Bayesian Vector Analysis and the Perception of Hierarchical Motion

Samuel J. Gershman[1] **(sjgershm@mit.edu), Frank Jäkel**[2] **(fjaekel@uos.de), Joshua B. Tenenbaum**[1] **(jbt@mit.edu)**
[1]Department of Brain and Cognitive Sciences, MIT
[2]Institute of Cognitive Science, University of Osnabrück

Abstract

Scenes filled with moving objects are often hierarchically organized: the motion of a migrating goose is nested within the flight pattern of its flock, the motion of a car is nested within the traffic pattern of other cars on the road, the motion of body parts are nested in the motion of the body. Humans perceive hierarchical structure even in stimuli with two or three moving dots. An influential theory of hierarchical motion perception holds that the visual system performs a "vector analysis" of moving objects, decomposing them into common and relative motions. However, this theory does not specify how to resolve ambiguity when a scene admits more than one vector analysis. We describe a Bayesian theory of vector analysis and show that it can account for classic results from dot motion experiments. Our theory takes a step towards understanding how moving scenes are parsed into objects.

Keywords: motion perception; Bayesian inference; structure learning

Introduction

Motion is a powerful cue for understanding the organization of a visual scene. Infants use motion to individuate objects, even when it contradicts property/kind information (Kellman & Spelke, 1983; Xu et al., 1999). The primacy of motion information is also evident in adult object perception (Mitroff & Alvarez, 2007). In addition to individuating and tracking objects, motion is used by the visual system to decompose objects into parts. In biological motion, for example, the motion of body parts are nested in the motion of the body. Object motion may be hierarchically organized into multiple layers: an arm's motion may be further decomposed into jointed segments, including the hand, which can itself be decomposed into fingers, and so on.

The hierarchical organization of motion presents a formidable challenge to current models of motion processing. It is widely accepted that the visual system balances motion integration over space and time (necessary for solving the aperture problem) and motion segmentation in order to perceive multiple objects simultaneously (Braddick, 1993). However, it is unclear how simple segmentation mechanisms can be used to build a hierarchically structured representation of a moving scene. Segmentation lacks a notion of *nesting*: when an object moves, its parts should move with it. To understand nesting, it is crucial to represent the underlying dependencies between objects and their parts.

The experimental and theoretical foundations of hierarchical motion perception were laid by the pioneering work of Johansson (1950), who demonstrated that surprisingly complex percepts could arise from simple dot motions. Johansson proposed that the visual system performs a "vector analysis" of moving scenes into common and relative motions between objects. In the example of biological motion (see Johansson, 1973), the global motion of the body is subtracted from the image, revealing the relative motions of body parts; these parts are further decomposed by the same subtraction operation.

While the vector analysis theory provides a compelling explanation of numerous motion phenomena (we describe several below), it is incomplete from a computational point of view, since it relies on the theorist to provide the underlying motion components and their organization; it lacks a mechanism for *discovering* a hierarchical decomposition from sensory data. This is especially important in complex scenes where many different vector analyses are consistent with the scene. Various principles have been proposed for how the visual system resolves this ambiguity. For example, Restle (1979) proposed a "minimum principle," according to which simpler motion interpretations (i.e., those with a shorter description length) are preferred over more complex ones. Gogel (1974) argued for an "adjacency principle," according to which the motion interpretation is determined by relative motion cues between nearby points. However, there is still no unified computational theory that can encompass all these ideas.

In this paper, we recast Johansson's vector analysis theory in terms of a Bayesian model of motion perception. The model discovers the hierarchical structure of a moving scene, resolving the ambiguity of multiple vector analyses using a set of probabilistic constraints. We show that this model can account for several classic phenomena in the motion perception literature that are challenging for existing models.

Bayesian vector analysis

In this section, we describe our computational model formally. We start by describing a probabilistic generative model of motion—a set of assumptions about the environment that we impute to the observer. The generative model can be thought of as stochastic "recipe" for generating moving images. We then describe how Bayesian inference can be used to invert this generative model and recover the underlying hierarchical structure from observations of moving images.

Generative model

Our model describes the process by which a sequence of two-dimensional visual element positions $\{\mathbf{s}_n(t)\}_{n=1}^N$ is generated, where $\mathbf{s}_n(t) = [s_n^x(t), s_n^y(t)]$ is the x and y position of element n at time step t.[1] Elements can refer to objects, parts or features;

[1] This representation assumes that basic perceptual preprocessing has taken place (e.g., the correspondence problem has been solved).

