Annual Review of Vision Science

Hierarchical Vector Analysis of Visual Motion Perception

Samuel J. Gershman,[1] Johannes Bill,[2] and Jan Drugowitsch[2]

[1]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, Massachusetts, USA; email: gershman@fas.harvard.edu

[2]Department of Neurobiology, Harvard Medical School, Boston, Massachusetts, USA

Annu. Rev. Vis. Sci. 2025. 11:411–22

First published as a Review in Advance on March 31, 2025

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-110323-031344

Keywords

motion perception, Bayesian inference, structure learning

Abstract

Visual scenes are often populated by densely layered and complex patterns of motion. The problem of motion parsing is to break down these patterns into simpler components that are meaningful for perception and action. Psychophysical evidence suggests that the brain decomposes motion patterns into a hierarchy of relative motion vectors. Recent computational models have shed light on the algorithmic and neural basis of this parsing strategy. We review these models and the experiments that were designed to test their predictions. Zooming out, we argue that hierarchical motion perception is a tractable model system for understanding how aspects of high-level cognition such as compositionality may be implemented in neural circuitry.

1. INTRODUCTION

Watching birds flocking or cars driving on the highway, one may notice that there is a hidden simplicity underlying the tangle of different motion directions: Some of the complexity can be subtracted away due to the fact that the birds tend to flock in the same direction and the cars tend to drive in the same direction. Our visual systems are attuned to both the global motion direction of the group and the local motion direction of the elements. This hierarchical structure seems intuitive, but discovering it from visual input is computationally nontrivial. The goal of this review is to summarize recent advances in our understanding of how the brain solves the motion parsing problem.

The foundations for contemporary work on this problem were laid by Gunnar Johansson in his 1950 dissertation, *Configurations in Event Perception: An Experimental Study* (collected in Johansson 1994). Johansson used simple dot configurations to reveal a hierarchical parsing strategy, which he called perceptual vector analysis. We summarize some of these studies in Section 2. Importantly, this early work left open exactly how the visual system selects an appropriate parse, since most dot configurations admit multiple possible parses. This question was addressed directly in computational and experimental work over the last decade. The key idea, which we cover in Section 3, is to formulate perceptual vector analysis as a form of probabilistic inference over tree-structured motion representations. This can be implemented with a circuit model employing linear and quadratic interactions between motion-sensitive neurons.

The study of hierarchical motion perception is important for several reasons. One reason is that it brings us closer to understanding how motion perception works in complex scenes. Much of what we know about motion perception comes from relatively simple displays (e.g., gratings or fields of coherently moving dots) with one or sometimes two motion components. The models that have been developed to explain these phenomena therefore lack generality. Another reason is that it may aid us in the quest to understand the neural basis of high-level cognition by capturing a form of compositionality within a tractable neural system. We return to this point in Section 5.

2. CLASSIC STUDIES OF HIERARCHICAL MOTION PERCEPTION

Johansson designed many ingenious moving dot configurations; without covering these exhaustively, we can convey his general approach in **Figure 1**. The central dot, flanked by two

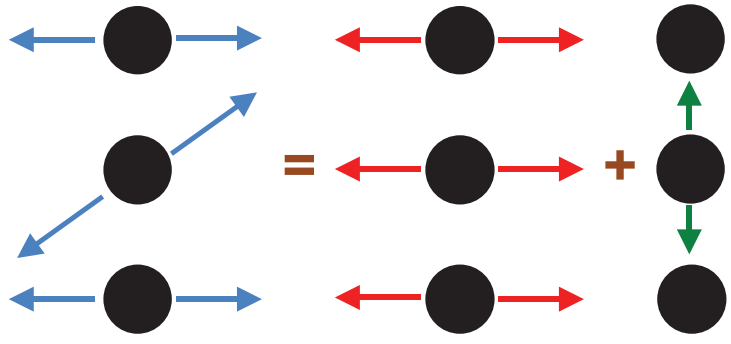

Figure 1

Perceptual vector analysis. A three-dot configuration created by Gunnar Johansson (1994). The arrows show motion directions. When the bottom and top dots oscillate horizontally and the central dot oscillates diagonally, human observers perceive the central dot oscillating vertically within a horizontally oscillating reference frame. Johansson proposed that the visual system decomposes the observed configuration into a sum of underlying vectors, as shown here.

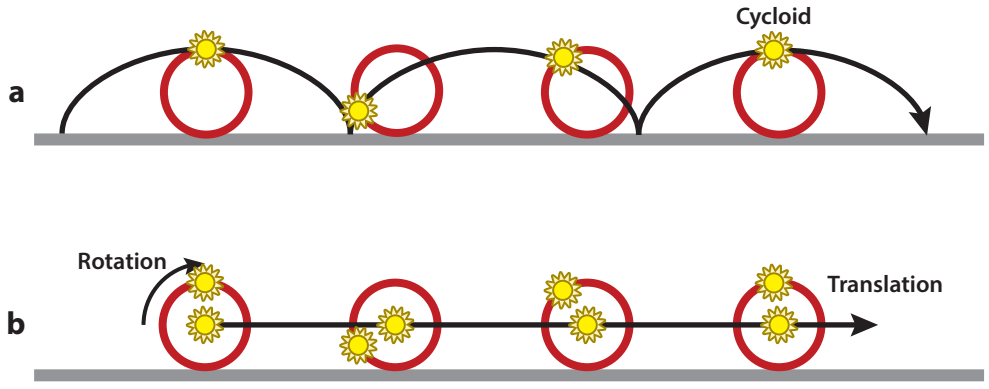

Figure 2

The Duncker wheel. (*a*) When viewed in darkness, a light attached to the rim of a wheel appears to move cycloidally. (*b*) Adding a light to the hub makes the rim light appear to rotate around a horizontally translating reference frame.

