Discovering hierarchical motion structure

Samuel J. Gershman*

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Frank Jäkel

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, as well as new experimental data. Our theory takes a step towards understanding how moving scenes are parsed into objects.

Keywords: motion perception, Bayesian inference, structure learning

*Corresponding address: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Telephone: 773-607-9817

Email addresses: sjgershm@mit.edu (Samuel J. Gershman), jbt@mit.edu (Joshua B. Tenenbaum), fjaekel@uos.de (Frank Jäkel)

1. 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 and Spelke, 1983; Xu and Carey, 1996; Xu et al., 1999). The primacy of motion information is also evident in adult object perception (Burke, 1952; Flombaum and Scholl, 2006; Mitroff and Alvarez, 2007) and non-human primates (Flombaum et al., 2004). For example, in the *tunnel effect* (Burke, 1952; Flombaum et al., 2004; Flombaum and Scholl, 2006), an object passing behind an occluder is perceived as the same object when it reappears despite changes in surface features (e.g., color), as long as it reappears in the time and place stipulated by a spatiotemporally continuous trajectory.

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 (Johansson, 1973).

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 (see also Shum and Wolford, 1983). In the example of biological motion (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 (see also Attneave, 1954; Hochberg and McAlister, 1953). While such description length approaches are formally related to the Bayesian approach described below, Restle only developed his model to explain a small set of parametrized motions under noiseless conditions. Gogel (1974) argued for an "adjacency principle," according to which the motion interpretation is determined by relative motion cues between nearby points. The "belongingness principle" (DiVita and Rock, 1997) holds that relative motion is determined by the perceived coplanarity of objects and their potential reference frames. 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—*Bayesian vector analysis*. 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 qualitatively for several classic phenomena in the motion perception literature that are challenging for existing models. We then report a new experiment to demonstrate that the model can also provide a good quantitative fit to human data.

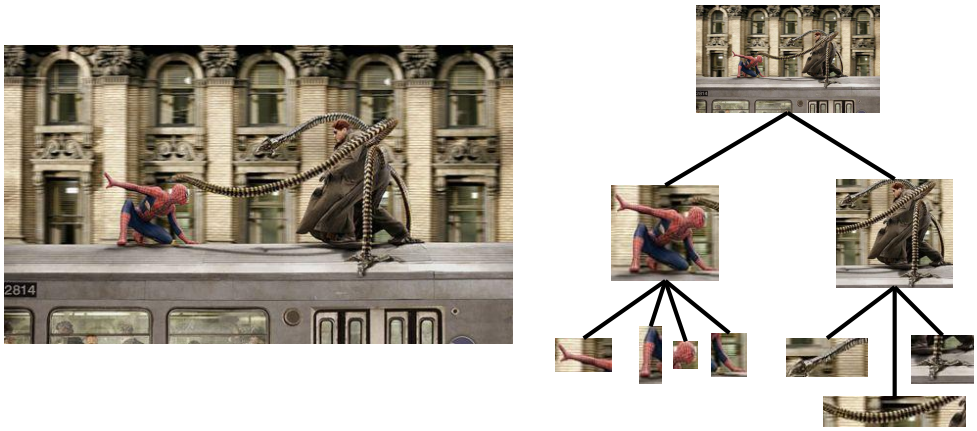


Figure 1: **Illustration of how a moving scene is decomposed into a motion tree**. Each node in the tree corresponds to a motion component. Each object in the scene traces a path through the tree, and the observed motion of the object is modeled as the superposition of motion components along its path.

2. Bayesian vector analysis

In this section, we describe our computational model formally.[1] 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, consisting of two parts: a probability distribution over trees, and a probability distribution over data (image sequences) given a particular tree. 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. Specifically, the goal of inference is to find the motion tree with highest posterior probability. According to Bayes' rule, the posterior $P(\text{tree}|\text{data})$ is proportional to the product of the likelihood $P(\text{data}|\text{tree})$ and the prior $P(\text{tree})$. The likelihood encodes the fit between the data and a hypothetical tree, while the prior encodes the "goodness" (in Gestalt terms) of the tree.

[1] Matlab code implementing the model is available at `https://github.com/sjgershm/hierarchical_motion`.

2.1. Generative model

The generative model describes the process by which a sequence of two-dimensional visual element positions $\{\mathbf{s}_n(t)\}_{n=1}^N$ is generated, where $\mathbf{s}_n(t) = [s_n^x(t), s_n^y(t)]$ encodes the x and y position of element n at time step t. Most experimental demonstrations of vector analysis have used moving dot displays. A good example are point-light walkers. For these demonstrations each moving dot is naturally represented by its 2-d position on the screen at each time point. This representation, of course, assumes that basic perceptual preprocessing has taken place and the correspondence problem has been solved. Although we will only model moving dot displays in this paper, and hence $\mathbf{s}_n(t)$ is usually the position of the n^{th} dot at time t, $\mathbf{s}_n(t)$ could also be the position of an object, a visual part, or a feature. In the following, we will simply refer to the elements whose movement we want to analyze as either dots or 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 is a generalization of the *Chinese restaurant process* (Aldous, 1985; Pitman, 2002), a distribution over partitions of objects. A tree can be understood as a nested partition of objects, where each layer of the tree defines a partition of objects, and thus a distribution over trees can be constructed by recursively sampling from a distribution over partitions. This is the logic underlying the nCRP construction.

The nCRP generates a motion tree by drawing, for each object n, a sequence of motion components, denoted by $\mathbf{c}_n = [c_{n1}, \ldots, c_{nD}]$, where D is the maximal tree depth.[2] The probability of assigning object n to component j at depth d is proportional to the number of previous objects assigned to component j (M_j). This induces a simplicity bias, whereby most objects tend to be assigned to a small number of motion components. With probability proportional to γ, an object can always be assigned to a new (previously unused) motion component. Thus, the model has "infinite capacity" in the sense that it can generate arbitrarily complex motion structures, but will probabilistically favor simpler structures. Mathematically, we can write the component assignment process as:

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

where J is the number of components currently in use (i.e., those for which $M_j > 0$). Importantly, 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

[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.