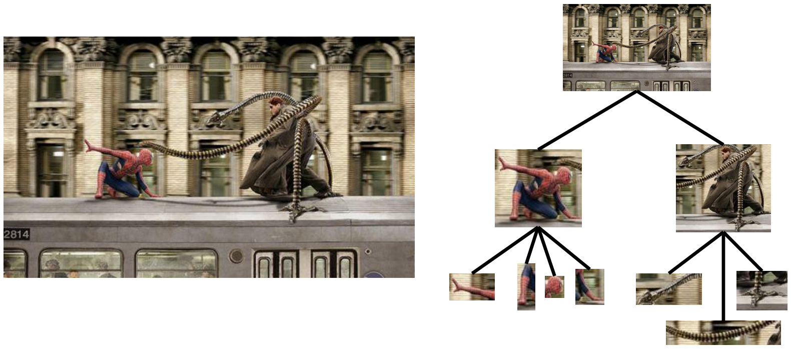

Figure 1: **Illustration of how a moving scene is decomposed into a motion tree**. Each node in the tree corresponds to a motion component. Each object in the scene traces a path through the tree, and the observed motion of the object is modeled as the superposition of motion components along its path.

in this paper we will simply refer to them as objects. The object positions are modeled as arising from a tree-structured configuration of motion components; we refer to this representation as the *motion tree*. Each motion component is a transformation that maps the current object position to a new position.

An illustration of a motion tree is shown in Figure 1. Each node in the tree corresponds to a motion component. The motion of the train relative to the background is represented by the top-level node. The motions of Spiderman and Dr. Octopus relative to the train are represented at the second-level nodes. Finally, the motions of each body part relative to the body are represented at the third-level nodes. The observed motion of Spiderman's hand can then be modeled as the superposition of the motions along the path that runs from the top node to the hand-specific node. The aim for our model is to get as inputs the retinal motion of pre-segmented objects—in this example, the motion of hands, feet, torsos, windows, etc.—and output a hierarchical grouping that reflects the composition of the moving scene.

The motion tree can capture the underlying motion structure of many real-world scenes, but inferring which motion tree generated a particular scene is challenging because different trees may be consistent with the same scene. To address this problem, we need to introduce a prior distribution over motion trees that expresses our inductive biases about what kinds of trees are likely to occur in the world. This prior should be flexible enough to accommodate many different structures while also preferring simpler structures (i.e., parsimonious explanations of the sensory data). These desiderata are satisfied by a nonparametric distribution over trees known as the *nested Chinese restaurant process* (nCRP; Blei et al., 2010). The nCRP generates a motion tree by drawing, for each object n, a sequence of motion components, denoted by $\mathbf{c}_n = [c_{n1}, \dots, c_{nD}]$, where D is the maximal tree depth.[2] The component assignments are drawn according to:

$$P(c_{nd} = j | \mathbf{c}_{1:n-1}) = \begin{cases} \frac{M_j}{n-1+\gamma} & \text{if } j \leq J \\ \frac{\gamma}{n-1+\gamma} & \text{if } j = J+1 \end{cases} \tag{1}$$

where j indexes motion components, M_j is the number of previous objects assigned to component j, and J is the number of components currently in use (i.e., those for which $M_j > 0$). The assignment at depth d is restricted to a unique set of components specific to the component assigned at depth $d-1$. In this way, the components form a tree structure, and $\mathbf{c}_n$ is a path through the tree. The parameter $\gamma \geq 0$ controls the branching factor of the motion tree. As γ decreases, different objects will tend to share the same motion components. Thus, the nCRP exhibits a preference for trees that use a small number of motion components.

Note that so far we have generated a path through a potentially very deep tree for each object. Each path has the same length D. Remember that each node in the tree will represent a motion component. We want each object n to be associated with a node in the tree and its overall motion to be the sum of all the motion components above it (including itself). Hence, for each object we need to sample an additional parameter $d_n \in \{1, \dots, D\}$ that determines to which level on the tree the object will be assigned. This depth specifies a truncation of $\mathbf{c}_n$, thereby determining which components along the path contribute to the observations. The depth assignments $\mathbf{d} = [d_1, \dots, d_N]$ are drawn from a Markov random field:

$$P(\mathbf{d}) \propto \exp\left\{ \alpha \sum_{m=1}^{N} \sum_{n>m}^{N} \mathbb{I}[d_m = d_n] - \rho \sum_{n=1}^{N} d_n \right\}, \tag{2}$$

where the indicator function $\mathbb{I}[\cdot] = 1$ if its argument is true and 0 otherwise. The parameter α controls the penalty for assigning objects to different depths, and the parameter ρ controls a penalty for deeper level assignments.

