Version: 3/10/23 (PsyArXiv)

Compositionality in Visual Perception

Alon Hafri[1], E.J. Green[2], and Chaz Firestone[3]

[1]*Department of Linguistics and Cognitive Science, University of Delaware* | alon@udel.edu
[2]*Department of Linguistics and Philosophy, Massachusetts Institute of Technology* | ejgr@mit.edu
[3]*Department of Psychological and Brain Sciences, Johns Hopkins University* | chaz@jhu.edu

Abstract

Quilty-Dunn et al.'s wide-ranging defense of LoT argues that vision traffics in abstract, structured representational formats. We agree: Vision, like language, is *compositional*—just as words compose into phrases, many visual representations contain discrete constituents that combine in systematic ways. Here, we amass evidence extending this proposal, and explore its implications for how vision interfaces with the rest of the mind.

The world we see is populated by colors, textures, edges, and countless other visual features. Yet we see more than a collection of features: we also see whole objects, and relations within and between those objects. How are these entities represented? Here, we advance the case for LoT-like representation in perception. We argue that at least two types of visual representations are compositional, and we explore their connections with the rest of the mind.

Consider the hands in Figure 1A. Although they differ in various superficial features, they appear to share something: their *structure*—specifically, their *skeletal structure*. The same parts are connected in the same ways, just in different poses. Similarly, the middle shape in Figure 1B shares its structure with the left shape but not the right shape, even though the middle and right shapes share other features. Skeletal representations describe shapes via their parts' intrinsic axes and connections, often in a hierarchical *tree* format, wherein certain parts 'descend' or 'offshoot' from others (Feldman & Singh, 2006). Copious evidence suggests that skeletal representations are psychologically real, implicated in detection (Kovács & Julesz, 1994; Wilder et al., 2016), discrimination (Lowet et al., 2018), categorization (Wilder et al., 2011), aesthetics (van Tonder et al., 2002), and more (Psotka, 1978; Firestone & Scholl, 2014).

We contend that skeletal representations exhibit several of Quilty-Dunn et al.'s LoT properties: *discrete constituents*, *role-filler independence*, and *abstract content*. First, skeletal representations contain discrete constituents that represent axis structure independently of surrounding boundaries, composing with boundary representations to describe overall shape. This may explain why infants (Ayzenberg & Lourenco, 2022)

and adults (Wilder et al., 2011) categorize novel shapes by skeletal structure despite differences in surface properties. Second, representations of individual parts exhibit role-filler independence, retaining identity over changes in position within the overall skeletal representation. Such *transportability* (Fodor, 1987) explains why we can easily determine when distinct shapes share the same parts, and why such shapes prime one another (Cacciamani et al., 2014). Third, skeletal representations are abstract, expressing aspects of shape that appear stable despite part articulations (Figure 1A), changes in surface properties (Figure 1B; Green, 2019), and sense modality (Green, 2022). Moreover, visual brain areas encode skeletal structure across surface changes (Hung et al., 2012; Lescroart & Biederman, 2013; Ayzenberg et al., 2022). Skeletal representations may also encode non-metric, categorical properties—e.g., *straight*/*curved*, *symmetric*/*asymmetric* (Amir et al., 2012; Green, 2017; Hafri et al., 2022).