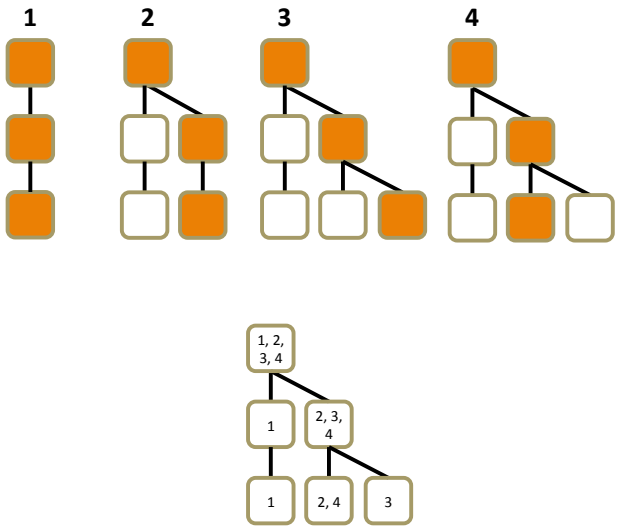


Figure 2: **Illustration of the tree-generating process**. (*Top*) Objects are successively added, going from left to right. Orange shading indicates the component assignments for each object. (*Bottom*) Alternative visualization showing which objects are assigned to each component.

motion components.

Figure 2 (top panel) shows how a tree is generated by successively adding objects. Starting from the left, a single object follows a path (indicated by orange shading) through 3 layers of the motion tree. Note that the initial object always follows a chain since no other branches have yet been created. The second object creates a new branch at layer 2. The third object creates a new branch at layer 3. The fourth object follows the same trajectory as the second object. Once all objects have been assigned paths through the tree, each layer of the tree defines a partition of objects, as shown in the bottom panel of 2.

Thus 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, not necessarily at depth D, 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 $d_n \in \{1, \ldots, D\}$ 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 $\mathbf{d} = [d_1, \ldots, d_N]$ 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 $\alpha > 0$ controls the penalty for assigning objects to different depths, and the parameter $\rho > 0$ controls a penalty for deeper level assignments. With this Markov random field objects tend to be placed high up in the tree and on the same level as other objects.

Each motion component (i.e. each node j 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 discrete time step t from a Gaussian process (see Rasmussen and Williams, 2006, for a comprehensive introduction):

$$P(f) = \mathrm{GP}(f; m, k), \tag{3}$$

where $m(\mathbf{s})$ is the *mean function* and $k(\mathbf{s}, \mathbf{s}')$ is the *covariance function*. The mean function specifies the average flow field, while the covariance function specifies the dependency between motion at different spatial locations: the stronger the covariance between spatial locations, the smoother flow fields become. We assumed $m(\mathbf{s}) = 0$ for all $\mathbf{s}$, and a squared exponential covariance function:

$$k(\mathbf{s}, \mathbf{s}') = \tau \exp\left\{-\frac{||\mathbf{s} - \mathbf{s}'||^2}{2\lambda}\right\}, \tag{4}$$

where $\tau > 0$ 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 sharing the same motion component. Examples of flow fields sampled from the prior are shown in Figure 3.

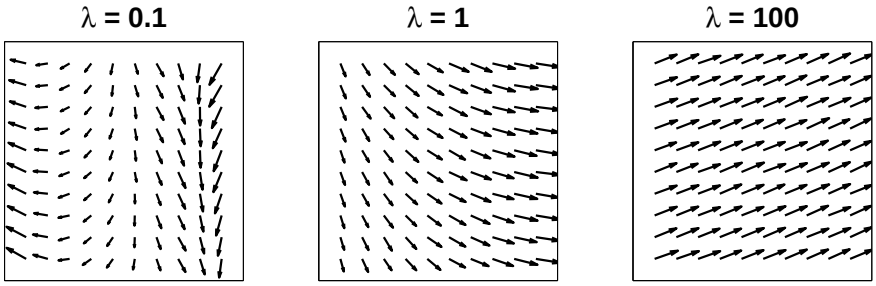


Figure 3: **Flow fields sampled from a Gaussian process**. Each panel shows a random flow field sampled with a different length-scale parameter (λ). As the length-scale parameter gets larger, the flow fields become increasingly rigid.

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. The position of object n 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. The function $\texttt{node}(n, d)$ picks out the index of the node of the tree that lies at depth d on path $\mathbf{c}_n$ and therefore

$$\mathbf{s}_n(t+1) = \mathbf{s}_n(t) + \sum_{d=1}^{d_n} \mathbf{f}_{\texttt{node}(n,d)}(\mathbf{s}_n(t), t) + \boldsymbol{\epsilon}_n(t), \tag{5}$$

where $\boldsymbol{\epsilon}_n(t) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ represents sensory noise with variance σ^2 and $\mathbf{I}$ is the identity matrix. This is equivalent to sampling displacements for each motion component separately and then adding up the displacements to form the next object position.

The additive form in Eq. 5 allows us to analytically marginalize the latent functions to compute the conditional distribution over image sequences given the component and depth assignments (see Rasmussen and Williams, 2006):

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

where $\mathbf{K}(t)$ is the Gram matrix of covariances between objects:

$$K_{mn}(t) = k(\mathbf{s}_m(t), \mathbf{s}_n(t))\phi_{mn}. \tag{7}$$

where the function ϕ_{mn} is the number of components shared by m and n (implicitly a function of $\mathbf{c}_m$ and $\mathbf{c}_n$). Intuitively, the covariance between two points counts the number of motion components shared between their paths, weighted by their proximity in space. Thus, the model captures two important Gestalt principles: grouping by proximity and common fate (Wertheimer, 1923).

