A space-variant model for motion interpretation across the visual field

Manuela Chessa
Department of Informatics, Bioengineering, Robotics and System Engineering, University of Genoa, Genoa, Italy

Guido Maiello
UCL Institute of Ophthalmology, University College London, London, UK
Department of Psychology, Northeastern University, Boston, MA, USA

Peter J. Bex
Department of Psychology, Northeastern University, Boston, MA, USA

Fabio Solari
Department of Informatics, Bioengineering, Robotics and System Engineering, University of Genoa, Genoa, Italy

We implement a neural model for the estimation of the focus of radial motion (FRM) at different retinal locations and assess the model by comparing its results with respect to the precision with which human observers can estimate the FRM in naturalistic motion stimuli. The model describes the deep hierarchy of the first stages of the dorsal visual pathway and is space variant, since it takes into account the retino-cortical transformation of the primate visual system through log-polar mapping. The log-polar transform of the retinal image is the input to the cortical motion-estimation stage, where optic flow is computed by a three-layer neural population. The sensitivity to complex motion patterns that has been found in area MST is modeled through a population of adaptive templates. The first-order description of cortical optic flow is derived from the responses of the adaptive templates. Information about self-motion (e.g., direction of heading) is estimated by combining the first-order descriptors computed in the cortical domain. The model's performance at FRM estimation as a function of retinal eccentricity neatly maps onto data from human observers. We further show that the model is able to estimate direction of heading in real-world scenes, thus validating the model's potential application to neuromimetic robotic architectures. More broadly, we provide a framework in which to model complex motion integration across the visual field in real-world scenes.

Introduction

In this article, we examine the role of central and peripheral vision in the dorsal visual pathway (Goodale & Westwood, 2004; G. A. Orban, 2008) for the processing of self-motion and object motion (Koenderink, 1986; G. Orban et al., 1992; Perrone & Stone, 1994; Simoncelli & Heeger, 1998; Watson, Ahumada, et al., 1985). More specifically, we propose a model that is able to compute the focus of expansion (FOE), which is the point from which all motion vectors expand, indicating the direction of heading when eye and head positions are known. Humans can accurately judge direction of heading by observing the location of the FOE in the flow field (Warren & Hannon, 1988). The focus of radial motion (FRM) is a generalization of the FOE: It is the point from which all motion vectors either expand or contract. Bex and Falkenberg (2006)
showed that the accuracy with which human observers can estimate the FRM in random-dot stimuli decreases as a function of retinal eccentricity and, using equivalent-noise analysis, that this effect may be due to an increase in local-motion-detector noise rather than undersampling.

We aim to model the processes of direction detection and motion pooling which support the perception of optic flow. We propose a neural model which implements the deep hierarchy of the first stages of the dorsal visual pathway (Solari, Chessa, & Sabatini, 2014). Such a model (see Figure 1) is space variant, since it takes into account the retina-cortical transformation of the primate visual system through log-polar mapping that produces a cortical representation of the visual signal to the retina. We study how this space-variant approach affects optic-flow computations across the visual field. Specifically, we ask whether a model which implements the known processing stages of the dorsal visual pathway is sufficient to account for human perception. In particular, in this article we (a) propose a neural model for the estimation of the FRM at different retinal locations and (b) assess the model by comparing its results with respect to (c) the experimental evidence related to the precision with which human observers can estimate the FRM in naturalistic, dead-leaves stimuli (Bordenave, Gousseau, & Roueff, 2006; Lee, Mumford, & Huang, 2001).

The first stage of the proposed model implements the retino-cortical transformation by mapping the Cartesian retinal image onto log-polar cortical coordinates. The log-polar transform of the retinal image is the input to the cortical motion-estimation stage, where optic flow is computed. The cortical motion-estimation stage consists of a three-layer population of cells. A population of spatiotemporal-oriented Gabor filters approximates the simple cells of area V1 (first layer), which are combined into complex cells as motion energy units (second layer). The responses of the complex cells are pooled (third layer) to encode the magnitude and direction of velocities, as in the extrastriate motion pathway, between areas MT and MST. The sensitivity to complex motion patterns that has been found in area MST is modeled through a population of adaptive templates, and from the responses of such a population the first-order description of optic flow is derived. Information about self-motion (e.g., direction of heading) is estimated by combining such first-order descriptors computed in the cortical domain.

