For $D=2$, input v_k has fixed polar coordinates (R_k, ϑ_k) with radial distance R_k and angle ϑ_k, relative to the pivot point. Denoting radial and rotational motion sources by s_r and s_φ, we obtain for the noise-free part of v_k in Cartesian coordinates: $v_{k,x} = s_r\cos\vartheta_k - s_\varphi R_k \sin\vartheta_k$, and $v_{k,y} = s_r\sin\vartheta_k + s_\varphi R_k\cos\vartheta_k$. Since R_k and ϑ_k are fixed coefficients, the mapping $(s_r, s_\varphi)\mapsto(v_{k,x}, v_{k,y})$ is linear and, thus, can be described by the component matrix $\boldsymbol{C}$. The full derivation and an illustration of the velocity relations in polar coordinates are provided in *Supplemental Information, Section 2.5*.

Online inference algorithm. The goal of motion structure inference is to simultaneously infer the value of motion sources, $\boldsymbol{s}_t$, and the underlying structure, $\boldsymbol{\lambda}$, from a stream of velocity observations. The number of spatial dimensions, D, component matrix, $\boldsymbol{C}$, time constant τ_s, and observation noise σ_{obs} are assumed to be known. The EM algorithm leverages that changes in $\boldsymbol{s}_t$ and $\boldsymbol{\lambda}$ (if changing at all) occur on different time scales, τ_s and τ_λ, respectively. For $\tau_\lambda \gg \tau_s$, the EM algorithm treats $\boldsymbol{\lambda}$ as a constant for inferring $\boldsymbol{s}_t$ (E-step), and optimizes an estimate, $\boldsymbol{\lambda}_t$, online based on the inferred motion strengths (M-step).

E-Step. For fixed $\boldsymbol{\lambda}$, the posterior $p(\,\boldsymbol{s}_t\,|\,\boldsymbol{v}_{0:t};\,\boldsymbol{\lambda}\,)$ is always a multivariate normal distribution, $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, and can be calculated by a Kalman-Bucy filter [63, 64]; see *Supplemental Information, Section 3.1, 3.2, and 3.3.1* for the derivation. This yields coupled differential equations for the time evolution of $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$. To reduce the computational complexity of the system, we perform an adiabatic approximation on the posterior covariance, $\boldsymbol{\Sigma}_t$, by assuming (a) that it has always converged to its stationary value, and (b) that off-diagonal values in $\boldsymbol{\Sigma}_t$ are zero, that is, we ignore correlations in uncertainty about latent motion sources in the posterior distribution. As shown in the full derivation in *Supplemental Information, Section 3.3*, the first assumption is warranted because the stationary value of $\boldsymbol{\Sigma}_t$ depends only on the current structure estimate, $\boldsymbol{\lambda}_t$; then, because $\boldsymbol{\Sigma}_t$ decays to stationarity at time scale $\tau_s/2$, it can always follow any changes in $\boldsymbol{\lambda}_t$ which happen at time scale $\tau_\lambda \gg \tau_s$. The second assumption is a modeling assumption: that biological agents might disregard the subtle (and complicated) interactions between the uncertainties of different motion sources and rely on their individual uncertainties, instead. Using the two assumptions we derive a closed-form solution for the posterior variance,

$$\Sigma_{mm} = \frac{\sigma^2_{\text{obs}}}{\tau_s\,\|\boldsymbol{c}_m\|^2}\left(-1+\sqrt{1+\frac{{\tau_s}^2\,\|\boldsymbol{c}_m\|^2}{\sigma^2_{\text{obs}}}\,\lambda^2_m}\right) =: f_\Sigma(\lambda^2_m)\ , \tag{5}$$

with $\|\boldsymbol{c}_m\|^2 = \sum_{k=1}^{K} C^2_{km}$ denoting the vector-norm of the m-th column of $\boldsymbol{C}$. This is eqn. (3) of the main text. The plot in **Fig. 1i** has parameters $\|\boldsymbol{c}_m\|^2=4$, $\tau_s=300\,\text{ms}$, and $\sigma_{\text{obs}}=0.05$. By plugging the adiabatic approximation of the variance

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

into the time evolution of $\boldsymbol{\mu}_t$, we arrive at eqn. (2) of the main text (see *Supplemental Information, Section 3.3.4* for the derivation).