horizontally oscillating dots, oscillates diagonally so that all three dots always maintain collinearity vertically. Observers typically perceive the central dot as oscillating vertically, not diagonally, nested within a horizontally oscillating reference frame. Johansson argued that this perceptual phenomenon arose because the observer's visual system decomposed the configural motion into the sum of two vectors (one horizontal and one vertical).

Another classic example, the Duncker wheel (Duncker 1929), is shown in **Figure 2**. In this case, the same dot can appear as either moving cycloidally or rotating around a horizontally translating reference frame, depending on the absence or presence of a central horizontally translating dot. Duncker provided other examples of what he called induced motion (context-induced changes in perceived motion). In fact, discussion of induced motion goes all the way back to Ptolemy in the second century AD, who noted that a stationary boat in a moving river may be perceived as a moving boat in a stationary river (Smith 1996).

Loomis & Nakayama (1973) reported another striking example of induced motion, simultaneous velocity contrast: Two dots moving at the same speed will appear to be moving at different speeds if one dot is surrounded by slowly moving dots (and therefore perceived as moving faster) while the other dot is surrounded by quickly moving dots (and therefore perceived as moving more slowly). Anyone who has been stuck driving in a slow lane next to a fast lane will have a vivid appreciation of this phenomenon.

Our final example is the phenomenon of motion transparency (**Figure 3**). When two fields of moving dots are superimposed, they will appear to be moving together (along the average direction) provided the directional difference is small; if the directional difference is sufficiently large, two transparent fields are perceived as simultaneously moving in different directions, sometimes with a repulsive bias (Braddick et al. 2002).

The common insight running through these examples is that the visual system can parse the same motion patterns in different ways depending on the context. The parses may be simple segregation (as in simultaneous velocity contrast and motion transparency) or hierarchical organization (as in Johansson's triplet and the Duncker wheel). Johansson's proposal that this is accomplished through a perceptual vector analysis was pivotal but informal nonetheless. In the absence of a formal model, it is unclear how the brain resolves the inherent ambiguity of the parsing problem: Any given motion configuration may admit multiple vector decompositions. A number of researchers have attempted to specify additional principles that the visual system may use to partially resolve

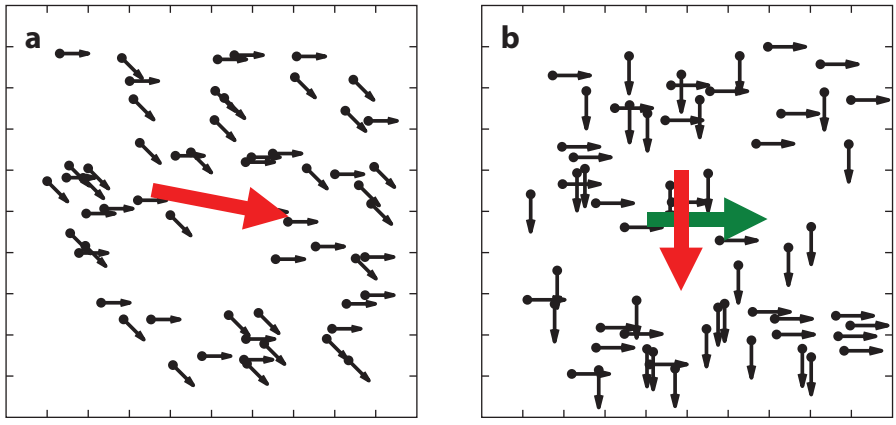

Figure 3

Motion transparency. (*a*) When the directional difference between two superimposed motion fields is small, the dots appear to be moving along the average direction. (*b*) When the directional difference is sufficiently large, the two fields are perceptually segregated. Arrows show the direction of perceived motion.

this ambiguity (e.g., Gogel 1974, Restle 1979, DiVita & Rock 1997). For example, Gogel (1974) proposed that the motion parse is determined by relative motion cues between nearby points (the adjacency principle). DiVita & Rock (1997) proposed that the motion parse is determined by the perceived coplanarity of objects and their potential reference frames (the belongingness principle). While there is empirical evidence for all of these proposals, they are somewhat ad hoc and do not provide a comprehensive account of all the relevant phenomena. In the next section, we describe a modern approach to this problem grounded in probabilistic inference.