We thus investigate whether a fully functional, space-variant model of the known processing stages of the dorsal visual pathway accounts for human performance across the visual field. Such a model must take real image sequences as input, perform space-variant computations on the input images, and output behaviorally relevant metrics such as the estimate of the FRM. From our investigation we conclude that our model shows qualitatively and quantitatively similar patterns of results to the human behavioral data. The proposed neural model thus captures essential aspects of the neural computations that occur in the cortical motion pathway. We also discuss which aspects of human perception are not captured by the model and might be interesting starting points for further investigation. Our work has clear applications in neuromi-
metic robotic architectures, and more broadly, we provide a framework in which to model complex motion integration across the visual field.

Related work

From neurophysiological studies it is well known that neurons in the dorsal part of the medial superior temporal area (MSTd) are selective for optic-flow patterns (Tanaka & Saito, 1989) generated by the relative motion between the observer and the 3-D world. MSTd neurons are known to respond to elementary (expansion, contraction, rotation, and translation) as well as complex patterns of optic flow (Duffy & Wurtz, 1991; Graziano, Andersen, & Snowden, 1994). It has been shown that activity in MSTd neurons that are selective for expanding optic-flow fields is linked with behavioral performance in heading judgments (Gu, DeAngelis, & Angelaki, 2012; Gu, Fetsch, Adeyemo, DeAngelis, & Angelaki, 2010). Xu, Wallisch, and Bradley (2014) conclude that a variety of subpopulations of MSTd neurons, differently tuned for expansion, rotation, or spiral motion, contribute to heading judgments. Graziano et al. (1994) analyzed the idea that navigation is achieved by decomposing the optic flow into these elementary components. They found that many cells in area MST are tuned to intermediate spiral motions, obtained by the combination of expansion and rotation components. These findings reinforced the idea that area MST might process complex configurations of visual motion, to obtain information on both navigation and motion of objects and surfaces. In the human brain, the analysis of visual motion continues in higher processing layers. Duffy and Wurtz (1995) found that many neurons in area MST respond to a combination of expansion and planar motions which shift the position of the FOE in the neuron’s receptive field. In particular, they found a continuum in the MST neurons’ responses, which combine expansion, rotation, and planar motions. Recently Xu et al. (2014) described the spiral space model: They propose that neurons in area MSTd are tuned to elementary optic-flow types (and consequently to motion in real-world space) in a continuous way, which we call spiral-space tuning. In particular, they propose that adding laminar motion (e.g., a translation to the left) to radial motion (e.g., expansion) yields tuning for a particular FRM location or direction of heading in real situation. Their findings suggest that MSTd neurons respond to all possible motion types in a 3-D space (called the spiral space) whose axes are the expansion, rotation, and laminar-motion components (see Xu et al., 2014, figure 1).

This rich array of neurophysiological findings has led to the development of multiple computational models of optic-flow processing and heading perception. Lappe and Rauschecker (1993) propose a neural model that detects the direction of ego-motion from optic flow in a fashion that is consistent with neurophysiological and psychophysical data. That model, however, does not implement optic-flow estimation, but describes the neural computations occurring after layer MT. A “proof of principle” of a template-like strategy for heading estimation by following a strategy similar to MSTd neurons is presented by Perrone and Stone (1994). The input units of such a model are idealized MT neurons. Grossberg, Mingolla, and Pack (1999) describe a neural model that provides both functional explanation and quantitative simulation of experimental data of MSTd cells. They interestingly take into account the retino-cortical mapping by providing log-polar representations of optic-flow patterns as inputs to their model. Beardsley and Vaina (2001) describe a model population of MST neurons capable of performing graded motion-pattern discrimination tasks. More recently, the ViSTARS neural model proposed by Browning, Grossberg, and Mingolla (2009) describes interaction among neurons in the primate magnocellular pathway and is capable of estimating heading direction from natural image sequences. The ViSTARS model, however, does not consider the retino-cortical transformation, and is thus space invariant.

We build on this body of previous work by developing a model which incorporates known aspects of neural processing relevant for motion estimation and interpretation. The retinal image input to our model is mapped onto cortical space, where optic flow is estimated and decoded directly in cortical coordinates to provide various behaviorally relevant metrics, including direction of heading.

Model description

The biologically inspired neural model of the dorsal motion pathway can be summarized as follows.