Each motion component, i.e. each node in the motion tree, is associated with a time-varying flow field, $\mathbf{f}_j(\mathbf{s}, t) = [f_j^x(\mathbf{s}, t), f_j^y(\mathbf{s}, t)]$. We place a prior on flow fields that enforces spatial smoothness but otherwise makes no assumptions about functional form. In particular we assume that f_j^x and f_j^y are spatial functions drawn independently at each time discrete time step t from a zero-mean Gaussian process with covariance function

$$k(\mathbf{s}, \mathbf{s}') = \tau \exp\left\{ -\frac{||\mathbf{s} - \mathbf{s}'||^2}{2\lambda} \right\}, \tag{3}$$

[2] As described in Blei et al. (2010), trees drawn from the nCRP can be infinitely deep, but we impose a maximal depth for simplicity.

where τ is a global scaling parameter and $\lambda > 0$ is a length-scale parameter controlling the smoothness of the flow field. When λ is large, the flow field becomes rigid. Smoothness is only enforced between objects covered by the same node in the motion tree.

To complete the generative model, we need to specify how the motion tree gives rise to observations, which in our case are the positions of the N objects over time. For each object, the dot position at the next time step is set by sampling a displacement from a Gaussian whose mean is the sum of the flow fields along path $\mathbf{c}_n$ truncated at d_n:

$$\mathbf{s}_n(t+1) = \mathbf{s}_n(t) + \sum_{d=1}^{d_n} \mathbf{f}_{c_{nd}}(\mathbf{s}_n(t), t) + \varepsilon_n(t), \qquad (4)$$

where $\varepsilon_n(t) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

This generative model contains a number of important special cases under particular parameter settings. When $\gamma = 0$, only one motion component will be generated; in this case, the prior on flow-fields—favoring local velocities close to 0 that vary smoothly over the image—resembles the "slow and smooth" model proposed by Weiss & Adelson (1998). When $\gamma = 0$ and $\lambda \to \infty$, we obtain the "slow and rigid" model of Weiss et al. (2002). When $D = 1$, the model will generate multiple motion components, but these will all exist at the same level of the hierarchy (i.e., the motion tree is flat, with no nesting), resulting in a form of transparent layered motion (Wang & Adelson, 1993; Weiss, 1997).

Inference

The goal of inference is to compute the posterior over the motion tree given a set of observations.[3] Because we are mainly interested in the highest probability tree, we use annealed Gibbs sampling to search for the posterior mode. The algorithm alternates between holding the depth assignments fixed while sampling the node assignments, and holding the node assignments fixed while sampling the depth assignments. By raising the conditional probabilities to a power $\beta > 1$, the posterior becomes peaked around the mode. We gradually increase β, so that the algorithm eventually settles on a high probability tree. We repeat this procedure 10 times (with 500 sampling iterations on each run) and pick the tree with the highest posterior probability. Below, we derive the conditional distributions used by the sampler.

The conditional distribution over $\mathbf{c}_n$ is given by:

$$P(\mathbf{c}_n|\mathbf{c}_{-n}, \mathbf{s}, \mathbf{d}) \propto P(\mathbf{c}_n|\mathbf{c}_{-n}) P(\mathbf{s}|\mathbf{c}, \mathbf{d}), \qquad (5)$$

where $\mathbf{c}_{-n}$ denotes the set of all paths excluding $\mathbf{c}_n$. The first factor in Eq. 5 is the nCRP prior (Eq. 1). The second factor in Eq. 5 is the likelihood of the data, given by:

$$P(\mathbf{s}|\mathbf{c}, \mathbf{d}) = \prod_t \prod_{z \in \{x,y\}} \mathcal{N}(\mathbf{s}^z(t+1); \mathbf{s}^z(t), \mathbf{K}(t) + \sigma^2\mathbf{I}) \qquad (6)$$

