Journal Club

Editor's Note: These short reviews of recent *JNeurosci* articles, written exclusively by students or postdoctoral fellows, summarize the important findings of the paper and provide additional insight and commentary. If the authors of the highlighted article have written a response to the Journal Club, the response can be found by viewing the Journal Club at www.jneurosci.org. For more information on the format, review process, and purpose of Journal Club articles, please see http://jneurosci.org/content/jneurosci-journal-club.

A Normative Account of 3D Motion Perception from Natural Scenes

Zhe-Xin Xu
Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115
Review of Herrera-Esposito and Burge

Two monks were watching a banner fluttering in the wind.

One said, "The wind is moving."

The other said, "The banner is moving."

They argued back and forth, unable to agree.

Huineng approached and said, "It is neither the wind that moves, nor the banner that moves. It is your mind that moves."

—Platform Sutra

Perception is not a passive reflection of reality, but an active inference. Our visual system receives 2D images on the retinae and must infer the underlying 3D layout of the environment. This inference is challenging because the same retinal image can be produced by many distinct 3D structures. To resolve this ambiguity, the brain uses multiple depth cues, such as binocular disparity, motion parallax, occlusion, blur, and perspective distortion (Howard and Rogers, 2008). The neural mechanisms underlying some of these cues are well studied, while others remain less clear (Howard and Rogers, 2008; Anzai and DeAngelis, 2010).

Received May 2, 2025; revised June 18, 2025; accepted June 19, 2025.
The author declares no competing financial interests.
Correspondence should be addressed to Zhe-Xin Xu at brian_xu@hms.harvard.edu.
https://doi.org/10.1523/JNEUROSCI.0866-25.2025
Copyright © 2025 the authors

Beyond estimating static 3D structures, the ability to detect movement in the environment is also critical for survival. Many species rely on motion perception to navigate their environment, avoid predators, and capture prey. Motion in the fronto-parallel plane—i.e., 2D motion—has been extensively studied (Park and Tadin, 2018). These studies have revealed that in the primate brain, object motion processing begins in the primary visual cortex (V1) and proceeds along the dorsal visual pathway, including the middle temporal (MT) and medial superior temporal (MST) areas (Park and Tadin, 2018). Less is known about how the visual system processes 3D motion (or motion-in-depth, MID), however. A particular challenge arises when objects move toward or away from us (or when we move in depth relative to an object). In such cases, our brain must integrate both motion and depth information to infer the object's movement in the world.

Previous studies have identified two major visual cues thought to contribute to MID perception: inter-ocular velocity difference (IOVD), the differences in retinal motion signals in the two eyes as a result of their spatial offset, and changing disparity (CD), the change in binocular disparity over time as the object's depth changes (Cormack et al., 2017). By carefully controlling visual stimuli using random-dot stereograms, psychophysical studies have demonstrated the influence of each cue, though their relative contributions can vary depending on the stimulus design (Cormack et al., 2017).

At the neural level, a substantial number of neurons in the primate visual area MT are selectively tuned to MID, responding to both IOVD and CD cues, with some evidence suggesting a stronger weighting of the IOVD signal (Sanada and DeAngelis, 2014; Joo et al., 2016). A recent computational study by Bonnen et al. (2020) showed that tuning curves for MID stimuli in MT neurons can be predicted by a model derived from 3D geometric principles and that this model also accounts for some counterintuitive perceptual errors observed in human MID perception.

While research has examined the perception and neural correlates of MID, there remains a lack of models that directly compute MID from natural images in a biologically plausible manner. In particular, it is still unclear how the visual system infers MID in the face of variability, ambiguity, and noise in natural sensory inputs. Addressing this gap requires models that not only adapt to the statistical structure of the natural environment but also respect the biological constraints of early visual processing. Ideal-observer models provide a principled framework for this goal, as they characterize the statistically optimal solution to a perceptual inference problem given the information available in the sensory inputs and task demands, thereby offering a normative account of many perceptual phenomena (Geisler, 2011).

Herrera-Esposito and Burge (2025) recently developed a novel computational model that directly computes MID from a series of naturalistic binocular images. Their model incorporates biologically inspired constraints on retinal image

encoding, including an approximation of the optical blur caused by the human eye's optics, a temporal response function simulating that of the photoreceptors, and contrast normalization. After the encoding steps, this model uses a set of spatiotemporal filters to decode 3D speed or direction. Unlike traditional approaches that rely on hand-designed feature extraction, their model learns optimal filters from the structure of natural binocular videos, using local operations that extract information from only a small region of the image. It estimates either the 3D direction or speed of motion in a statistically optimal manner given the intrinsic variability of sensory inputs and constraints in image encoding (Herrera-Esposito and Burge, 2025).