3. MODERN THEORIES AND NEW EXPERIMENTAL TESTS

Computational models of perception have evolved considerably since Johansson's time. A large class of models has been derived from the unifying idea that the brain probabilistically inverts the causes of its sensations (Knill & Richards 1996, Yuille & Kersten 2006). The inversion is accomplished by applying Bayes' rule,

$$P(\text{cause}|\text{sensation}) \propto P(\text{sensation}|\text{cause})P(\text{cause}). \quad 1.$$

Bayes' rule stipulates how an observer should update their prior belief, $P(\text{cause})$, to their posterior belief, $P(\text{cause}|\text{sensation})$. The evidence driving belief updating is the likelihood $P(\text{sensation}|\text{cause})$, which quantifies the match between sensory data and hypothetical hidden causes.

To apply Bayes' rule, one needs to specify what sensory data one is conditioning on and what the hypothesis space is, and then one needs to define the prior and likelihood. Weiss et al. (2002) proposed an influential Bayesian model of motion perception, where the sensory data are image time series (restricted to the surface of a rigidly moving object) and the hypothesis space is the object's velocity vector. The prior assumes that objects tend to move slowly, and the likelihood assumes that changes in image intensity are approximately linear functions of velocity, corrupted by Gaussian sensory noise. Despite its simplicity, this model was able to capture several important perceptual phenomena, such as the effects of contrast on the perception of speed and direction.

The model of Weiss et al. (2002) served as the starting point for a generalization to superpositions of motion vectors—what Gershman et al. (2016) dubbed Bayesian vector analysis. The key

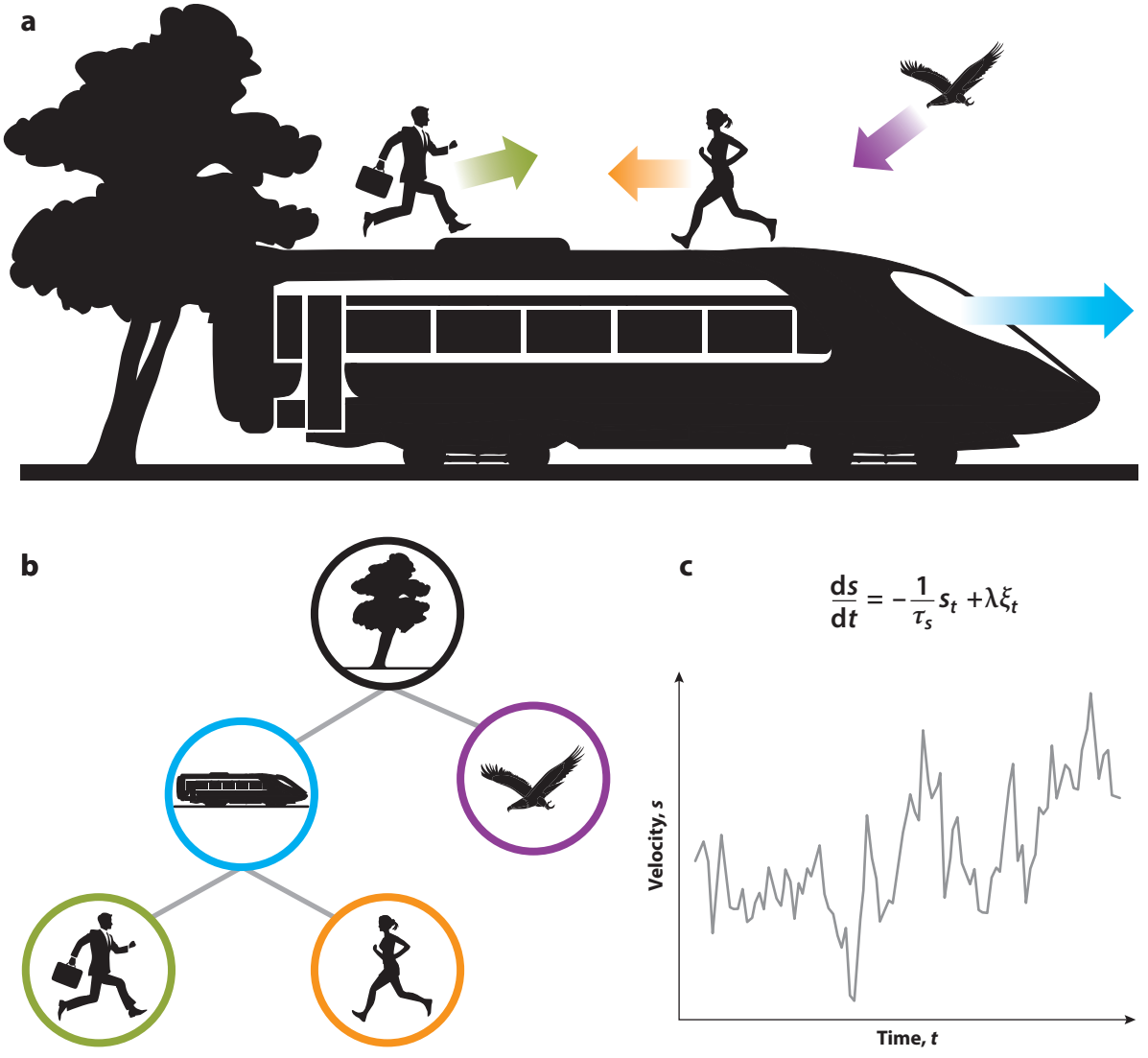

Figure 4

Hierarchical motion parsing framework. (*a*) A complex moving scene. Arrows show motion sources defining relative motion components. For example, the train is moving to the right relative to the background. (*b*) Motion tree representation, where each node corresponds to a source and edges denote hierarchical relationships between sources. The observed motion of an object is the sum of source velocities contributing to that object, following a path from the root node to the node corresponding to the object. (*c*) A source velocity s_t is assumed to change slowly over time t. In this example, the velocity time series follows an Ornstein–Uhlenbeck process with timescale τ_s and motion strength λ modulating a Gaussian perturbation ξ_t.