The generative model described here contains a number of important special cases under particular parameter settings. When $\gamma = 0$ and $D = 1$, 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 and Adelson (1998). When $\gamma = 0, D = 1$ and $\lambda \to \infty$, we obtain the "slow and rigid" model of Weiss et al. (2002). When $D = 1$ and $\gamma > 0$, 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, also known as "smoothness in layers" (Wang and Adelson, 1993; Weiss, 1997).

2.2. Inference

The goal of inference is to compute the maximum *a posteriori* motion tree given a set of observations. Recall that the motion tree is completely described by the component assignments $\mathbf{c} = \{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$ and depths $\mathbf{d} = \{d_1, \ldots, d_N\}$. The posterior is given by Bayes' rule:

$$P(\mathbf{c}, \mathbf{d}|\mathbf{s}) \propto P(\mathbf{s}|\mathbf{c}, \mathbf{d})P(\mathbf{c})P(\mathbf{d}), \tag{8}$$

where $\mathbf{s}$ denotes the observed trajectory of object positions. As described above, the latent motion components can be marginalized analytically using properties of Gaussian processes.

We use annealed Gibbs sampling to search for the posterior mode. The algorithm alternates between holding the depth assignments fixed while sampling from the conditional

distribution over component assignments, and holding the component assignments fixed while sampling from the conditional distribution over 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 component assignments $\mathbf{c}_n$ is given by:

$$P(\mathbf{c}_n|\mathbf{c}_{-n}, \mathbf{s}, \mathbf{d}) \propto P(\mathbf{c}_n|\mathbf{c}_{-n})P(\mathbf{s}|\mathbf{c}, \mathbf{d}), \tag{9}$$

where $\mathbf{c}_{-n}$ denotes the set of all paths excluding $\mathbf{c}_n$. The first factor in Eq. 9 is the nCRP prior (Eq. 1). The second factor in Eq. 9 is the likelihood of the data, given by Eq. 6.

The conditional distribution over depth d_n is given by:

$$P(d_n|\mathbf{c}, \mathbf{s}, \mathbf{d}_{-n}) \propto P(d_n|\mathbf{d}_{-n})P(\mathbf{s}|\mathbf{c}, \mathbf{d}), \tag{10}$$

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

$$P(d_n|\mathbf{d}_{-n}) \propto \exp\left\{\alpha \sum_{m \neq n} \mathbb{I}[d_m = d_n] - \rho d_n\right\}. \tag{11}$$

This completes the Gibbs sampler.

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):

$$\mathbb{E}[f_j^x(\mathbf{s}_n(t), t)] = \mathbf{k}_{nj}^\top (\mathbf{K}(t) + \sigma^2 \mathbf{I})^{-1} (\mathbf{s}^x(t+1) - \mathbf{s}^x(t)), \tag{12}$$

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).

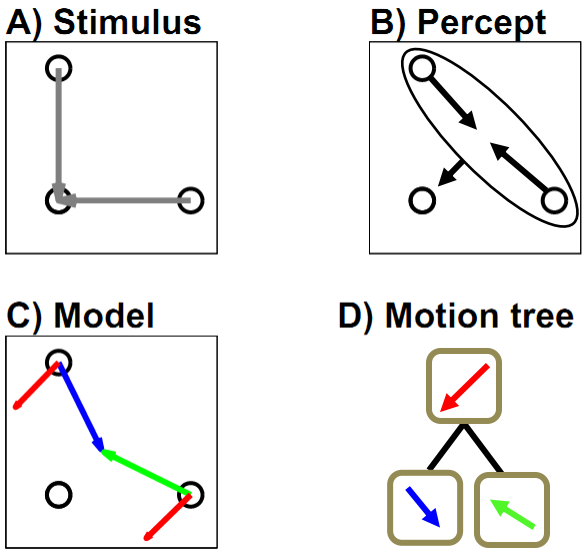


Figure 4: **Johansson (1950) two dot experiment**. (A) Veridical motion vectors. (B) Perceived motion. (C) Inferred motion vectors. Each color corresponds to a different component in the motion tree (D), but note that a component will predict different vectors depending on spatial location.

3. Simulations

In this section, we show how Bayesian vector analysis 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: $D = 3, \sigma^2 = 0.01, \tau = 1, \lambda = 100, \alpha = 1, \rho = 0.1, \gamma = 1$. The interpretation of σ^2 and λ depend on the spatial scale 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 4A 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 4B).

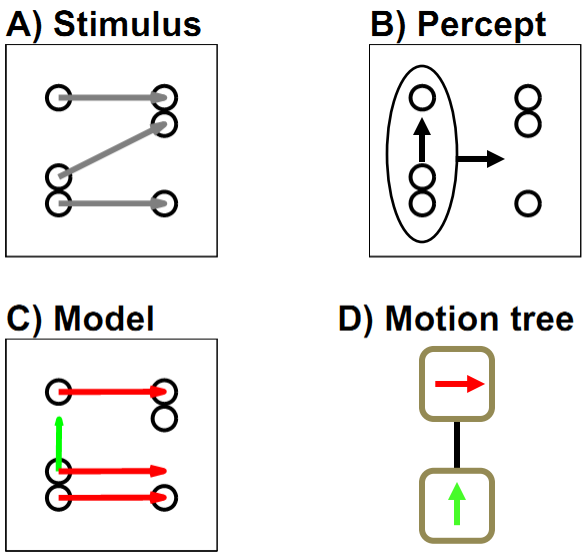


Figure 5: **Johansson (1973) three dot experiment**. (A) Veridical motion vectors. (B) Perceived motion. (C) Inferred motion vectors. (D) Inferred motion tree.