Remarkably, two distinct subpopulations of filters emerged after training: in a 3D speed estimation task, one subpopulation of filters extracted the IOVD cue, and another extracted the CD cue, and in a 3D direction estimation task, two different subpopulations emerged, with one extracting motion information in the frontoparallel plane and the other extracting toward-away motion (Herrera-Esposito and Burge, 2025). These results are consistent with neurophysiological evidence that different subpopulations of MT neurons selectively respond to the IOVD and CD cues, respectively (Sanada and DeAngelis, 2014; Joo et al., 2016). Importantly, the model was not designed to explicitly extract the IOVD and CD cues; rather, these subpopulations of filters naturally emerged after training. Therefore, these findings explain how the IOVD and CD cues can arise from optimal processing of 3D motion in natural scenes.

Notably, this model also exhibited confusion between toward and away motion, similar to that reported in human psychophysics by Bonnen et al. (2020). In the model, this toward–away confusion is related to bimodal tuning curves with peak responses to both directions. This result suggests that human MID perception may reflect optimal computation adapted to natural scene statistics. Another interesting aspect of this model is that the weighted quadratic combination of its spatiotemporal filters closely resembles the classic energy model. The energy model is a widely used descriptive framework that characterizes how neurons respond to various visual stimuli, although it does not claim that such responses are optimal in any statistical sense. In contrast, the model developed by Herrera-Esposito and Burge (2025) was specifically optimized for MID computation from natural images. The similarity between these two models therefore suggests that the traditional energy model may in fact be optimal under natural scene statistics.

Despite its strengths, the model has some limitations. For example, the spatial extent of its local computations is smaller than the receptive-field size of most neurons in area MT, an area implicated in MID computation (Herrera-Esposito and Burge, 2025). Moreover, this model does not consider top-down feedback or long-range horizontal connections within the area. While this suggests that strictly local computations suffice for MID perception, future work that incorporates local and global processes would provide a more biologically plausible account of MID processing.

Several important questions about MID processing remain. First, to what extent are the spatiotemporal filters task specific? Herrera-Esposito and Burge (2025) developed separate ideal observers for estimating 3D speed and 3D direction, which revealed that different subpopulations of filters are emphasized in each task. However, natural behavior may require flexible coding of 3D motion vectors, including both speed and direction. Understanding how this general motion processing might arise within a single model remains an open challenge. Second, how invariant or robust are the learned filters across varying viewing contexts? Natural scenes contain rich 3D structures, and self-motion drastically changes the retinal image motion. Correctly attributing visual inputs to the 3D layout and motion in the scene requires the brain to flexibly perform different computations based on the inferred viewing geometry (Xu et al., 2024). Clarifying how generalizable these computations are across viewing conditions could lead to a better understanding of how the brain represents space and time.

In summary, the normative framework developed by Herrera-Esposito and Burge (2025) is a valuable step forward in our understanding of how the brain might optimally estimate 3D motion from natural visual input. Their ideal-observer model provides a principled account of how IOVD and CD cues could emerge from task-driven optimization. Their findings bridge the gap between ecological vision, Bayesian theory, human perception, and neural correlates. And it lays the foundation for a unified, normative account of 3D motion processing.

References

Anzai A, DeAngelis GC (2010) Neural computations underlying depth perception. Curr Opin Neurobiol 20:367–375.

Bonnen K, Czuba TB, Whritner JA, Kohn A, Huk AC, Cormack LK (2020) Binocular viewing geometry shapes the neural representation of the dynamic three-dimensional environment. Nat Neurosci 23:113–121.

Cormack LK, Czuba TB, Knoll J, Huk AC (2017) Binocular mechanisms of 3D motion processing. Annu Rev Vis Sci 3:297–318.

Geisler WS (2011) Contributions of ideal observer theory to vision research. Vision Res 51:771–781.

Herrera-Esposito D, Burge J (2025) Optimal estimation of local motion-in-depth with naturalistic stimuli. J Neurosci 45:e0490242024.

Howard IP, Rogers BJ (2008) *Seeing in depth: volume 1: basic mechanics/volume 2: depth perception 2-volume set*. United Kingdom: Oxford University Press.

Joo SJ, Czuba TB, Cormack LK, Huk AC (2016) Separate perceptual and neural processing of velocity-and disparity-based 3D motion signals. J Neurosci 36:10791–10802.

Park W, Tadin D (2018) Motion perception. In: *Stevens' handbook of experimental psychology and cognitive neuroscience* (Serences J, Wixted JT, eds), Vol. 2, pp 415–487. United Kingdom: Wiley.

Sanada TM, DeAngelis GC (2014) Neural representation of motion-in-depth in area MT. J Neurosci 34:15508–15521.

Xu Z-X, Pang J, Anzai A, DeAngelis GC (2024) Viewing geometry drives flexible perception of object motion and depth. bioRxiv, 2024.2010.2029.620928.