idea, illustrated in **Figure 4**, is to define the hypothesis space as the space of motion trees, where each node is a motion vector (or more generally a spatial vector field) and each edge corresponds to a hierarchical relationship. The observed motion of an object or part is the sum of motion vectors along the path from the root node (the background) to the relevant object node. Each motion vector is parameterized by a direction and magnitude (motion strength); a strength of 0 means that a particular component is absent.

An example is shown **Figure 4*a***. The train is moving to the right relative to a stationary background, and the female figure is moving to the left relative to the train. The total observed motion for the female figure is thus equal to the sum of the train and female figure relative to motion vectors. This representation enables the representation of complex motions in terms of compositions of simpler motions. In some variations of the model (Bill et al. 2020, 2022), the velocity vectors may be time varying (**Figure 4*c***).

The inference problem for hierarchical motion is significantly more complex than the problem in Weiss et al. (2002), because the visual system needs to jointly infer the structure of the motion configuration (the motion tree) and the motion vector for each node. To constrain the vast hypothesis space, Gershman et al. (2016) posited a prior over motion trees that favors smaller trees while still allowing trees with unbounded depth and width (Blei et al. 2010). More complex trees are favored only when they match the observed motion patterns strongly. This prior can be understood as formalizing the Gestalt law of Prägnanz (Koffka 1935) for motion patterns: Good patterns are patterns that can be described with a small motion tree.

Structural inference allows the model to explain transitions between different structural interpretations as a function of motion evidence, for example, the transition from averaging to transparency as a function of directional difference between overlaid dot fields (Braddick et al. 2002). Using more complex configurations consisting of five dots, Gershman et al. (2016) demonstrated that the model could accurately predict human judgments about hierarchical relationships between dots (e.g., Is the blue dot moving relative to the red dot?), capturing over 96% of the variance on held-out trials after fitting the model on a separate set of trials.

It is important to emphasize here that a hierarchical representation of motion is central to understanding the relevant perceptual phenomena. One might be tempted to substitute a less structured decomposition method, such as principal component analysis or nonnegative matrix factorization, to the observed motion patterns. Indeed, some prior proposals (Beyeler et al. 2016, Chen et al. 2022) have posited that cortical motion processing implements nonnegative matrix factorization. While there is certainly merit to such proposals, they cannot explain how human subjects can report the hierarchical relationships between dot motions as in Gershman et al. (2016); the models simply do not represent such relationships.

Yang et al. (2021) investigated structural inference using a more rigorously controlled stimulus design (**Figure 5*a***; adapted from Bill et al. 2020, which we discuss below). By displaying dot motions on a circular track, the joint distribution of object velocities could be precisely matched for different motion structures. This was important for eliminating several potential confounds, such as the spatial arrangement and low-level stimulus statistics, which provide information about stimulus velocity and identity. Subjects were first trained to recognize four different motion structures (**Figure 5*b***). On each trial, subjects viewed a display and judged which structure generated the display, followed by feedback with the correct structure label. Subjects then completed a series of new test trials. Yang et al. (2021) showed that a Bayesian vector analysis model could accurately predict human structural inferences, including for highly ambiguous displays. An example fit between model and data is shown in **Figure 5*c***. The same model could also predict human confidence reports (modeled as the posterior probability of their structure judgment).

Other studies have focused on how people exploit hierarchical motion structure to guide other aspects of object perception. Using the circular motion display described above, Bill et al. (2020) showed that motion structure strongly constrains both multiple object tracking and motion prediction. Human performance was much better when presented with hierarchically structured displays than with independently moving dots (for converging evidence with a noncircular display, see also Xu et al. 2017). These performance improvements could be captured by a Bayesian vector analysis model equipped with the correct motion structures but not by a model that assumed