Retino-cortical mapping

Each Cartesian moving image sequence is transformed into its cortical representation through a log-polar transformation that implements the central blind-spot model. Such a transformation mimics the retino-cortical mapping of the primate visual system (Schwartz, 1977; Traver & Pla, 2008; Wilkinson, Anderson, Bradley, & Thibos, 2016).

In the central blind-spot model, the mapping from the Cartesian domain \((x, y)\) to the cortical domain of
coordinates \((\xi, \eta)\) is described by the following equations:

\[
\begin{align*}
\xi &= \log_a \left( \frac{\rho}{\rho_0} \right) \\
\eta &= q \psi,
\end{align*}
\]

where \(a\) parameterizes the nonlinearity of the mapping, \(q\) is related to the angular resolution, \(\rho_0\) is the radius of the central blind spot, and \((\rho, \psi) = \left(\sqrt{x^2 + y^2}, \arctan(y/x)\right)\) are the polar coordinates derived from the Cartesian ones. All points with \(\rho < \rho_0\) are ignored (hence the central blind spot), thus \(\rho_0\) must be small with respect to the size of the image.

### Discrete log-polar mapping

Because the transformation is implemented on digital images, given a Cartesian image of \(N_x \times N_y\) pixels and defined \(\rho_{\max} = 0.5 \min(N_x, N_y)\), we obtain an \(R \times S\) (rings \(\times\) sectors) discrete cortical image of coordinates \((u, v)\) by taking:

\[
\begin{align*}
\u &= \lfloor \xi \rfloor \\
\v &= \lfloor \eta \rfloor,
\end{align*}
\]

where \(\lfloor \cdot \rfloor\) denotes the integer part, \(q = S/(2\pi)\), and \(a = \exp(\ln(\rho_{\max}/\rho_0)/R)\). Moreover, we can define the compression ratio (CR) of the cortical image with respect to the Cartesian one as

\[
\text{CR} = \frac{N_x \times N_y}{R \times S}. \quad (3)
\]

Figure 2 shows the log-polar pixels (which can be thought of as the log-polar receptive fields) in the Cartesian and cortical domains. The red circular curve in Figure 2a, with radius \(S/2\pi\), represents the locus where the size of log-polar pixels is equal to the size of Cartesian pixels. In particular, in the area inside the red circular curve (the fovea) a single Cartesian pixel contributes to many log-polar pixels (oversampling), whereas outside this region multiple Cartesian pixels will contribute to a single log-polar pixel. This property is highlighted in the log-polar pixel bordering in violet in Figure 2a. The retinal area (i.e., the log-polar pixel) that refers to a given cortical pixel defines the cortical pixel's receptive field (RF). To avoid spatial aliasing due to the undersampling, we employ overlapping RFs (see RF shape, later, for details). An example of transformation from the Cartesian domain to the cortical domain and back to the retinal one is shown in Figure 2, for the standard image Lena and for a frame of the dead-leaves stimuli used in the experiments described later. It is worth noting that the cortical image (Figure 2c) shows the effects of the log-polar mapping: In particular, the zoomed cortical image (Figure 2c) shows that the eye, which is in fovea in Figure 2d, is overrepresented (left-bottom of Figure 2c), whereas the hat feathers, which are in periphery in Figure 2d, is underrepresented (right-middle of Figure 2c)—i.e., few neural units code the visual information of the periphery. Looking at the backward mapped image (Figure 2f), the eye has full resolution, since it is in fovea, whereas the hat feathers have low resolution due to the neural underrepresentation.

Through inverting Equation 1, the centers of the RFs can be computed, and these points present a nonuniform distribution throughout the retinal plane (see the yellow and black circles overlying the Cartesian images of Lena and dead leaves, respectively, in Figure 2d through g).

The optimal relationship between \(R\) and \(S\) is the one that optimizes the log-polar pixel aspect ratio \(\gamma\), making it as close as possible to 1 (see Motion estimation, later, for details). It can be shown (Solari, Chessa, & Sabatini, 2012; Traver & Pla, 2008) that for a given \(R\), the optimal rule is \(S = 2\pi/(a - 1)\).