M-step. Using the posterior from the E-step, motion strengths, $\boldsymbol{\lambda}$, are optimized to maximize the likelihood of the observed velocities. This optimization further incorporates prior distributions, $p(\lambda_m^2)$, most conveniently formulated over the squared motion strengths, for which we employ a scaled inverse chi-squared distribution,

$$p(\lambda_m^2) = \mathcal{I}\chi(\lambda_m^2; \nu_m, \kappa_m^2) = \frac{1}{\lambda_m^{(2+\nu_m)}} \exp\left[-\frac{\nu_m \kappa_m^2}{2\lambda_m^2} - A(\nu_m, \kappa_m^2)\right] , \tag{6}$$

owing to its conjugacy to estimating the variance of s_m (this is what λ_m^2 controls). The prior features two hyper-parameters, ν_m and κ_m^2, which give rise to an intuitive interpretation as ν_m pseudo-observations of average value κ_m^2. The partition function, $A(\nu_m, \kappa_m^2)$, only serves for normalization. By default, we employ a Jeffreys prior ($\nu_m = \kappa_m^2 = 0$), which is a typical choice as a non-informative prior in Bayesian statistics and promotes a preference for finding simple structures by assigning higher beliefs to small values of λ_m (and highest to $\lambda_m = 0$). The only exception is the motion strength assigned to self motion, λ_{self}, for which we employ a uniform prior distribution, formally by setting $\nu_{\text{self}} = -2$ and $\kappa_{\text{self}}^2 = 0$. These choices reflect the a-priori belief that motion components supported by $\boldsymbol{C}$ will usually be absent or small in any given scene—with the exception of self-motion-induced velocity on the retina, which occurs with every saccade and every turn of the agent's head (see *Supplemental Information, Section 3.1.2* for the formal calculation of the M-step).

In the online formulation of EM (see *Supplemental Information, Section 3.2.3 and 3.3.4* for the derivation of the online EM algorithm and of the proposed adiabatic inference algorithm, respectively), these priors give rise to the low-pass filtering dynamics in eqn. (1) for updating λ_m^2, with constants

$$\alpha_m = \frac{2}{\tau_s{}^2 \left(2 + \nu_m + \tau_\lambda / \tau_s\right)} \text{ , and} \tag{7}$$

$$\beta_m = \frac{\nu_m \kappa_m^2}{\tau_\lambda \left(2 + \nu_m + \tau_\lambda / \tau_s\right)} . \tag{8}$$

This completes the derivation of the online algorithm for $D = 1$ spatial dimensions. The extension to multiple dimensions is straightforward and provided in *Supplemental Information, Section 3.1.3, 3.2.3 and 3.3.4* alongside the respective derivations.

Preference for simple structures. The above Jeffreys prior on motion strengths, $p(\lambda_m^2)$, facilitates the discovery of sparse structures. This property is important when a large "reservoir" of possible motion components in $\boldsymbol{C}$ is considered: the algorithm will recruit only a small number of components from the reservoir. In *Supplemental Figure S2*, we demonstrate this ability for the example of the Johansson experiment from **Fig. 2b–d** by duplicating the shared motion component, i.e., the first two columns in $\boldsymbol{C}$ are all 1's. As *Supplemental Figure S2* shows, the online algorithm recruits only one of the two identical components and discards the other. This example of identical components in the reservoir represents the theoretically hardest scenario for maintaining a sparse structure.

Computer simulations. Computer simulations and data analysis were performed with custom Python code (Python 3.8, Numpy 1.21, Scipy 1.7, scikit-learn 0.24, Matplotlib 3.4, Pandas 1.3, xarray 0.19). The code package supports most of the extensions presented in *Supplemental Information, Section 4*.

For the numerical simulation, input was drawn with observation noise variance $\sigma_{\text{obs}}^2 / \delta t$, at the time points of input frames (every δt). The drawn input remained stable until the next frame. Between frames, the differential equations for online hierarchical inference were integrated with SciPy's explicit Runge-Kutta method "RK45" which adapts the step size. This integration method combines numerical accuracy with a parameterization that is almost invariant to the input frame rate. The default parameters that we used are listed in **Table 1**.