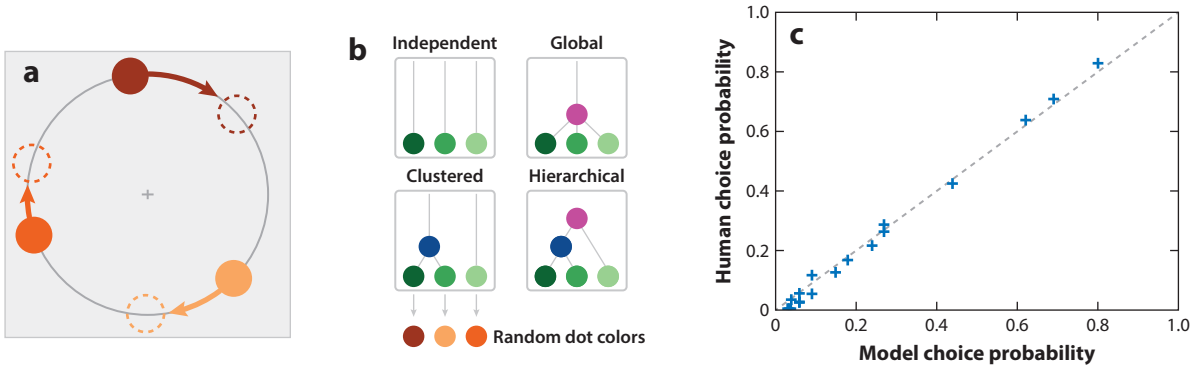

Figure 5

Circular motion display. (*a*) Dots rotate stochastically around a circle. This display was used in different ways by Bill et al. (2020) and Yang et al. (2021) to study hierarchical motion perception. (*b*) Motion trees underlying the generation of dot motions. (*c*) Human choice probabilities closely match model predictions from Bill et al. (2022), $R^2 = 0.99$. The data are taken from Yang et al. (2021), experiment 1. Each cross corresponds to the conditional probability of choosing one of the four motion structures conditional on a particular ground truth structure. Panels *a* and *b* adapted from Bill et al. (2020) (CC BY 4.0).

independent motion. Importantly, performance improvement is accompanied by systematic errors in the direction of relative motion, as explored in detail by Xu et al. (2017); the hierarchical structure provides an inductive bias that both is useful and results in systematic errors.

Shivkumar et al. (2023) have provided another line of evidence using random dot kinematograms (RDKs) arranged into concentric rings. This kind of display is appealing because it builds on a substantial body of psychophysical and neurophysiological work using RDKs, making it a useful tool for pursuing the neural basis of hierarchical motion perception (see Section 4). With a two-level hierarchy (central RDK plus one surrounding ring), the perceived motion direction for the center dots was biased toward the surround when the difference was small and repulsed when the difference was large. This result is consistent with a Bayesian vector analysis model that infers a single motion source until the evidence for a multisource structure is strong enough. The repulsive effect arose from the inference that both the center and the surround moved relative to a global motion source (such as self-motion). This implies that observers should additionally use a global motion source (perceived or implicit) in their inferred motion tree, which should reduce the perceived velocity of the components; as far as we know, whether such a global motion source was perceived was not directly tested in this study. Shivkumar et al. (2023) also studied a more complex display with two outer rings, allowing them to show that the visual system can select different reference frames for the central motion depending on the directional differences between the center and the two surrounds.

The studies reviewed so far have relied on relatively simple stimuli with a small number of motion components. Natural scenes can be much more complex, and therefore it is necessary to consider the generality of theoretical claims based on simple stimuli. Yang et al. (2023) investigated motion perception in naturalistic movies by asking subjects to adjust the direction and speed of a matching noise stimulus so that it coincided with the flow at a probed location in the movie frame. They found that subjects often produced flow judgments that were consistent with the ground truth, but sometimes made errors consistent with vector analysis, such as perceiving illusory object motion induced by background motion.

An important question raised by these studies is how the brain could plausibly carry out the required computations in an efficient, online manner. This question was addressed by Bill et al. (2022), who derived an online expectation–maximization algorithm that inferred both the motion

tree and the velocity vectors. This model could capture both the classic experimental results reviewed above and the more recent results with the circular motion display. Bill et al. (2022) also showed how the algorithm could be implemented in a biologically plausible recurrent neural network with linear and quadratic interactions between neurons. In the next section, we discuss which brain circuits might implement the proposed architecture.

4. IN SEARCH OF NEURAL MECHANISMS

The problem of parsing multiple motions has long exercised computational neuroscientists. This work has focused mostly on two cases: (*a*) parsing multiple object motions under transparency and occlusion and (*b*) parsing self-motion and object motion. We discuss each case in turn and then speculate about their implications for the neural implementation of hierarchical vector analysis.

The earliest cortical stages of motion processing are in primary visual cortex (V1) followed by the middle temporal area (MT, also known as V5). In response to a transparent motion stimulus (two superimposed fields of dots moving in different directions, as in **Figure 3**), direction-selective V1 neurons are activated by their preferred direction regardless of transparency, whereas MT neurons, also activated by their preferred direction, are additionally suppressed by transparency (Snowden et al. 1991). A study by Qian & Andersen (1994) took advantage of the fact that locally pairing dots moving in opposite directions eliminates the perception of motion transparency (Qian et al. 1994a); they found that V1 neurons do not discriminate between transparent (unpaired) and nontransparent (paired) conditions, whereas MT neurons do. These findings are consistent with models in which V1 neurons extract local motion energy signals, which are then converted into relative local velocity signals in MT via subtractive or divisive inhibition (Qian et al. 1994b, Nowlan & Sejnowski 1995, Simoncelli & Heeger 1998, Koechlin et al. 1999).