Thus, observers perceive the stimulus as organized into common and relative motions. This percept is reproduced by Bayesian vector analysis (Figure 4C); the inferred motion tree (shown in Figure 4D) 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 because there is uncertainty about the decomposition, leading to partial sharing of structure across components.

Another example studied by Johansson (1973) is shown in Figure 5A (see also Hochberg and Fallon, 1976). 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 5B). Consistent with this percept, 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 5C-D).

Duncker (1929) showed that if a light is placed on the rim of a rolling wheel in a dark

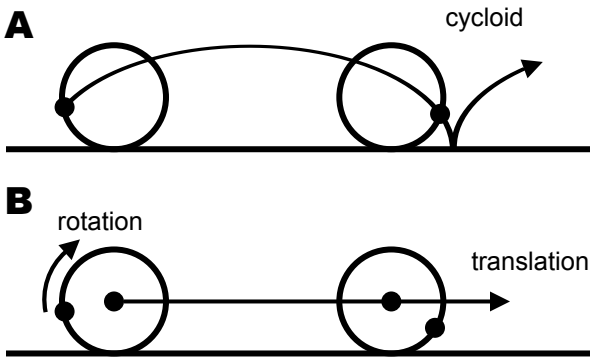


Figure 6: **Duncker wheel**. (*A*) A light on the rim of a rolling wheel, viewed in darkness, produces cycloidal motion. (*B*) Adding a light on the hub produces rolling motion (translation + rotation).

room, cycloidal motion is perceived (Figure 6A), but if another light is placed on the hub then rolling motion is perceived (Figure 6B). Simulations of these experiments are shown in Figure 7. 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.[3] 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 7, bottom).

Bayesian vector analysis can also illuminate the computations underlying motion transparency (Snowden and Verstraten, 1999). 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., two layers in our model) for a range of relative directions using 20 dots. As the relative direc-

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

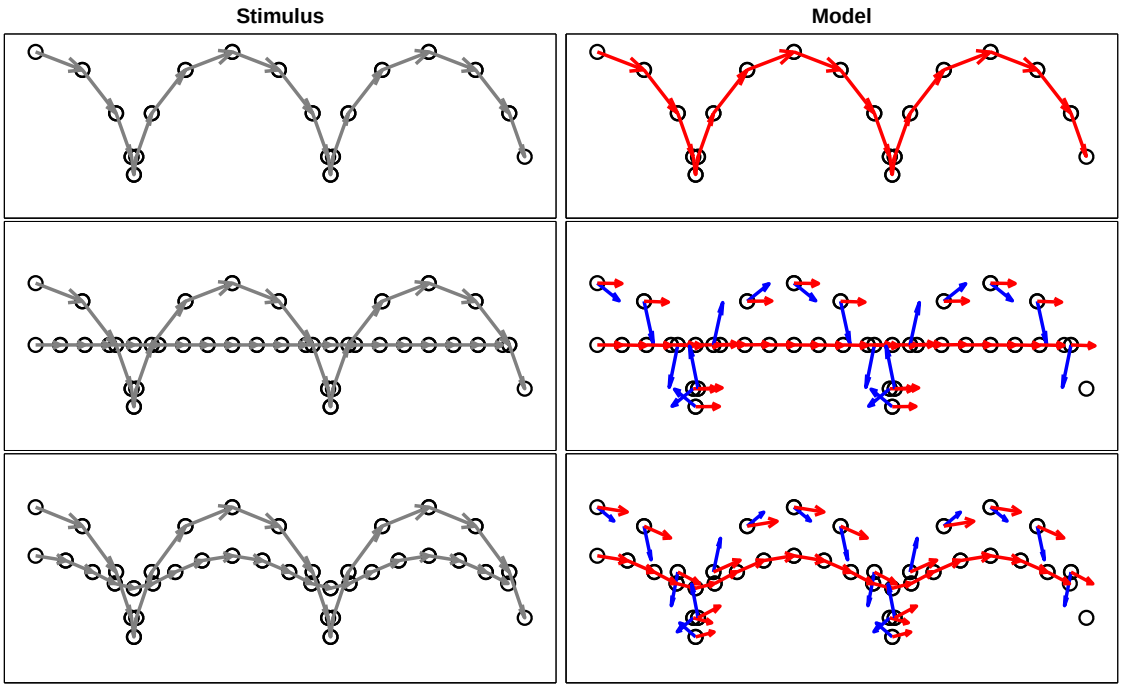


Figure 7: **Simulations of the Duncker wheel**. (*Top*) A single light on the rim produces one vector following a cycloidal path. (*Middle*) Adding a light on the hub produces two vectors: translation + rotation, giving rise to the percept of rolling motion. (*Bottom*) Placing the light on the interior of the wheel produces weaker rolling motion: the translational component is no longer perfectly horizontal.

tion increases, the statistical evidence in favor of two separate layers increases, resulting in a smoothly changing probability (Figure 8). Our simulations are in qualitative agreement with the results of (Braddick et al., 2002).

Inferences about the motion hierarchy may interact with the spatial structure of the scene. The phenomenon of motion contrast, originally described by Loomis and Nakayama (1973), provides an illustration: The perceived motion of a dot depends on the motion of surrounding "background dots" (the black dots in Figure 9A). 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 9A) 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.