Hierarchical motion experiments (Fig. 2). For the Johansson experiment, all $K = 3$ dots followed sinusoidal velocities with frequency 0.5 Hz. Horizontal amplitudes were $2\sqrt{\tau_s}$ for all dots; vertical amplitudes were 0 for the outer dots and $\cos(45°) \cdot 2\sqrt{\tau_s}$ for the inner dot. For the Duncker wheel, we set the wheel radius to $R = 1$ and the rotation frequency to 1 Hz. This leads to the hub velocity $v_{\text{hub,y}} = 0$ and $v_{\text{hub,x}} = 2\pi\,\text{s}^{-1}$ because the hub must travel $2\pi R$ during one period for slip-free rolling. For the rim velocities, being the derivatives of location, we thus find $v_{\text{rim,x}} = v_{\text{hub,x}} + R\omega\cos(\omega t)$ and $v_{\text{rim,y}} = -R\omega\sin(\omega t)$, with $\omega = 2\pi\,\text{s}^{-1}$. For the simulation, we increased the observation noise to $\sigma_{\text{obs}} = 0.15$ and set $\lambda_m(t=0) = 0.1$ to highlight the gradual discovery of the motion components.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Description	Variable	Object-indexed	Location-indexed	Network
Time const. motion sources	τ_s	0.300 s	0.100 s	0.100 s
Time const. motion strengths	τ_λ	1.000 s	0.333 s	0.333 s
Inv. observation frame rate	δt	1/60 s	1/60 s	1/120 s
Observation noise	σ_{obs}	0.05	$0.017 = 0.05 \div 3$	0.017
Initial motion strength	$\lambda_m(t{=}0)$	0.5	0.5	0.5
No. of pseudo observation	ν_m	0	0 / −1	0
Val. of pseudo observations	κ_m	0	0	0
Vestibular input	v_{vst}	–	0	–
Obs. noise for vestibular input	σ_{vst}	–	0.05	–
Time const. for pred. err.	τ_ϵ	–	–	0.050 s

Table 1. Default parameters of the computer simulations. Most parameters are maintained throughout all computer experiments. Deviations from these parameters are listed in the respective experiment description. The value $\nu_m = -1$ in location-indexed experiments relates to self-motion. For $D{=}2$ spatial dimensions, $\nu_{\text{self}} = -2/D = -1$ yields a uniform prior distribution (see *Supplemental Information*, *Section 3.1.3*).

Structure classification (Fig. 3). The stimulus data and human responses were released by [17] on GitHub. The experiment is described in detail in [17]. There were 12 participants with each participant performing 200 trials. Each trial consisted of three dots moving on a circle for 4 s. Dots had different colors to prevent their confusion, but colors did not convey any information on the dots' roles within the structure. No data was excluded. Trials were generated stochastically from the same generative model that is considered in this work, with uniform probability for each of the four structures (Independent, Global, Clustered, Hierarchical) to underlie the trial. Motion strengths were chosen such that all dots had identical marginal velocity distributions, $p(v_k)$, across all structures—leaving motion relations as the only distinguishing information (see [17], for detailed stimulus parameters and λ-values of all structures). Like [17], we treated the experiment as one-dimensional ($D{=}1$), operating directly on the angular velocities. Noise-free angular velocities were calculated from the circular distance of subsequent stimulus frames, and we set $1/\delta t{=}50\,\text{Hz}$ to match the experiment's frame rate.

For presenting the trials to our inference algorithm, we initialized each of the λ_m at its average value (average taken across the ground truth of all structures). At trial end, the inference algorithm yielded $M{=}7$-dimensional $\boldsymbol{\lambda}$-vectors associated with 1 shared component, 3 cluster components (one per possible pair), and 3 individual components (see *Supplemental Figure S3* for example trials). For logistic regression, we calculated 5 features, T_i, from $\boldsymbol{\lambda}$, namely:

$$T_1 = \lambda_1 \Big/ \sum_i \lambda_i \qquad \text{“Does shared motion stand out?”} \tag{9}$$

$$T_2 = \max\{\lambda_2,\lambda_3,\lambda_4\} \Big/ \sum_{m=2,3,4} \lambda_m \qquad \text{“Does one cluster dominate the others?”}$$

$$T_3 = \max\{\lambda_5,\lambda_6,\lambda_7\} \Big/ \sum_{m=5,6,7} \lambda_m \qquad \text{“Does one individual component stand out?”}$$

$$T_4 = \lambda_c^2 \Big/ \sum_{m=c,\text{Ch}_1(c),\text{Ch}_2(c)} \lambda_m^2 \quad \text{with } c = \text{argmax}(\lambda_2,\lambda_3,\lambda_4) \qquad \text{“Does the strongest cluster dominate its children?”}$$

$$T_5 = \lambda_c^2 \Big/ \sum_{m=c,\neg\text{Ch}(c)} \lambda_m^2 \quad \text{with } c = \text{argmax}(\lambda_2,\lambda_3,\lambda_4) \qquad \text{“Does the strongest cluster dominate the 3rd dot?”}$$