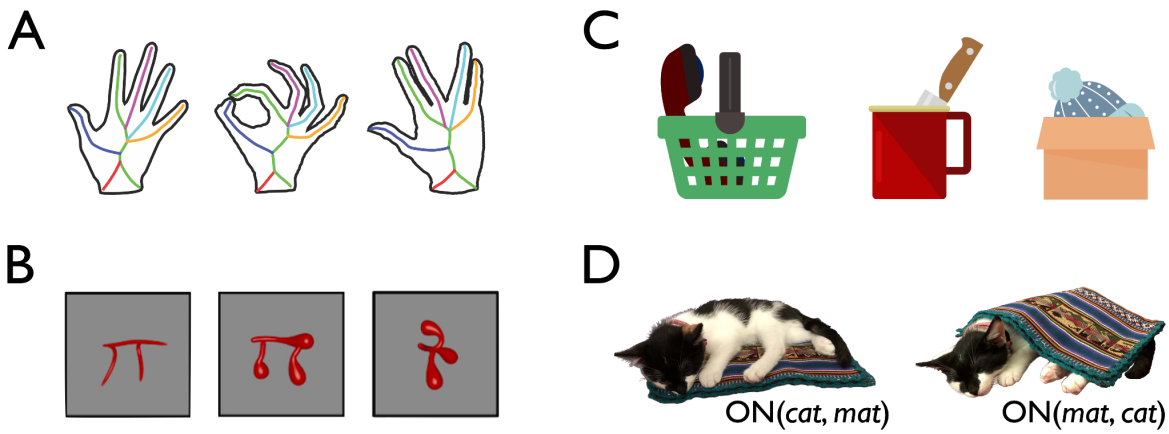

Figure 1. Demonstrations of compositionality in visual perception. (A) The three hands shown here differ in global shape, the locations of their boundaries, and other surface features; however, they appear to share something: their *structure*—specifically, their *skeletal structure* (indicated by the inset colored lines). The same parts have taken on different poses. Skeletal shape representations describe objects in terms of the axes of their parts, including how those parts are arranged with respect to one another, in ways that instantiate several core LoT properties. (Adapted from Lowet et al., 2018.) **(B)** Skeletal shape representations explain why infants and adults can see that the middle shape shares something with the leftmost shape that it does not share with the rightmost shape, even though the middle and rightmost shape share other features. (Adapted from Ayzenberg & Lourenco, 2019.) **(C)** The three object-pairs shown here differ in a variety of visual features, and even involve different objects—but each seems to instantiate the same relation: *containment*. Recent evidence suggests that the mind rapidly and automatically encodes such relations, representing the relation itself separately from the objects participating in it. (Adapted from Hafri et al., 2020.) **(D)** These two images depict the same objects (cat and mat) and the same relation (*support*), but differ in their *structure*—a cat on a mat is a very different scene from a mat on a cat. Put differently, 'argument order' matters: *R(x,y)* may be quite different than *R(y,x)*, and there is evidence that visual processing is sensitive to this difference in compositional structure. (Adapted from Hafri & Firestone, 2021.)

We suggest that these LoT properties make skeletal representations compositional: Discrete constituents encoding different geometrical elements and properties combine to form representations of global shape.

Compositionality in vision extends to relations *between* objects. Consider the object-pairs in Figure 1C. They appear to share something: the relation *containment*. Visual processing respects this commonality—it represents relations between objects, beyond the objects themselves (Hafri & Firestone, 2021). Such representations also exhibit several LoT properties. First, visual processing represents relations abstractly and categorically: Observers are more sensitive to metric changes across relational category boundaries (e.g., from containing to merely touching) than within (e.g., from one instance of containment to another; Lovett & Franconeri, 2017), and even 'confuse' instances of the same relation for one another (Hafri et al., 2020). Furthermore, visual brain areas encode eventive relations abstractly, generalizing across event participants (Wurm & Lingnau, 2015; Hafri et al., 2017).

Second, such representations contain discrete constituents and exhibit role-filler independence, in ways that augment Quilty-Dunn et al.'s discussion. Consider Figure 1D. Both images involve the same objects (cat and mat) and relation (*support*), but cat-on-mat differs from mat-on-cat in compositional structure. Thus, 'argument order' matters—the 'fillers' map to different roles. Recent work shows that vision is sensitive to this difference. When observers repeatedly reported the location of a target individual (e.g., blue-shirted man) in a stream of action photographs (e.g., blue-kicking-red, red-pushing-blue), a 'switching cost' emerged: slower responses when the target individual's role (*Agent*/*Patient*) switched (e.g., pusher on trial n–1 but kickee on trial n), suggesting that observers encoded relational structure automatically (Hafri et al., 2018).