Later motion processing stages in cortex are involved in the extraction of additional information from local velocity signals. In particular, the medial superior temporal (MST) area is divided principally into a lateroventral subdivision (MSTl), responsible for maintaining motion-dependent smooth pursuit eye movements (Dürsteler & Wurtz 1988), and a dorsomedial subdivision (MSTd), responsible for extraction of heading direction from optic flow and vestibular signals (Britten & van Wezel 1998; Page & Duffy 2003; Gu et al. 2007, 2010). We focus on MSTd, as its functions are the most relevant to questions concerning motion vector analysis. The core problem, as pointed out by Zemel & Sejnowski (1998), is that naturalistic optic flow patterns are complex combinations of observer self-motion with the motions of multiple objects. Zemel & Sejnowski proposed that MSTd solves this parsing problem by estimating each object's motion relative to the observer. The inputs to their model are local velocity signals computed by MT. MSTd was modeled as an autoencoder, trained to reconstruct the inputs passed through a hidden layer bottleneck. By combining sparsity regularization of the hidden layer with divisive normalization in the output layer, the model produces outputs that reflect assignments of local velocity signals to motion sources (for a related approach based on nonnegative matrix factorization, see also Beyeler et al. 2016).

While the model of Zemel & Sejnowski (1998) is feedforward, Layton & Fajen (2016) have proposed a model in which feedback connections from MST to MT help MT correct its initial velocity estimates. Anatomical evidence for this feedback pathway has been reported by Maunsell & van Essen (1983), but its functional implications remain relatively unexplored. A related approach by Vafaii et al. (2023) trained autoencoders on the flow fields underlying motion scenes, similar to Zemel & Sejnowski. However, rather than relating the network's latent activations to MSTd, they showed that MT activity could be predicted accurately without modeling feedback from MST.

Grossberg et al. (2011) integrated several of these ideas to develop a model of hierarchical motion parsing in the MT/MST circuit (for similar ideas, primarily emphasizing Gestalt grouping principles, see also He & Öğmen 2023). This model is speculative because there are no neurophysiological studies of the Johansson and Duncker displays simulated by Grossberg and colleagues or of the related displays described in Section 3. Hopefully, such studies will be undertaken in the future, but in the meantime it is useful to think about what we might expect based on existing data and theoretical ideas. According to this model, MT carries out two processing stages (in different cortical layers): boundary selection of motion in depth, followed by long-range motion grouping. The second stage interacts recurrently with a competitive directional grouping process in MST. Feedback from MST to MT amplifies the activity of MT cells coding the winning direction and suppresses the activity of other cells.

How do these previous models connect to the model proposed by Bill et al. (2022)? In many respects the models are quite different, but they are all broadly compatible with the following architecture: A late motion processing stage (putatively MST) operates on a local velocity estimate (putatively MT) to parse motion sources via some form of competitive grouping. For the Bill et al. model, these local velocity estimates correspond either to stationary motion patches, compatible with MT's retinotopic architecture, or to moving objects that might change location within the visual scene, for which brain areas other than MT (not yet identified) might be involved. One hint of object-centered reference frames can be found in Olson & Gettner (1995), which describes neurons in the supplementary eye field (a high-order oculomotor control region) that fired selectively when monkeys saccaded to the left or right end of an object, regardless of that object's position in the visual field. There is also abundant evidence for object-centered position representations in parietal cortex (e.g., Committeri et al. 2004, Chafee et al. 2007, Crowe et al. 2008). It is thus plausible that neurons exist that track changes in these position representations across time.

The Bill et al. (2022) model assumes that task-relevant variables are encoded in a population code that can be decoded by linear readouts. Evidence for linear decodability of object motion and self-motion in MSTd (Sasaki et al. 2017) is suggestive of the possibility that this area implements the form of hierarchical vector analysis posited by Bill et al. Future experimental work is needed to directly test this hypothesis.

5. CONCLUSIONS

Our review documents a range of empirical evidence for the hierarchical nature of motion perception. Recent theoretical work has shown that a fully formalized version of Johansson's perceptual vector analysis hypothesis can capture many of these empirical phenomena. This has also led to new experimental tests of theoretical predictions. The general picture that has emerged from this line of research is that the visual system parses complex motion patterns into compositions of simpler components. Ambiguity about the appropriate parse is resolved through Bayesian inference, which represents a distribution over possible parses. This distribution can be efficiently approximated using a biologically plausible neural circuit. Whether the brain actually uses such a circuit for hierarchical motion perception remains an open question.

Another exciting direction of research is the design of more humanlike artificial systems for computer vision. Recent work, partly inspired by studies of humans, has shown that such systems can be trained to extract part-based decompositions from videos and to use these decompositions to solve challenging tasks such as synthesizing future frames (Pérez-Rúa et al. 2016, Xu et al. 2019).