Here, $\text{Ch}_{1,2}(c)$ denote the individual motion components of the two dots within the cluster component c, and $\neg\text{Ch}(c)$ denotes the dot not being in cluster c. The features were hand-designed based on the reasoning that they may be useful for structure classification. Their most important property is that all information about a trial is conveyed through $\boldsymbol{\lambda}$ as a bottleneck. A multinomial logistic regression classifier was trained with L1-regularization on the feature vectors, $(T_1,..,T_5)$, to classify the ground truth structures of the trials. Then, we fitted the same choice model as [17] to the human choices, but replaced the ideal observer log-probability, $\log p(S \,|\, \boldsymbol{v}_{0:T})$, which was used in [17], with the class probability from the trained classifier, $\log p(S \,|\, \boldsymbol{\lambda})$:

$$P(\text{choice}{=}S) = \pi_{\text{L}} \frac{1}{4} + (1-\pi_{\text{L}}) \exp\left[\beta \left(\log p(S \,|\, \boldsymbol{\lambda}) + b_S\right)\right] \Big/ \text{Norm.}\,, \tag{10}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

with lapse probability, π_{L}, inverse temperature, β, and biases, b_S, for all structures, $S = G, C, H$, relative to the independent structure ($b_I = 0$ by convention). Note that, in contrast to [17], we do not need to consider structure multiplicities here because the features are already symmetric with regard to the three possible cluster assignments. Like [17], we did not apply observation noise to the presented velocities, but maintained a non-zero observation noise parameter, σ_{obs}, for the inference algorithm. Observation noise, σ_{obs}, and lapse probability, π_{L}, were shared parameters for all participants and were fitted jointly via a simple grid search. We obtained $\sigma_{\mathrm{obs}} = 0.04$ and $\pi_{\mathrm{L}} = 4\%$ (compared to 14% in [17]). The remaining 4 parameters, $\{\beta, b_G, b_C, b_H\}$, were fitted via maximum likelihood for each participant. All reported confusion matrices and log-likelihoods were obtained by fitting the 4 per-participant parameters using leave-one-out cross-validation. The log-chance level in **Fig. 3f** is $200 \cdot \log(1/4)$ since each participant performed 200 trials.

Location-indexed experiments (Fig. 4). To support self-motion, we introduce a column of −1's in $\boldsymbol{C}$ as an additional component, which is connected to all visual velocity inputs and to a vestibular input v_{vst}. In our simulations, the vestibular input is always stationary, but noisy: $v_{\mathrm{vst}} \sim \mathcal{N}\left(0, \sigma_{\mathrm{vst}}^2\right)$. The associated self-motion strength, λ_{self}, uses a uniform prior (see discussion under eqn. (6)). Perceived velocities are the sum over all-except-self-motion: $\boldsymbol{v}_{\mathrm{perceived}} = \sum_{m \neq \mathrm{self}} \boldsymbol{C}_{*m} \boldsymbol{\mu}_m$.

In the RDK experiments, input was modeled as $K = 3$ velocities: two for the two groups of dots, plus the vestibular input. Repulsion angles were estimated from 20 repetitions of 30 s long trials, with $\boldsymbol{v}_{\mathrm{perceived}}$ averaged over the last 10 s of each trial. Error bars from the simulations were too small to be shown in **Fig. 4e–g**.

In **Fig. 4e**, the velocities for opening angle, γ, were given by $(v_x, v_y) = v_0 \cdot (\cos(\gamma/2), \sin(\gamma/2))$ for the first group, with $v_0 = 2\sqrt{\tau_s}$, and $v_0 \cdot (\cos(\gamma/2), -\sin(\gamma/2))$ for the second group. As in Figure 3 of [36], the repulsion bias was measured with respect to the full opening angle.

In **Fig. 4f**, increasing contrast of the second group was modeled as dividing the observation noise variance by a factor, f, between 0.001 and 10, leading to variance $\sigma_{\mathrm{obs}}^2 / f$ for this group's input. As in [37], the repulsion bias was measured only with respect to the first group's perceived direction. The expressed similarity to experimental data refers to the "2-motion condition" in Figure 7 of [37].