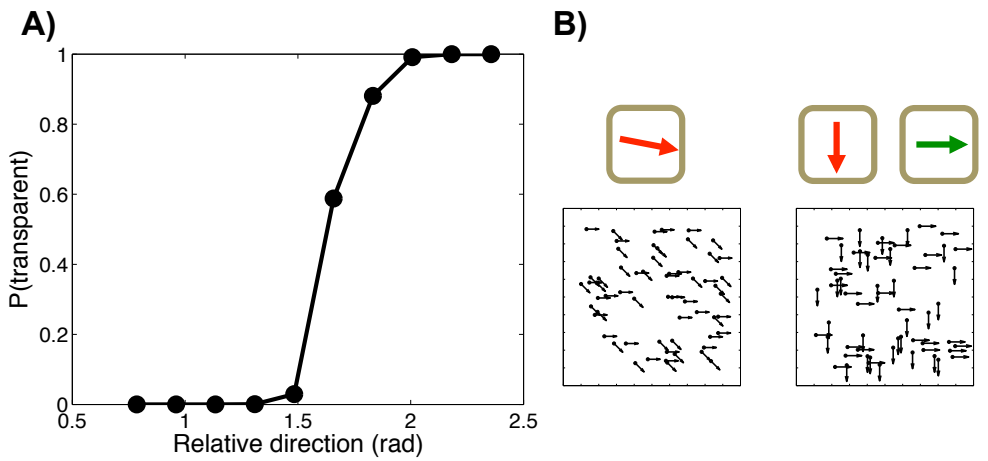


Figure 8: **Simulations of transparent motion**. (*A*) Transparency—the probability of a motion tree with two independent components, each corresponding to a motion layer—increases as a function of direction difference between two superimposed groups of dots. (*B*) Maximum *a posteriori* motion trees for two different motion displays: 45 degree (left) and 90 degree (right) direction difference.

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 9B).

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 and Shiffrar, 2007). We trained our model (with the same parameters) on point light displays derived from the CMU human motion capture database.[4] 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 10: the first layer of motion (not shown) captures the overall trajectory of the

[4] `http://mocap.cs.cmu.edu/`

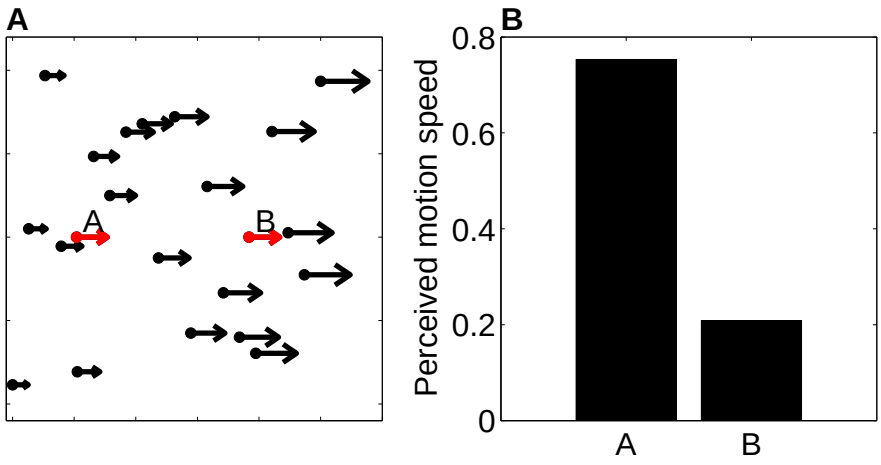


Figure 9: **Motion contrast**. (*A*) The velocity of the background (black) dots increases along the horizontal axis. Although A and B have the same velocity, A is perceived as moving faster than B. (*B*) Model simulation.

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.

Overall, and without much tweaking of the parameters, our model is able to qualitatively capture a wide range of the phenomena that have been observed in hierarchical motion organization. There are two components that are central to the model. The first is Johansson's idea of vector analysis. The representation of the motions is given by a tree-structure where the observed motions are the sum of the motion components of all nodes on the path from the root to the respective object node. The second is that the motion on each node of the tree is represented as a flow field that imposes spatiotemporal smoothness and a preference for slow motions (Weiss and Adelson, 1998; Weiss et al., 2002). The idea of using flow-fields on different motion layers without a hierarchical structure is well established (Koechlin et al., 1999; Nowlan and Sejnowski, 1994; Wang and Adelson, 1993; Weiss, 1997). Our model combines these two ideas in a Bayesian framework and treats the problem as inference over the hierarchical structure. Together with the prior over trees, the model can capture the quali-