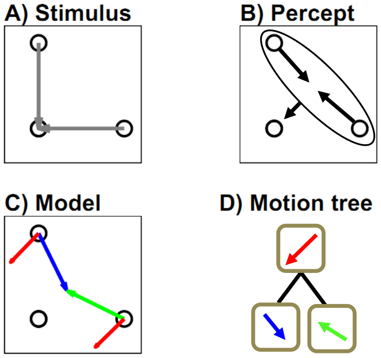

Figure 2: **Johansson (1950) two dot experiment**. (*A*) Veridical motion vectors. (*B*) Perceived motion. (*C*) Inferred motion vectors. Each color corresponds to a different component in the motion tree (*D*), but note that a component will predict different vectors depending on spatial location.

where

$$K_{mn}(t) = k(\mathbf{s}_m(t), \mathbf{s}_n(t)) \sum_j \mathbb{I}[j \in \mathbf{c}_m \wedge j \in \mathbf{c}_n]. \qquad (7)$$

Intuitively, the covariance between two points counts the number of nodes shared between their paths, weighted by their proximity in space.

The conditional distribution over d_n is given by:

$$P(d_n|\mathbf{c}, \mathbf{s}, \mathbf{d}_{-n}) \propto P(d_n|\mathbf{d}_{-n}) P(\mathbf{s}|\mathbf{c}, \mathbf{d}_{-n}, d_n), \qquad (8)$$

where $\mathbf{d}_{-n}$ denotes the level assignments excluding d_n and

$$P(d_n|\mathbf{d}_{-n}) \propto \exp\left\{\alpha \sum_{m \neq n} \mathbb{I}[d_m = d_n] - \rho d_n\right\}. \qquad (9)$$

To visualize the motion components that are given by a grouping through $\mathbf{d}_n$ and $\mathbf{c}_n$, we can calculate the posterior predictive mean for object n at each component j (shown here for the x dimension):

$$\mathbb{E}[f_j^x(\mathbf{s}_n(t), t)] = \mathbf{k}_{nj}^\top (\mathbf{K}(t) + \sigma^2\mathbf{I})^{-1} (\mathbf{s}^x(t+1) - \mathbf{s}^x(t)), \quad (10)$$

where $\mathbf{k}_{nj}$ is the N-dimensional vector of covariances between $\mathbf{s}_n(t)$ and the locations of all the objects whose paths pass through node j (if an object does not pass through node j then its corresponding entry in $\mathbf{k}_{nj}$ is 0).

Simulations

In this section, we show how the Bayesian vector analysis model can account for several classic experimental phenomena. These experiments all involve stimuli consisting of moving dots, so for present purposes $\mathbf{s}_n(t)$ corresponds to the position of dot n at time t. In these simulations we use the following parameters: $D = 3, \sigma^2 = 0.01, \tau = 1, \lambda = 100, \alpha = 1, \rho = 0.1$. The interpretation of σ^2 and λ depend on the spatial scale

[3] The latent motion components can be marginalized analytically using properties of Gaussian processes.

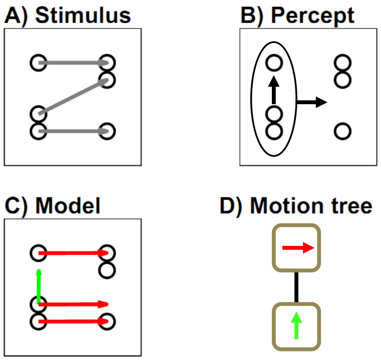

Figure 3: **Johansson (1973) three dot experiment**. (*A*) Veridical motion vectors. (*B*) Perceived motion. (*C*) Inferred motion vectors. (*D*) Inferred motion tree.

of the data; in general, we found that changing these parameters (within the appropriate order of magnitude) had little influence on the posterior. We set λ to be large enough so that objects assigned to the same layer moved near-rigidly.

Johansson (1950) demonstrated that a hierarchical motion percept can be achieved with as few as two dots. Figure 2A shows the stimulus used by Johansson, consisting of two dots translating orthogonally to meet at a single point. Observers, however, do not perceive the orthogonal translation. Instead, they perceive the two dots translating along a diagonal axis towards each other, which itself translates towards the meeting point (Figure 2B). Thus, observers perceive the stimulus as organized into common and relative motions. This percept is reproduced by the Bayesian vector analysis model (Figure 2C); the inferred motion tree (shown in Figure 2D) represents the common motion as the top level component and the relative motions as subordinate components. The subordinate components are not perfectly orthogonal to the diagonal motion, consistent with the findings of Wallach et al. (1985); this arises in our model through a form of "explaining away"—i.e., posterior coupling between the motion layers implied by Eq. 10.