In **Fig. 4g**, the velocity of the second group was multiplied by a factor between 0 and 2, and the repulsion bias was measured only with respect to the first group's perceived direction. A direct quantitative comparison to the human data from Figure 4b in [36] is difficult because they had measured the point of subjective equality (PSE) to a 90° opening angle for this stimulus condition, finding a 10° bias for the full opening angle.

For the Lorenceau illusion in **Fig. 4h–k**, we modeled each dot's velocity as a separate input owing to the spatially distributed nature of the stimulus. As in [38], the two groups of 10 dots each oscillated at a frequency of 0.83 Hz. For the oscillation amplitude, we chose $R = 1/2$ (arbitrary units), leading to velocities $v_x(t) = R\,\omega \cos(\omega\, t)$ for the horizontal group and $v_y(t) = -R\,\omega \sin(\omega\, t)$ for the vertical group, with $\omega = 2\pi \cdot 0.83\,\mathrm{s}^{-1}$. As shown in *Supplemental Figure S5*, the inference algorithm decomposes this stimulus into a deeply nested structure comprising self-, shared-, group-, and individual motion. For the noise-free stimulus condition, we used the default simulation parameters. For the condition with motion noise, the observation noise, σ_{obs}, of the visual inputs (not the vestibular input) was multiplied by 25.

Network implementation (Fig. 5). A detailed derivation of how to implement the online inference algorithm in a neural network model is provided in *Supplemental Information, Section 5*. In the following, we will focus on the specific model parameters used in the simulations of **Fig. 5**.

For both simulations (the demonstration in **Fig. 5e–j** and the proposed experiment in **Fig. 5k–m**), there were $K = 6$ location-indexed input variables in $D = 2$ spatial dimensions. Input was encoded according to the model of area MT presented in *Supplemental Information, Section 5.6*. Each velocity, $\boldsymbol{v}_k$, was encoded by a population of 192 neurons, with tuning centers organized on a polar grid with $N_\alpha = 16$ preferred directions, and $N_\rho = 12$ preferred speeds (sketched in **Fig. 5c** for smaller values of N_α and N_ρ). Each neuron in each of the K populations thus has "coordinates" (n_α, n_ρ) describing its preferred direction and speed. To account for the reported bias of MT tuning toward slow speed [52], the density of preferred speeds became sparser for higher speeds, which we modeled in *Supplemental Information, eqn. (70)* by $\mu_\rho(n_\rho) = \rho_{\min} + d_\rho\, n_\rho^{1.25}$, with $d_\rho = (\rho_{\max} - \rho_{\min})/(N_\rho - 1)^{1.25}$, and $\rho_{\min} = 0.1$, $\rho_{\max} = 8.0$, for neurons $n_\rho = 0, .., N_\rho - 1$. Preferred directions covered the circle equidistantly. The remaining parameters in the tuning function were $\kappa_\alpha = 1/0.35^2$ and $\sigma_\rho^2 = 0.35^2$ for the angular and radial tuning widths, respectively, and $\psi = 0.1$ Hz for the overall firing rate scaling factor. For the network simulations, we increased the frame rate to $\delta t = 1/120$ Hz for the sake of a higher sampling rate on the x-axis in **Fig. 5h–j** (the simulation software stores firing rates only at the time of frames).

The distributed population comprised 100 neurons. Readout vectors, $\boldsymbol{a}_x$, for all variables represented by this

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

population were drawn i.i.d. from a standard normal distribution, $\mathcal{N}(0,1)$, for each vector element. Adjoint matrices were calculated numerically to fulfill the required orthonormality conditions (see *Supplemental Information, Section 5.4*). The low-pass filtering time constant of the prediction error was $\tau_e = \tau_s/2 = 0.050\,\mathrm{s}$, such that the prediction error could react to changes in $\boldsymbol{\mu}_t$.

The one-to-one population comprised M neurons (or small populations; $M=8$ for the demo, and $M=7$ for the proposed experiment), one per function value, $f_\Sigma(\lambda_m^2)$. The proportionality constant for the readout was $f_\Sigma(\lambda_m^2) = 0.001\, r_{\text{1-to-1},m}$.

Given the parameters and decoding vectors, the simulation software automatically transforms the differential equations of the online algorithm into the corresponding neural dynamics, according to the rules stated in *Supplemental Information, Section 5.4*. Numerical integration of neural dynamics was performed by the same "RK45" method used in the previous simulations.