These properties make representations of categorical between-object relations compositional: Discrete constituents encoding entities and relations combine to form representations of structured situations.

The prospect of LoT-like, compositional visual representations impacts broader debates about perception's *format*. Many claim that perceptual representations are constitutively iconic, analog, or 'picture-like' (Dretske 1981; Kosslyn et al., 2006; Carey, 2009; Burge, 2022). However, while LoT-like formats clearly suffice to encode categorical, non-degreed relations (e.g., containment), many iconic formats may not—particularly accounts requiring perceptual icons to mirror graded degrees of difference in perceptible properties (e.g., orientation or brightness; Block, 2023).

This perspective also raises exciting questions and research directions. For example, it may partially explain how information from perception is 'readily consumed' by cognitive and linguistic systems (due to the similar formats of some perceptual and higher-level representations; Quilty-Dunn, 2020; Cavanagh, 2021). Recent work explores these connections explicitly: skeletal shape representations impact aesthetic preferences and linguistic descriptions of shapes (Sun & Firestone, 2022a, 2022b), and representations of symmetry and roles may be shared across perception and language (Strickland, 2017; Hafri et al., 2018; Rissman & Majid, 2019; Hafri et al., 2022). One could also investigate the 'psychophysics' of compositional processes—the timing/ordering of how relational representations are built from their parts.

Nevertheless, LoT-like perceptual representations may not be *fully* language-like. While perception plausibly predicates properties of individuals (Quilty-Dunn & Green, 2021), it may lack the full *expressive freedom* of first-order logic (Camp, 2018), especially logical connectives needed for truth-functional completeness (Mandelbaum et al., 2022). Perception may be able to represent that an object is red but not that it is *not* red. Moreover, certain perceptual formats may impose constraints on which properties are attributable to which individuals—constraints absent from higher-level cognition. Perhaps perception cannot explicitly represent relations between non-adjacent object parts, or eventive relations of long durations (e.g., a jack slowly lifting a car).

Because perception and thought confront multifarious tasks with different computational demands, we contend that they comprise a multiplicity of formats (Marr, 1982; Yousif, 2022), each optimized for different computations, and some more LoT-like than others. Thus, any theory positing a single privileged format for perception or thought should be met with suspicion. Instead, researchers should heed Quilty-Dunn et al.'s advice to "let a thousand representational formats bloom."

References

Amir, O., Biederman, I., & Hayworth, K. J. (2012). Sensitivity to nonaccidental properties across various shape dimensions. *Vision Research, 62*, 35–43.

Ayzenberg, V., Kamps, F. S., Dilks, D. D., & Lourenco, S. F. (2022). Skeletal representations of shape in the human visual cortex. *Neuropsychologia, 164*, 108092.

Ayzenberg, V., & Lourenco, S. F. (2019). Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports*, *9*, 1–13.

Ayzenberg, V., & Lourenco, S. (2022). Perception of an object's global shape is best described by a model of skeletal structure in human infants. *eLife, 11*, e74943.

Block, N. (2023). *The Border Between Seeing and Thinking*. Oxford: Oxford University Press.

Burge, T. (2022). *Perception: First Form of Mind.* Oxford: Oxford University Press.

Cacciamani, L., Ayars, A. A., & Peterson, M. A. (2014). Spatially rearranged object parts can facilitate perception of intact whole objects. *Frontiers in Psychology, 5*, 482.

Camp, E. (2018). Why maps are not propositional. In A. Grzankowski & M. Montague (eds.), *Non-Propositional Intentionality*, 19-45. Oxford: Oxford University Press.

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

Cavanagh, P. (2021). The language of vision. *Perception*, *50*, 195–215.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences, 103*, 18014–18019.