Another example studied by Johansson (1973) is shown in Figure 3A. Here the bottom and top dot translate horizontally while the middle dot translates diagonally such that all three dots are always collinear. The middle dot is perceived as translating vertically as all three dots translate horizontally (Figure 3B). Consistent with this percept, the Bayesian vector analysis assigns all three dots to a common horizontal motion component, and additionally assigns the middle dot to a vertical motion component (Figure 3C-D).

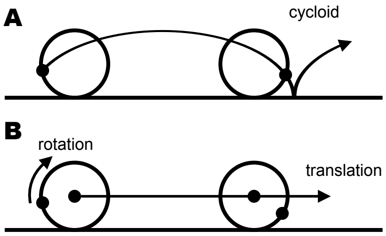

Figure 4: **Duncker wheel**. (*A*) A light on the rim of a rolling wheel produces cycloidal motion. (*B*) Adding a light on the hub produces rolling motion (translation + rotation).

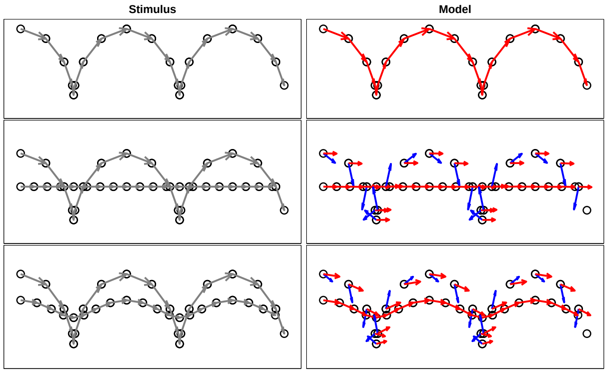

Figure 5: **Simulations of the Duncker wheel**. (*Top*) A single light on the rim produces one vector following a cycloidal path. (*Middle*) Adding a light on the hub produces two vectors: translation + rotation, giving rise to the percept of rolling motion. (*Bottom*) Placing the light on the interior of the wheel produces weaker rolling motion: the translational component is no longer perfectly horizontal.

Duncker (1929) showed that if a light is placed on the rim of a rolling wheel in a dark room, cycloidal motion is perceived (Figure 4A), but if another light is placed on the hub then rolling motion is perceived (Figure 4B). Simulations of these experiments are shown in Figure 5. When a light is placed only on the rim, there is strong evidence for a single cycloidal motion component, whereas stronger evidence for a two-level hierarchy (translation + rotation) is provided by the hub light.[4] It has also been observed that placing a light in between the rim and the hub produces weaker rolling motion (i.e., the translational component is no longer perfectly horizontal; Proffitt et al., 1979), a phenomenon that is reproduced by Bayesian vector analysis (Figure 5, bottom).

So far, we have been considering qualitative characterizations of various motion phenomena, but one advantage of a computational model is its ability to make quantitative predictions. We illustrate the quantitative power of Bayesian vector analysis for the case of motion transparency. When two groups of randomly moving dots are superimposed, observers may see either transparent motion (two planes of motion sliding past each other) or non-transparent motion (all dots moving in the direction of the average motion of the two groups). Which percept prevails depends on the relative direction of the two groups (Braddick et al., 2002): as the direction difference increases, transparent motion becomes more perceptible. We computed the probability of transparent motion (i.e.,

[4] Note that the model does not explicitly represent rotation but instead represents the tangential motion component in each time step.

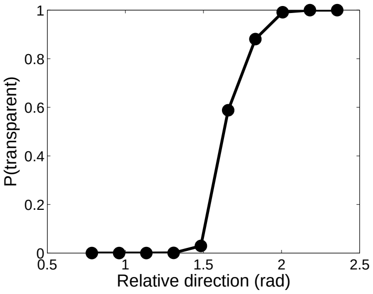

Figure 6: **Simulations of transparent motion**. Transparency increases as a function of direction difference between two superimposed groups of dots.