For the demonstration in **Fig. 5e–j**, inputs were arranged on a ring of radius $R_k = 1$ with angular location $\vartheta_k = 60^\circ \cdot k$ (measured from the x-axis in counter-clockwise direction). Presented velocities were $(v_x, v_y) = (-2\sin(\vartheta_k), 2\cos(\vartheta_k))$ for $t \leq 1\,\mathrm{s}$, $(2\sin(\vartheta_k), -2\cos(\vartheta_k))$ for $1\,\mathrm{s} < t \leq 2\,\mathrm{s}$, and $(2+2\sin(\vartheta_k), -2\cos(\vartheta_k))$ for $2\,\mathrm{s} < t$. In the $\boldsymbol{C}$-matrix underlying the network construction, the shared polar visual component was constructed according to *Supplemental Information, eqn. (6)*. The shared translational and the 6 individual components were Cartesian (as in all other experiments).

In the proposed experiment in **Fig. 5k–m**, all motion components and the input were "Cartesian" such that input location played no role (formally, we maintained the circular arrangement of the previous network simulation). Input was generated from the inference algorithm's underlying generative model for a motion tree comprising 1 shared component and 6 individual components. For a given fraction of shared motion, q, we set $\lambda^2_{\text{shared}} = 2^2\, q$ and $\lambda^2_{\text{ind}} = 2^2\,(1-q)$. Maintaining constant total squared motion strength, $\lambda^2_{\text{shared}} + \lambda^2_{\text{ind}} = 4$, ensures that the (marginal) input velocity distributions are statistically identical across all input locations and all values of q. In total, seven fractions, $q = 1/8, 2/8, .., 7/8$, of shared motion were presented. Per simulation run, each fraction was presented for 10 s, and simulations were repeated for 10 runs. For the subsequent data analysis, the neural responses of only the 2nd half ($5\,\mathrm{s} \leq t$) of the stimulus presentation were considered to avoid potential initial transients. A standard linear regression model (with intercept; `class sklearn.linear_model.LinearRegression`) was trained to decode the correct q from the distributed population's response, r_{dis}, for the fractions $q = 1/8$ and $q = 7/8$. The resulting linear readout (with intercept) was employed to decode q from r_{dis} for the remaining stimuli in **Fig. 5m**.

Acknowledgments

We thank Anna Kutschireiter for valuable discussions and feedback on the theory.

Funding Information

This research was supported by grants from the Harvard Brain Science Initiative (Collaborative Seed Grant, J.B., J.D. & S.J.G.), the Center for Brains, Minds, and Machines (CBMM; funded by NSF STC award CCF-1231216), and a James S. McDonnell Foundation Scholar Award for Understanding Human Cognition (Grant 220020462, J.B. & J.D.).

Data availability

No new data was produced for this study. The behavioral data from [17] is available with the original publication. The behavioral data for [36] has been digitized by the authors and is included in the software repository.

Code availability

Computer simulations, data analyses and visualization have been performed with custom Python code which will be released in a public repository upon acceptance of the manuscript.

Competing interests

The authors declare no competing interests.

Author contributions

J.B., S.J.G., and J.D. conceived the study; J.B. developed the theory; J.B. performed the computer simulations; J.B. and J.D. analyzed and discussed the data; J.B., S.J.G., and J.D. wrote the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

[1] Daniel Kaiser et al. "Object vision in a structured world". In: *Trends in cognitive sciences* (2019).

[2] Steven Yantis. "Multielement visual tracking: Attention and perceptual organization". In: *Cognitive psychology* (1992).

[3] Jon Driver, Peter McLeod, and Zoltan Dienes. "Motion coherence and conjunction search: Implications for guided search theory". In: *Perception & Psychophysics* (1992).

[4] Constance S Royden and Ellen C Hildreth. "Human heading judgments in the presence of moving objects". In: *Perception & psychophysics* (1996).

[5] Geniva Liu et al. "Multiple-object tracking is based on scene, not retinal, coordinates." In: *Journal of Experimental Psychology: Human Perception and Performance* (2005).

[6] Haokui Xu et al. "Seeing "what" through "why": Evidence from probing the causal structure of hierarchical motion." In: *Journal of Experimental Psychology: General* (2017).

[7] Kalpana Dokka et al. "Causal inference accounts for heading perception in the presence of object motion". In: *Proceedings of the National Academy of Sciences* (2019).