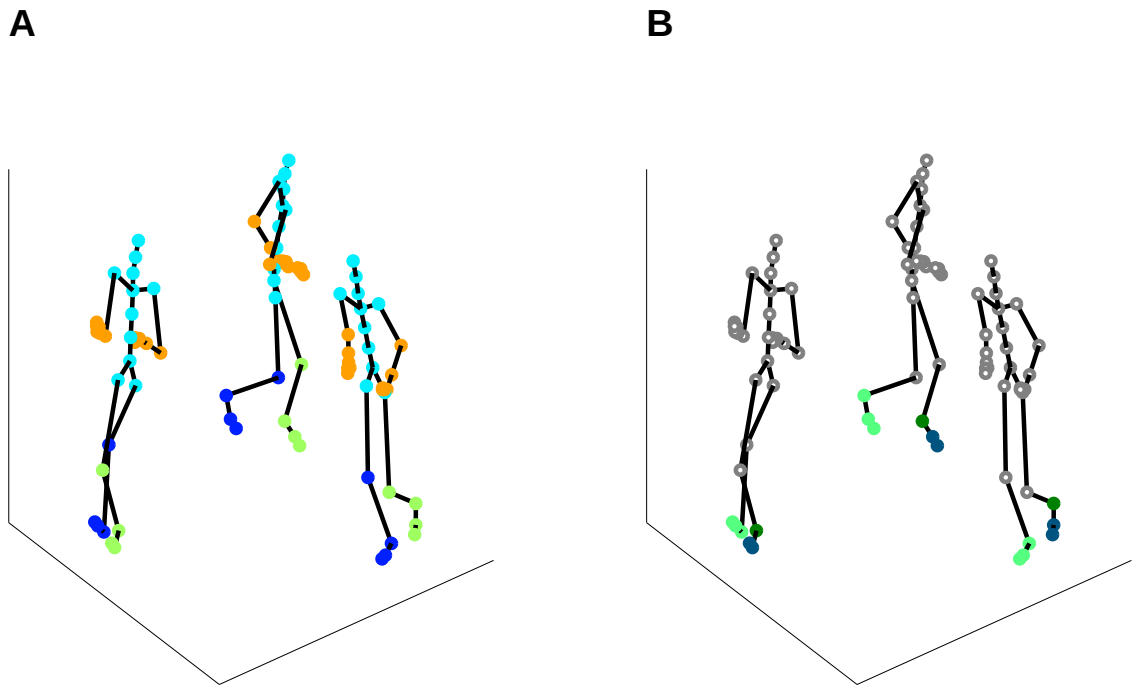


Figure 10: **Analysis of human motion capture data**. Each color represents the assignment of a node to a motion component. All nodes are trivially assigned to the first layer (not shown). In addition, all nodes were assigned to the second layer (A). A subset of the nodes were also assigned components 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.

tative effects of the all the phenomena discussed above. In the following we will present an experiment to test whether the model can also provide quantitative fits to human data.

4. Experiment

In this section, we report a new psychophysical experiment aimed at testing the descriptive adequacy of Bayesian vector analysis. The stimuli were generated using more complicated motion trees than stimuli used in previous research (e.g., Johansson, 1950), and thus provided a more complex challenge for our model.[5]

[5]Demos can be viewed at `https://sites.google.com/site/hierarchicalmotionperception/quintets-iii`.

4.1. Subjects

Four naive, young adult subjects participated in the experiment. All had normal or corrected-to-normal acuity. All subjects gave informed consent. The work was carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

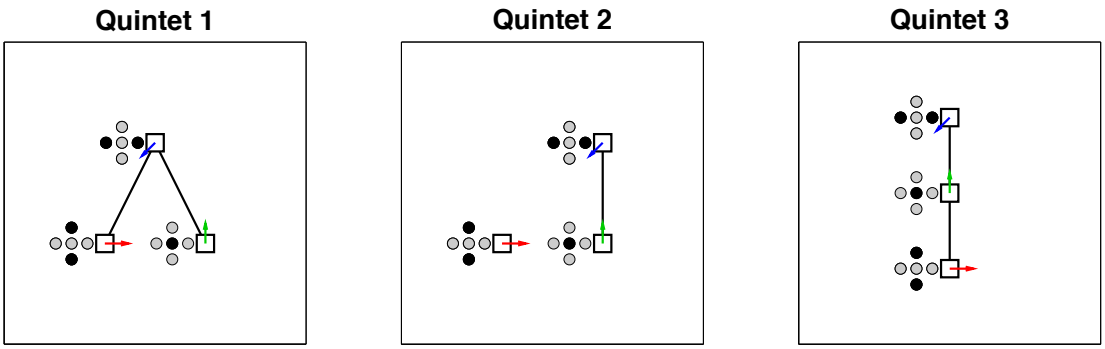
4.2. Stimuli and Procedure

Participants, seated approximately 60 cm from the computer screen (resolution: 1600 × 900; frame rate: 60 Hz), viewed oscillating dot quintets, as schematized in Figure 11. Each dot was circular, subtending a visual angle of 0.43°. An oscillating stimulus was chosen because this is a convenient way to present a single motion structure for a prolonged period of time, and is in keeping with traditional displays (e.g., Johansson, 1950). The quintets were synthesized by combining three different motion components, whose parameters are shown in the bottom panel of Figure 11. The quintets varied in their underlying motion tree structure, while always keeping the underlying motion components the same. In particular, we explored a simple 2-layer tree (Quintet 1), a 2-layer chain plus an independent motion (Quintet 2), and a 3-layer chain (Quintet 3). This allowed us to explore the model predictions for several qualitatively different structures.

On each trial, subjects viewed a single quintet presented on a gray background, with one dot colored red, one dot colored blue, and the rest colored white. The dots are numbered as follows:

1. Top
2. Bottom
3. Center
4. Left
5. Right

Irrespective of the condition (i.e. of the three differing motion trees) the five dots formed a cross at the beginning of each trial. The motion on each component of the motion tree was sinusoidal so that after one cycle the original cross shape would reappear.



	Frequency (deg/sec)	Amplitude (deg)
Component 1	[0.012, 0.012]	[1.17, 1.17]
Component 2	[0.012, 0]	[2.13, 0]
Component 3	[0, 0.024]	[0, 0.90]

Figure 11: **Schematic of dot quintet stimuli**. Motion trees are shown using the same convention as in earlier figures, alongside the spatial arrangement of the quintet. Dots assigned to each component are highlighted in black. For example, the motion of the top dot in Quintet 1 is the sum of the diagonal component (root node, blue vector) and the horizontal component (red vector). Parameters of each component are shown in the table.

The task was to make one of three responses: (1) red moving relative to blue; (2) blue moving relative to red; (3) neither. Subjects were instructed that "red moving relative to blue" should be taken to mean that the blue dot's motion forms a reference frame for the red dot's motion, and hence the red dot is "nested" in the blue dot motion. Subjects were asked to make their response as quickly as possible, but there was no response deadline and the stimulus kept moving on the screen until the subject had responded. Each quintet was shown 32 times, with 8 different dot comparisons.