We conclude with some reflections on the broader scope of this research area. Hierarchical motion perception is important to understand on its own terms, but we also believe that these studies tell us something about a more general strategy used by the brain and how it might be

realized neurally. Hierarchical structure shows up in many other domains: language into phrase structure, plans into subplans, narratives into events, objects into parts. The ability to discover and represent such structure is often considered a hallmark of high-level cognition (Fodor 1975, Tenenbaum et al. 2011, Gershman 2021). Yet the neural mechanisms underlying this competence remain elusive, in large part because we lack adequate understanding of how the basic components (primitives) are represented. As a consequence, even the most comprehensive attempts to address this question (e.g., Van der Velde & De Kamps 2006, Eliasmith 2013) remain speculative. In contrast, the neural mechanisms underlying visual motion perception have been studied for decades, supplying a set of primitives grounded in specific brain regions, cell types, and circuit mechanisms (Albright & Stoner 1995, Andersen 1997, Clark & Demb 2016, Nishida et al. 2018). This opens the door to using motion perception as a tractable model system for elucidating how aspects of high-level cognition are implemented in the brain.

SUMMARY POINTS

1. The brain parses complex moving scenes into hierarchically organized relative motion components.
2. Probabilistic inference can be used to solve the motion parsing problem and can be implemented in biologically plausible neural circuits similar to those found in cortical motion processing areas.

FUTURE ISSUES

1. Hierarchical motion parsing assumes that motion is represented in object-centered coordinates, but little is known about such representations in the brain.
2. How can hierarchical motion parsing, operating on pixel inputs, scale to natural movies?
3. How can we apply the lessons from the study of motion perception to other neural systems (e.g., language, concept learning) where compositional computations are thought to be implemented?

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Some of the work described here was supported by a seed grant from the Harvard Brain Science Initiative. S.J.G. is supported by the Kempner Institute for the Study of Natural and Artificial Intelligence and by a Schmidt Science Polymath Award. J.D. is supported by a National Institutes of Health/National Institute of Neurological Disorders and Stroke grant from the BRAIN Initiative (U19NS118246).

LITERATURE CITED

Albright TD, Stoner GR. 1995. Visual motion perception. *PNAS* 92:2433–40

Andersen RA. 1997. Neural mechanisms of visual motion perception in primates. *Neuron* 18:865–72

Beyeler M, Dutt N, Krichmar JL. 2016. 3D visual response properties of MSTd emerge from an efficient, sparse population code. *J. Neurosci.* 36:8399–415
Bill J, Gershman SJ, Drugowitsch J. 2022. Visual motion perception as online hierarchical inference. *Nat. Commun.* 13:7403
Bill J, Pailian H, Gershman SJ, Drugowitsch J. 2020. Hierarchical structure is employed by humans during visual motion perception. *PNAS* 117:24581–89
Blei D, Griffiths T, Jordan M. 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. Account. Comput. Mach.* 57:1–30
Braddick O, Wishart K, Curran W. 2002. Directional performance in motion transparency. *Vis. Res.* 42:1237–48
Britten KH, van Wezel RJ. 1998. Electrical microstimulation of cortical area MST biases heading perception in monkeys. *Nat. Neurosci.* 1:59–63
Chafee MV, Averbeck BB, Crowe DA. 2007. Representing spatial relationships in posterior parietal cortex: Single neurons code object-referenced position. *Cereb. Cortex* 17:2914–32
Chen K, Beyeler M, Krichmar JL. 2022. Cortical motion perception emerges from dimensionality reduction with evolved spike-timing-dependent plasticity rules. *J. Neurosci.* 42:5882–98
Clark DA, Demb JB. 2016. Parallel computations in insect and mammalian visual motion processing. *Curr. Biol.* 26:R1062–72
Committeri G, Galati G, Paradis AL, Pizzamiglio L, Berthoz A, LeBihan D. 2004. Reference frames for spatial cognition: Different brain areas are involved in viewer-, object-, and landmark-centered judgments about object location. *J. Cogn. Neurosci.* 16:1517–35
Crowe DA, Averbeck BB, Chafee MV. 2008. Neural ensemble decoding reveals a correlate of viewer- to object-centered spatial transformation in monkey parietal cortex. *J. Neurosci.* 28:5218–28
DiVita JC, Rock I. 1997. A belongingness principle of motion perception. *J. Exp. Psychol. Hum. Percept. Perform.* 23:1343–52
Duncker K. 1929. Uber inducierte bewegung [about induced motion]. *Psychol. Forsch.* 12:180–259
Dürsteler M, Wurtz R. 1988. Pursuit and optokinetic deficits following chemical lesions of cortical areas MT and MST. *J. Neurophysiol.* 60:940–65
Eliasmith C. 2013. *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford Univ. Press
Fodor JA. 1975. *The Language of Thought*. Harvard Univ. Press
Gershman SJ. 2021. *What Makes Us Smart: The Computational Logic of Human Cognition*. Princeton Univ. Press
Gershman SJ, Tenenbaum JB, Jäkel F. 2016. Discovering hierarchical motion structure. *Vis. Res.* 126:232–41
Gogel WC. 1974. Relative motion and the adjacency principle. *Q. J. Exp. Psychol.* 26:425–37
Grossberg S, Léveillé J, Versace M. 2011. How do object reference frames and motion vector decomposition emerge in laminar cortical circuits? *Atten. Percept. Psychophys.* 73:1147–70
Gu Y, DeAngelis GC, Angelaki DE. 2007. A functional link between area MSTd and heading perception based on vestibular signals. *Nat. Neurosci.* 10:1038–47
Gu Y, Fetsch CR, Adeyemo B, DeAngelis GC, Angelaki DE. 2010. Decoding of MSTd population activity accounts for variations in the precision of heading perception. *Neuron* 66:596–609
He D, Öğmen H. 2023. A neural model for vector decomposition and relative-motion perception. *Vis. Res.* 202:108142
Johansson G. 1994. *Perceiving Events and Objects*. Lawrence Erlbaum Assoc.
Knill D, Richards W. 1996. *Perception as Bayesian Inference*. Cambridge Univ. Press
Koechlin E, Anton JL, Burnod Y. 1999. Bayesian inference in populations of cortical neurons: a model of motion integration and segmentation in area MT. *Biol. Cybernet.* 80:25–44
Koffka K. 1935. *Principles of Gestalt Psychology*. Harcourt, Brace
Layton OW, Fajen BR. 2016. A neural model of MST and MT explains perceived object motion during self-motion. *J. Neurosci.* 36:8093–102
Loomis J, Nakayama K. 1973. A velocity analogue of brightness contrast. *Perception* 2:425–27
Maunsell J, van Essen DC. 1983. The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J. Neurosci.* 3:2563–86
Nishida S, Kawabe T, Sawayama M, Fukiage T. 2018. Motion perception: from detection to interpretation. *Annu. Rev. Vis. Sci.* 4:501–23