[8] Andrew D Bolton et al. "Elements of a stochastic 3D prediction engine in larval zebrafish prey capture". In: *ELife* (2019).

[9] Yair Weiss, Eero P Simoncelli, and Edward H Adelson. "Motion illusions as optimal percepts". In: *Nature neuroscience* (2002).

[10] Alan A. Stocker and Eero P. Simoncelli. "Noise characteristics and prior expectations in human visual speed perception". In: *Nature Neuroscience* (Apr. 2006). Number: 4 Publisher: Nature Publishing Group.

[11] Alan A Stocker and Eero P Simoncelli. "Sensory adaptation within a Bayesian framework for perception". In: *Advances in neural information processing systems*. Citeseer. 2006.

[12] Andrew E. Welchman, Judith M. Lam, and Heinrich H. Bülthoff. "Bayesian motion estimation accounts for a surprising bias in 3D vision". In: *Proceedings of the National Academy of Sciences* (2008).

[13] Edward Vul et al. "Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model". In: *Advances in neural information processing systems* (2009).

[14] James H Hedges, Alan A Stocker, and Eero P Simoncelli. "Optimal inference explains the perceptual coherence of visual motion stimuli". In: *Journal of vision* (2011).

[15] Samuel J Gershman, Joshua B Tenenbaum, and Frank Jäkel. "Discovering hierarchical motion structure". In: *Vision research* (2016).

[16] Johannes Bill et al. "Hierarchical structure is employed by humans during visual motion perception". In: *Proceedings of the National Academy of Sciences* (2020).

[17] Sichao Yang et al. "Human visual motion perception shows hallmarks of Bayesian structural inference". In: *Scientific reports* (2021).

[18] HB Barlow and William R Levick. "The mechanism of directionally selective units in rabbit's retina." In: *The Journal of physiology* (1965).

[19] Michael Steven Graziano, Richard A Andersen, and Robert J Snowden. "Tuning of MST neurons to spiral motions". In: *Journal of Neuroscience* (1994).

[20] Christopher C Pack et al. "End-stopping and the aperture problem: two-dimensional motion signals in macaque V1". In: *Neuron* (2003).

[21] Richard T Born and David C Bradley. "Structure and function of visual area MT". In: *Annu. Rev. Neurosci.* (2005).

[22] Patrick J. Mineault et al. "Hierarchical processing of complex motion along the primate dorsal visual pathway". In: *Proceedings of the National Academy of Sciences* (2012).

[23] Kang Li et al. "Neurons in primate visual cortex alternate between responses to multiple stimuli in their receptive field". In: *Frontiers in Computational Neuroscience* (2016).

[24] Max Wertheimer. "Laws of organization in perceptual forms". In: *A Sourcebook of Gestalt Psychology*. Ed. by W.B. Ellis. Harcourt, Brace, 1938.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[25] Gunnar Johansson. "Visual perception of biological motion and a model for its analysis". In: *Perception & psychophysics* (1973).

[26] Walter C. Gogel. "Relative motion and the adjacency principle". In: *Quarterly Journal of Experimental Psychology* (1974).

[27] Stephen Grossberg, Jasmin Léveillé, and Massimiliano Versace. "How do object reference frames and motion vector decomposition emerge in laminar cortical circuits?" In: *Attention, Perception, & Psychophysics* (2011).

[28] Elizabeth S Spelke. "Principles of object perception". In: *Cognitive Science* (1990).

[29] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* (1977).

[30] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[31] Olivier Cappé and Eric Moulines. "On-line expectation–maximization algorithm for latent data models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2009).

[32] Keiji Tanaka, Yoshiro Fukada, and Ha Saito. "Underlying mechanisms of the response specificity of expansion/contraction and rotation cells in the dorsal part of the medial superior temporal area of the macaque monkey". In: *Journal of neurophysiology* (1989).

[33] Jonathan I Flombaum and Brian J Scholl. "A Temporal Same-Object Advantage in the Tunnel Effect: Facilitated Change Detection for Persisting Objects". In: *Journal of Experimental Psychology: Human Perception and Performance* (2006).

[34] Crispin Gardiner. *Stochastic methods*. Springer Berlin, 2009.

[35] Karl Duncker. "Über induzierte Bewegung". In: *Psychologische Forschung* (1929).

[36] Oliver J Braddick, Keith A Wishart, and William Curran. "Directional performance in motion transparency". In: *Vision research* (2002).