4.3. Results

To model our data, we used the same parameters as in our simulations reported above, except that we selected τ (the scaling parameter for the Gaussian process prior) to fit the data. Small values of τ diminish the strength of the prior on smooth flow fields, allowing more complex motion patterns. In addition, to more flexibly capture the response patterns, we raised the model's choice probability (i.e., the probability of choosing one of the 3 response

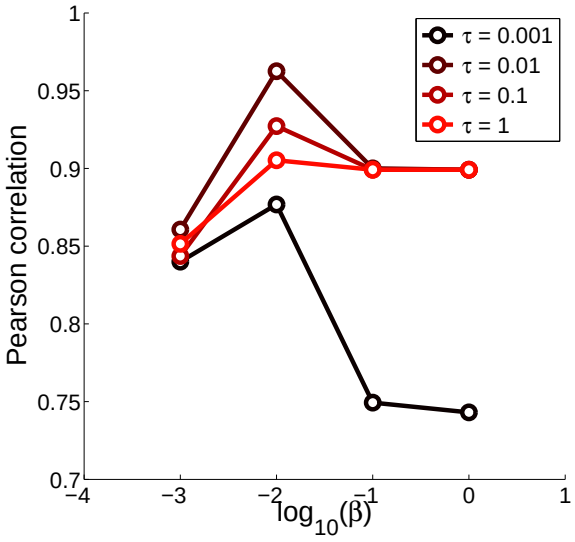


Figure 12: **Cross-validation results**. Pearson correlation between model predictions and experimental data for different values of the scaling parameter τ and the inverse temperature parameter β. The correlation values are computed on held-out (split-half) data and averaged across splits.

options given to subjects) to a power specified by a free parameter, β, and renormalized:

$$\tilde{P}(\mathcal{T} = t|\mathbf{s}) = \frac{P(\mathcal{T} = t|\mathbf{s})^{\beta}}{\sum_{t'} P(\mathcal{T} = t'|\mathbf{s})^{\beta}}, \tag{13}$$

where $\mathcal{T}$ denotes the chosen tree. The consequence of this distortion is to disperse the probability distribution when $\beta < 1$, thus allowing us to model the effect of stochasticity in responding. This kind of response scaling is a common way to map model predictions to subjects' responses in mathematical psychology and Bayesian modeling, in particular (Acerbi et al., 2014; Navarro, 2007). Note that this distortion of probability does not affect ordinal preferences. In practice, when computing the probabilities we did not evaluate all possible motion trees, but instead evaluated the three trees shown in Figure 11, since the data-generating trees are likely to possess most of the posterior probability mass.

To select τ and β, we conducted a coarse grid search, splitting our data into two halves and using one half of the data to select the parameters which maximized the Pearson correlation between model predictions and average choice probabilities. We then computed the correlation on the held-out data as a cross-validated assessment of model fit. The correla-

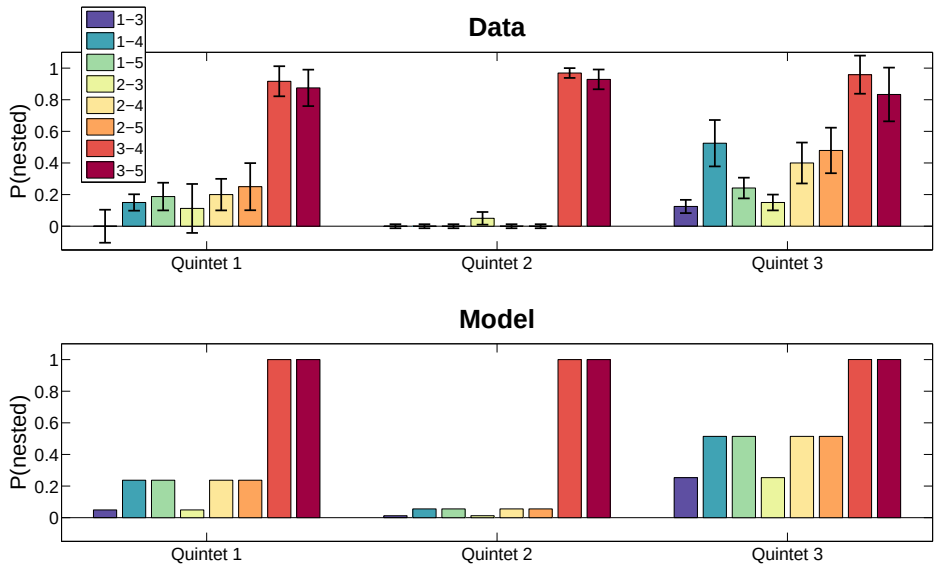


Figure 13: **Experimental data and model predictions**. Each bar represents one dot motion comparison, as denoted in the legend (e.g., "1−3" means "dot 1 nested in dot 3"). The y-axis shows the average probability of reporting that one dot motion is nested in the other dot's motion. Error bars represent standard error of the mean.

tions for the entire parameter range are shown in Figure 12. The optimal parameter setting was found to be $\tau = 0.01$ and $\beta = 0.01$, but the correlations are high for parameters that vary over several orders of magnitude.

Our behavioral data and model predictions are shown in Figure 13 for all 8 comparisons in each quintet. Generally speaking, subjects are able to extract the underlying hierarchical structure, but there is substantial uncertainty; some comparisons are more difficult than others. The entire pattern of response probabilities is very well-captured by our model ($r = 0.98, p < 0.0001$). As a further test of the model, we reasoned that greater uncertainty (measured as the entropy of the choice probability distribution) would require more evidence and hence longer response times. Consistent with this hypothesis, we found that entropy was significantly correlated with response time ($r = 0.73, p < 0.0001$; Figure 14).