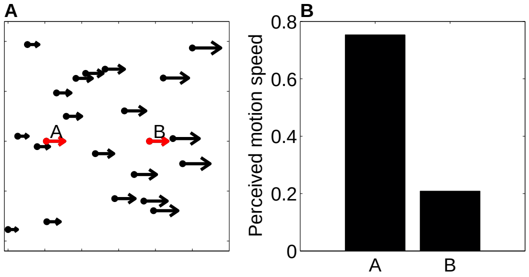

Figure 7: **Motion contrast**. (*A*) The velocity of the background (black) dots increases along the horizontal axis. Although A and B have the same velocity, A is perceived as moving faster than B. (*B*) Model simulation.

two layers in our model) for a range of relative directions using 20 dots. As the relative direction increases, the statistical evidence in favor of two separate layers increases, resulting in a smoothly changing probability (Figure 6).

Inferences about the motion hierarchy may interact with the spatial structure of the scene. The phenomenon of motion contrast, originally described by Loomis & Nakayama (1973), provides an illustration: The perceived motion of a dot depends on the motion of surrounding "background dots" (the black dots in Figure 7A). If a set of dots moves on a screen such that the dots on the left move more slowly than dots on the right, they form a velocity gradient. Two "target" dots that move with the same velocity and keep a constant distance (the red dots in Figure 7A) can still be perceived as moving with radically different speeds, depending on the speed of the dots close by. In our model, most of the motion of the velocity gradient is captured by the Gaussian process on the top-level motion component. However, this top-level component does not capture all of the motion of each dot. The target dots (in red), in particular, are each endowed with their own motion component and move relative to the top-level node. This relative motion differs depending on where along the gradient the target dot is located, resulting in motion contrast (Figure 7B).

How does our model scale up to more complex displays? An interesting test case is biological motion perception: Johansson (1973) showed that observers can recognize human motions like walking and running from lights attached to the joints. Later work has revealed that a rich variety of information can be discriminated by observers from point light displays, including gender, weight and even individual identity (Blake & Shiffrar, 2007). We trained our model (with the same parameters) on point light displays derived from the CMU human motion capture database.[5] These displays consisted of the 3-dimensional positions of 31 dots, including walking, jogging and sitting motions. The resulting motion parse is illustrated in Figure 8: the first layer of motion (not shown) captures the overall trajectory of the body, while the second and third layers capture more fine-grained structure, such as the division into limbs and smaller jointed body parts. Note that the model knows nothing about the underlying skeletal structure; it infers body parts directly from the dot positions. This demonstrates that Bayesian vector analysis can scale up to more complex and realistic motion patterns.

Conclusion

How does the visual system parse the hierarchical structure of moving scenes? In this paper, we have developed a Bayesian framework for modeling hierarchical motion perception, building upon the seminal work of Johansson (1950). The key idea of our theory is that a moving scene can be interpreted in terms of an abstract graph—the motion tree—encoding the dependencies between moving objects. Bayesian vector analysis is the process of inferring the motion tree from a sequence of images. Our simulations demonstrated that this formalism is capable of capturing a number of classic phenomena in the literature on hierarchical motion perception.

Two limitations of our theory need to be addressed. First, the generative model assumes that motion components combine through summation, but this is not adequate in general. For example, a better treatment of the Duncker wheel would entail modeling the *composition* of rotation and translation. In its current form, the model approximates rotation by inferring motion components that are tangent to the curve traced by the rotation. We are currently investigating a version of the generative model in which motion transformation compose with one another, which would allow for nonlinear interactions. Second, although we described an algorithm for finding the optimal motion tree, Bayesian vector analysis is really specified at the computational level; our simulations are not illuminating about the mechanisms by which the vector analysis is carried out. Nor does it commit to any particular neural implementation. More work is needed to connect all these levels of analysis. Grossberg et al. (2011) have described a detailed theory of how vector analysis could be performed by the visual cortex, and their efforts offer a possible starting point.