[37] Yuzhi Chen et al. "Effects of attention on motion repulsion". In: *Vision research* (2005).

[38] Jean Lorenceau. "Motion integration with dot patterns: effects of motion noise and structural information". In: *Vision Research* (1996).

[39] William Marshak and Robert Sekuler. "Mutual repulsion between moving visual targets". In: *Science* (1979).

[40] Jeounghoon Kim and Hugh R Wilson. "Direction repulsion between components in motion transparency". In: *Vision Research* (1996).

[41] Jessica N Cali, Patrick J Bennett, and Allison B Sekuler. "Phase integration bias in a motion grouping task". In: *Journal of Vision* (2020).

[42] Th Brandt, Johannes Dichgans, and Ellen Koenig. "Differential effects of central versus peripheral vision on egocentric and exocentric motion perception". In: *Experimental brain research* (1973).

[43] Dora E Angelaki, Yong Gu, and Gregory C DeAngelis. "Visual and vestibular cue integration for heading perception in extrastriate visual cortex". In: *The Journal of physiology* (2011).

[44] Kaoru Amano, Brian A Wandell, and Serge O Dumoulin. "Visual field maps, population receptive field sizes, and visual field coverage in the human MT+ complex". In: *Journal of neurophysiology* (2009).

[45] Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. "Marginalization in neural circuits with divisive normalization". In: *Journal of Neuroscience* (2011).

[46] Emilio Salinas and Laurence F Abbott. "A model of multiplicative neural responses in parietal cortex". In: *Proceedings of the national academy of sciences* (1996).

[47] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series, 2001.

[48] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[49] Hidehiko Komatsu and Robert H Wurtz. "Relation of cortical areas MT and MST to pursuit eye movements. I. Localization and visual properties of neurons". In: *Journal of neurophysiology* (1988).

[50] Charles J Duffy and Robert H Wurtz. "Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli". In: *Journal of neurophysiology* (1991).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.21.465346; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[51] Gregory C DeAngelis and Takanori Uka. "Coding of horizontal disparity and velocity by MT neurons in the alert macaque". In: *Journal of neurophysiology* (2003).

[52] Harris Nover, Charles H Anderson, and Gregory C DeAngelis. "A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance". In: *Journal of Neuroscience* (2005).

[53] Adam Kohn and J Anthony Movshon. "Adaptation changes the direction tuning of macaque MT neurons". In: *Nature neuroscience* (2004).

[54] Bart Krekelberg, Richard JA Van Wezel, and Thomas D Albright. "Interactions between speed and contrast tuning in the middle temporal area: implications for the neural code for speed". In: *Journal of Neuroscience* (2006).

[55] Eric Castet et al. "Perceived speed of moving lines depends on orientation, length, speed and luminance". In: *Vision Research* (1993).

[56] John Allman, Francis Miezin, and EveLynn McGuinness. "Direction-and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT)". In: *Perception* (1985).

[57] Xin Huang, Thomas D Albright, and Gene R Stoner. "Stimulus dependency and mechanisms of surround modulation in cortical area MT". In: *Journal of Neuroscience* (2008).

[58] Mark Nawrot and Robert Sekuler. "Assimilation and contrast in motion perception: Explorations in cooperativity". In: *Vision research* (1990).

[59] Dora E Angelaki, Yong Gu, and Gregory C DeAngelis. "Multisensory integration: psychophysics, neurophysiology, and computation". In: *Current opinion in neurobiology* (2009).

[60] Katsumasa Takahashi et al. "Multimodal coding of three-dimensional rotation and translation in area MSTd: comparison of visual and vestibular selectivity". In: *Journal of Neuroscience* (2007).

[61] Jocelyne Ventre-Dominey. "Vestibular function in the temporal and parietal cortex: distinct velocity and inertial processing pathways". In: *Frontiers in integrative neuroscience* (2014).

[62] Syed A Chowdhury et al. "Does the middle temporal area carry vestibular signals related to self-motion?" In: *Journal of Neuroscience* (2009).

[63] Rudolph E Kalman and Richard S Bucy. "New results in linear filtering and prediction theory". In: *Journal of Basic Engineering* (1961).

[64] Anna Kutschireiter, Simone Carlo Surace, and Jean-Pascal Pfister. "The Hitchhiker's guide to nonlinear filtering". In: *Journal of Mathematical Psychology* (2020).