Nowlan SJ, Sejnowski TJ. 1995. A selection model for motion processing in area MT of primates. *J. Neurosci.* 15:1195–214

Olson CR, Gettner SN. 1995. Object-centered direction selectivity in the macaque supplementary eye field. *Science* 269:985–88

Page WK, Duffy CJ. 2003. Heading representation in MST: sensory interactions and population encoding. *J. Neurophysiol.* 89:1994–2013

Pérez-Rúa JM, Crivelli T, Pérez P, Bouthemy P. 2016. Discovering motion hierarchies via tree-structured coding of trajectories. In *Proceedings of the British Machine Vision Conference 2016*, ed. RC Wilson, ER Hancock, WAP Smith. BMVC Press

Qian N, Andersen RA. 1994. Transparent motion perception as detection of unbalanced motion signals. II. Physiology. *J. Neurosci.* 14:7367–80

Qian N, Andersen RA, Adelson EH. 1994a. Transparent motion perception as detection of unbalanced motion signals. I. Psychophysics. *J. Neurosci.* 14:7357–66

Qian N, Andersen RA, Adelson EH. 1994b. Transparent motion perception as detection of unbalanced motion signals. III. Modeling. *J. Neurosci.* 14:7381–92

Restle F. 1979. Coding theory of the perception of motion configurations. *Psychol. Rev.* 86:1–24

Sasaki R, Angelaki DE, DeAngelis GC. 2017. Dissociation of self-motion and object motion by linear population decoding that approximates marginalization. *J. Neurosci.* 37:11204–19

Shivkumar S, DeAngelis GC, Haefner RM. 2023. Hierarchical motion perception as causal inference. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2023.11.18.567582v1**

Simoncelli EP, Heeger DJ. 1998. A model of neuronal responses in visual area MT. *Vis. Res.* 38:743–61

Smith AM. 1996. *Ptolemy's Theory of Visual Perception: An English Translation of the* Optics *with Introduction and Commentary*. Am. Philos. Soc. Press

Snowden RJ, Treue S, Erickson RG, Andersen RA. 1991. The response of area MT and V1 neurons to transparent motion. *J. Neurosci.* 11(9):2768–85

Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. 2011. How to grow a mind: statistics, structure, and abstraction. *Science* 331:1279–85

Vafaii H, Yates JL, Butts DA. 2023. Hierarchical VAEs provide a normative account of motion processing in the primate brain. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2023.09.27.559646v2.full**

Van der Velde F, De Kamps M. 2006. Neural blackboard architectures of combinatorial structures in cognition. *Behav. Brain Sci.* 29:37–70

Weiss Y, Simoncelli E, Adelson E. 2002. Motion illusions as optimal percepts. *Nat. Neurosci.* 5:598–604

Xu H, Tang N, Zhou J, Shen M, Gao T. 2017. Seeing "what" through "why": evidence from probing the causal structure of hierarchical motion. *J. Exp. Psychol. Gen.* 146:896–909

Xu Z, Liu Z, Sun C, Murphy K, Freeman WT, et al. 2019. Unsupervised discovery of parts, structure, and dynamics. Preprint, arXiv:1903.05136 [cs.CV]

Yang S, Bill J, Drugowitsch J, Gershman SJ. 2021. Human visual motion perception shows hallmarks of Bayesian structural inference. *Sci. Rep.* 11:3714

Yang YH, Fukiage T, Sun Z, Nishida S. 2023. Psychophysical measurement of perceived motion flow of naturalistic scenes. *iScience* 26:108307

Yuille A, Kersten D. 2006. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* 10:301–8

Zemel RS, Sejnowski TJ. 1998. A model for encoding multiple object motions and self-motion in area MST of primate visual cortex. *J. Neurosci.* 18:531–47