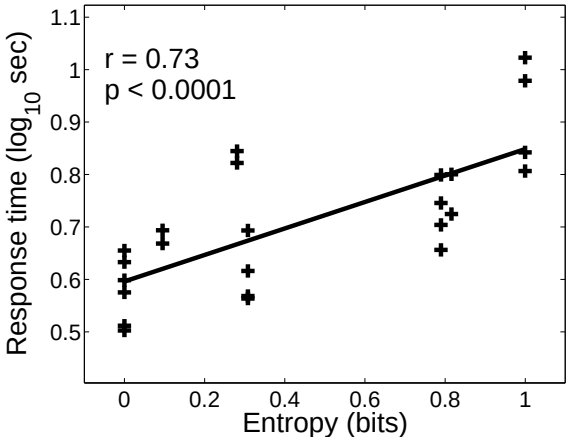


Figure 14: **Response times increase with the entropy of the predicted probability distribution over choices**.

5. Discussion

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. Furthermore, we showed that our model could quantitatively predict complex motion percepts in new experiments. Importantly, the probabilistic representation of uncertainty furnished by our model allowed it to capture variability in choices and response times that we observed in our experimental data.

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. Stochastic search algorithms similar to the one we proposed have been used to model various aspects of visual perception, such as multistability of ambiguous figures (Sundareswara and Schrater, 2008; Gershman et al., 2012). The observation that hierarchically structured motions can also produce multistability (Vanrie et al., 2004) suggests that stochastic search may be a viable algorithmic description, but more work is needed to explore this hypothesis. Our theory also does not commit to any particular neural implementation. 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. Alternatively, neural implementations of stochastic search algorithms (Buesing et al., 2011; Moreno-Bote et al., 2011) would allow us to connect the algorithmic and neural levels.

We view hierarchical motion as a model system for studying more general questions about structured representations in mind and brain (Austerweil et al., 2015; Gershman and Niv, 2010). 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, Liz Spelke, Jeff Beck, Alex Pouget, Yair Weiss, Ted Adelson, Rick Born, 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, the MIT Intelligence Initiative, and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. A preliminary version of this work was presented at the 35th annual Cognitive Science Society meeting (Gershman et al., 2013).

References

Acerbi, L., Vijayakumar, S., and Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology*, 20(6):1–23.

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII*, pages 1–198. Springer, Berlin.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61:183–193.

Austerweil, J., Gershman, S., Tenenbaum, J., and Griffiths, T. (2015). Structure and flexibility in Bayesian models of cognition. In Busemeyer, J., Townsend, J., Wang, Z., and Eidels, A., editors, *Oxford Handbook of Computational and Mathematical Psychology*. Oxford University Press, Oxford.

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

Blei, D., Griffiths, T., and Jordan, M. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57:1–30.

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

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

Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7:e1002211.

Burke, L. (1952). On the tunnel effect. *Quarterly Journal of Experimental Psychology*, 4:121–138.

DiVita, J. C. and Rock, I. (1997). A belongingness principle of motion perception. *Journal of Experimental Psychology: Human Perception and Performance*, 23:1343–1352.

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

Flombaum, J. I., Kundey, S. M., Santos, L. R., and Scholl, B. J. (2004). Dynamic object individuation in rhesus macaques: a study of the tunnel effect. *Psychological Science*, 15:795–800.

Flombaum, J. I. and Scholl, B. J. (2006). A temporal same-object advantage in the tunnel effect: facilitated change detection for persisting objects. *Journal of Experimental Psychology: Human Perception and Performance*, 32:840–853.

Gershman, S. J., Jäkel, F., and Tenenbaum, J. B. (2013). Bayesian vector analysis and the perception of hierarchical motion. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Gershman, S. J. and Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, 20:251–256.

Gershman, S. J., Vul, E., and Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, 24:1–24.

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

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

Hochberg, J. and Fallon, P. (1976). Perceptual analysis of moving patterns. *Science*, 194:1081–1083.

Hochberg, J. and McAlister, E. (1953). A quantitative approach, to figural "goodness". *Journal of Experimental Psychology*, 46:361–364.

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*, 14:201–211.

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

Koechlin, E., Anton, J. L., and Burnod, Y. (1999). Bayesian inference in populations of cortical neurons: a model of motion integration and segmentation in area mt. *Biological Cybernetics*, 80:25–44.

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

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

Moreno-Bote, R., Knill, D. C., and Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108:12491–12496.

Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, 51:85–98.

Nowlan, S. and Sejnowski, T. (1994). Filter selection model for motion segmentation and velocity integration. *JOSA A*, 11:3177–3200.

Pitman, J. (2002). *Combinatorial Stochastic Processes*. Notes for Saint Flour Summer School. Techincal Report 621, Dept. Statistics, UC Berkeley.

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

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

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

Shum, K. H. and Wolford, G. L. (1983). A quantitative study of perceptual vector analysis. *Perception & Psychophysics*, 34:17–24.

Snowden, R. J. and Verstraten, F. A. (1999). Motion transparency: making models of motion perception transparent. *Trends in Cognitive Sciences*, 3:369–377.

Sundareswara, R. and Schrater, P. R. (2008). Perceptual multistability predicted by search model for bayesian decisions. *Journal of Vision*, 8:1–19.

Vanrie, J., Dekeyser, M., and Verfaillie, K. (2004). Bistability and biasing effects in the perception of ambiguous point-light walkers. *Perception*, 33:547–560.

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

Wang, J. and Adelson, E. (1993). Layered representation for motion analysis. In *Computer Vision and Pattern Recognition*, pages 361–366. IEEE.

Weiss, Y. (1997). Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *Computer Vision and Pattern Recognition*, pages 520–526. IEEE.

Weiss, Y. and 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., and Adelson, E. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5:598–604.

Wertheimer, M. (1923). Untersuchungen zur lehre von der gestalt, ii. [investigations in gestalt theory: Ii. laws of organization in perceptual forms]. *Psychologische Forschung*, 4:301–350.

Xu, F. and Carey, S. (1996). Infants metaphysics: The case of numerical identity. *Cognitive Psychology*, 30:111–153.

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