We view hierarchical motion as a model system for study-

[5] http://mocap.cs.cmu.edu/

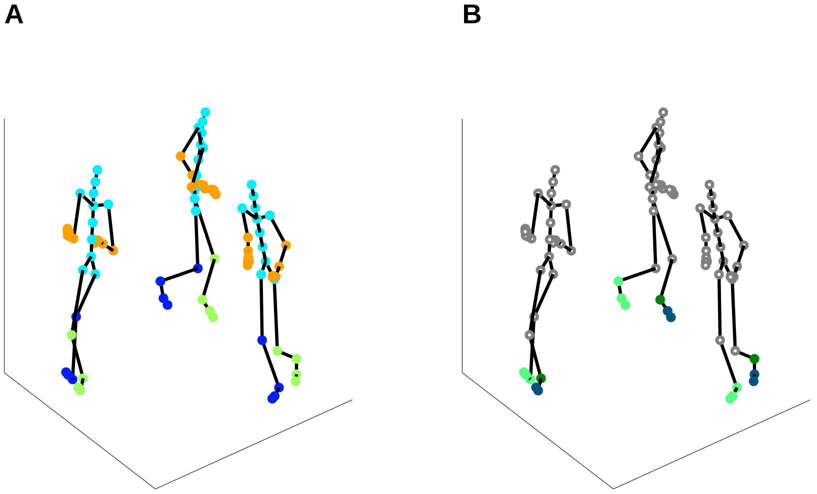

Figure 8: **Analysis of human motion capture data**. Each color represents the assignment of a node to a motion component. All nodes are trivially assigned to the first layer (not shown). In addition, all nodes were assigned to the second layer (*A*). A subset of the nodes were also assigned components in the third layer (*B*). Unfilled nodes indicate that no motion component was assigned at that layer. The skeleton is shown here for display purposes; the model was trained only on the dot positions.

ing more general questions about structured representations in mind and brain. The simplicity of the stimuli makes them amenable to rigorous psychophysical and neurophysiological experimentation, offering hope that future work can isolate the neural computations underlying structured representations like motion trees.

Acknowledgments

We thank Ed Vul and Peter Battaglia for helpful discussions. This work was supported by the Deutsche Forschungsgemeinschaft (DFG JA 1878/1-1), ONR MURI N00014-07-1-0937, IARPA ICARUS program, and the MIT Intelligence Initiative.

References

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology*, *58*, 47–73.

Blei, D., Griffiths, T., & Jordan, M. (2010). The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*, 7.

Braddick, O. (1993). Segmentation versus integration in visual motion processing. *Trends in Neurosciences*, *16*, 263–268.

Braddick, O., Wishart, K., & Curran, W. (2002). Directional performance in motion transparency. *Vision Research*, *42*(10), 1237–1248.

Duncker, K. (1929). Über induzierte Bewegung. (Ein Beitrag zur Theorie optisch wahrgenommener Bewegung). *Psychologische Forschung*, *12*, 180-259.

Gogel, W. (1974). Relative motion and the adjacency principle. *The Quarterly Journal of Experimental Psychology*, *26*, 425–437.

Grossberg, S., Léveillé, J., & Versace, M. (2011). How do object reference frames and motion vector decomposition emerge in laminar cortical circuits? *Attention, Perception, & Psychophysics*, *73*(4), 1147–1170.

Johansson, G. (1950). *Configurations in event perception*. Almqvist & Wiksell.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception, & Psychophysics*, *14*(2), 201–211.

Kellman, P., & Spelke, E. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology*, *15*(4), 483–524.

Loomis, J., & Nakayama, K. (1973). A velocity analogue of brightness contrast. *Perception*, *2*(4), 425–427.

Mitroff, S., & Alvarez, G. (2007). Space and time, not surface features, guide object persistence. *Psychonomic Bulletin & Review*, *14*, 1199–1204.

Proffitt, D., Cutting, J., & Stier, D. (1979). Perception of wheel-generated motions. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 289–302.

Restle, F. (1979). Coding theory of the perception of motion configurations. *Psychological Review*, *86*(1), 1–24.

Wallach, H., Becklen, R., & Nitzberg, D. (1985). Vector analysis and process combination in motion perception. *Journal of Experimental Psychology: Human Perception and Performance*, *11*, 93–102.

Wang, J., & Adelson, E. (1993). Layered representation for motion analysis. In *Computer vision and pattern recognition* (pp. 361–366).

Weiss, Y. (1997). Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *Computer vision and pattern recognition* (pp. 520–526).

Weiss, Y., & Adelson, E. (1998). Slow and smooth: A bayesian theory for the combination of local motion signals in human vision. *AI Memo 1616, MIT*.

Weiss, Y., Simoncelli, E., Adelson, E., et al. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604.

Xu, F., Carey, S., Welch, J., et al. (1999). Infants' ability to use object kind information for object individuation. *Cognition*, *70*, 137–166.