Firestone, C., & Scholl, B. J. (2014). "Please tap the shape, anywhere you like": Shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science, 25*, 377–386.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* MIT Press.

Green, E. J. (2017). A layered view of shape perception. *The British Journal for the Philosophy of Science, 68*, 355–387.

Green, E. J. (2019). On the perception of structure. *Noûs, 53*, 564–592.

Green, E. J. (2022). The puzzle of cross-modal shape experience. *Noûs*, *56*, 867–896.

Hafri, A., Bonner, M. F., Landau, B., & Firestone, C. (2020). A phone in a basket looks like a knife in a cup: Role-filler independence in visual processing. *PsyArXiv*. https://psyarxiv.com/jx4yg

Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences, 25*, 475–492.

Hafri, A., Gleitman, L. R., Landau, B., & Trueswell, J. C. (2022). Where word and world meet: Language and vision share an abstract representation of symmetry. *Journal of Experimental Psychology: General.*

Hafri, A., Trueswell, J. C., & Epstein, R. A. (2017). Neural representations of observed actions generalize across static and dynamic visual Input. *The Journal of Neuroscience, 37*, 3056–3071.

Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition, 175*, 36–52.

Hung, C. C., Carlson, E. T., & Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron, 74*, 1099–1113.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The Case for Mental Imagery*. Oxford: Oxford University Press.

Kovács, I., & Julesz, B. (1994). Perceptual sensitivity maps within globally defined visual shapes. *Nature, 370*, 644–646.

Lescroart, M. D., & Biederman, I. (2013). Cortical representation of medial axis structure. *Cerebral Cortex, 23*, 629–637.

Lovett, A., & Franconeri, S. L. (2017). Topological relations between objects are categorically coded. *Psychological Science, 28*, 1408–1418.

Lowet, A. S., Firestone, C., & Scholl, B. J. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attention, Perception, & Psychophysics, 80*, 1278–1289.

Mandelbaum, E., Dunham, Y., Feiman, R., Firestone, C., Green, E. J., Harris, D., Kibbe, M. M., Kurdi, B., Mylopoulos, M., Shepherd, J., Wellwood, A., Porot, N., & Quilty-Dunn, J. (2022). Problems and mysteries of the many languages of thought. *Cognitive Science*, *46*, e13225.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman.

Psotka, J. (1978). Perceptual processes that may create stick figures and balance. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 101–111.

Quilty-Dunn, J. (2020). Concepts and predication from perception to cognition. *Philosophical Issues*, *30*, 273–292.

Quilty-Dunn, J., & Green, E. J. (2021). Perceptual attribution and perceptual reference. *Philosophy and Phenomenological Research*. DOI: 10.1111/phpr.12847.

Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin & Review, 26*, 1850–1869.

Strickland, B. (2017). Language reflects “core” cognition: A new theory about the origin of cross-linguistic regularities. *Cognitive Science*, *41*, 70–101.

Sun, Z., & Firestone, C. (2022a). Beautiful on the inside: Aesthetic preferences and the skeletal complexity of shapes. *Perception, 51*, 904–918.

Sun, Z., & Firestone, C. (2022b). Seeing and speaking: How verbal 'description length' encodes visual complexity. *Journal of Experimental Psychology: General, 151*, 82–96.

Van Tonder, G. J., Lyons, M. J., & Ejima, Y. (2002). Visual structure of a Japanese Zen garden. *Nature, 419*, 359–360.

Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition, 119*, 325–340.

Wilder, J., Feldman, J., & Singh, M. (2016). The role of shape complexity in the detection of closed contours. *Vision Research, 126*, 220–231.

Wurm, M. F., & Lingnau, A. (2015). Decoding Actions at Different Levels of Abstraction. *Journal of Neuroscience, 35*, 7727–7735.

Yousif, S. R. (2022). Redundancy and reducibility in the formats of spatial representations. *Perspectives on Psychological Science*, *17*, 1778–1793.