The size of the RFs increases as a function of the eccentricity (the distance between the center of the RF and the fovea). We can define the relationship between the RF size (in particular, the maximum RF size \(W_{\max}\)) and the parameters of the mapping as follows (Solari et al., 2012):

\[
W_{\max} = \rho_0 a^R (1 - a^{-1}). \quad (4)
\]

The parameters of the log-polar mapping also influence the proportion of cortical units used to overrepresent the fovea. In particular, it is possible to define the percentage of the cortical area used to represent the fovea \(q\). This can be derived from Equation 4 by setting the RF size to 1, inverting the equation to find the corresponding \(u\) (see Equation 2), and dividing for the overall size of the modeled cortex \(R\):

\[
\frac{1 - \log_a (\rho_0 (a - 1))}{R}. \quad (5)
\]

By exploiting Equations 4 and 5 we can control the growth of the size of the RFs and the overrepresentation of the fovea in order to reproduce data from the literature on size-to-eccentricity relationships (Freeman & Simoncelli, 2011; Wilkinson et al., 2016; Wurbs, Mingolla, & Yazdanbakhsh, 2013).

### RF shape

The shape of the RFs employed for the mapping affects both the quality of the transformation and its computational burden. In this article, we employed overlapping circular RFs (Bolduc & Levine, 1998;
Figure 2. The retino-cortical mapping. (a) Cartesian domain \((x, y)\) with overlying log-polar pixels—i.e., the RFs (the circles). (b) Cortical domain \((\xi, \eta)\), where the squares denote the neural units. The magenta and the blue areas in (a) represent two log-polar pixels at different angular and radial positions (thus with different \(w\) and \(h\)) that correspond to two cortical pixels—the magenta and the blue squares in (b)—of equal size. The orange circle of RFs and the green sector of RFs in the Cartesian domain (a) map to vertical and horizontal stripes of neural units, respectively, in the cortical domain (b). The red circular curve in (a) delimits the oversampling and undersampling areas, and the area inside it is the fovea. An example of image transformation from the Cartesian (d) to the cortical domain (e), enlarged in (c) to better appreciate the distortions, and backward to the retinal domain (f). The latter is shown for completeness, though it is not used in our approach. The RFs, the yellow circles in (d), are overlying the Cartesian image. The specific choices of parameters are \(R = 60\), \(S = 93\), \(\rho_0 = 5\), \(\rho_{\text{max}} = 512\), and \(\text{CR} = 47\). (g) In the green box, the image transformation applied to a dead-leaves stimulus image used in the experiments is shown. The specific choices of parameters are \(R = 110\), \(S = 176\), \(\rho_0 = 10\), \(\rho_{\text{max}} = 960\), and \(\text{CR} = 47\). The RFs and the circle delimiting the oversampling and undersampling areas are represented in black so as to not affect the color of the stimuli in the visualization.
Pamplona & Bernardino, 2009), which are the most biologically plausible and optimally preserve image information (Chessa, Sabatini, Solari, & Tatti, 2011). To implement the log-polar mapping, the cortical plane is divided in two regions: the fovea and the periphery. The periphery is defined as the part of the plane in which the distance between the centers of two RFs on the same radius is greater than 1 pixel (undersampling). To obtain the cortical image we use overlapping Gaussian RFs, as shown in Figure 2a. The fovea (in which we have an oversampling — i.e., the distance between two consecutive RFs is less than 1 pixel) is handled by using fixed-size RFs, whereas in the periphery the size of the RFs grows. The standard deviation of the RF Gaussian profile is a third of the spatial support — i.e., the width of the RF — is 6 times the standard deviation. As a consequence of this choice, adjacent RFs overlap. A cortical pixel $C_i$ is computed as a Gaussian weighted sum of the Cartesian pixels $P_j$ in the $i$th RF: $C_i = \sum w_{ij} P_j$, where the weights $w_{ij}$ are the values of a normalized Gaussian centered on the $i$th RF. A similar approach is used to compute the inverse log-polar mapping that produces the retinal image, where the space-variant effect of the log-polar transformation has a negligible effect on the RFs of the neural model cells. Thus we can exploit the advantages of log-polar mapping without the drawbacks (i.e., ad hoc modifications of the model) related to the distortions of the filters in the cortical domain. In particular, we adopt specific constraints on the aspect ratio of the log-polar pixel (i.e., $\gamma = 1$) and on the spatial support of the filter in order to obtain undistorted RFs, since the first stage of the proposed feed-forward architecture is based on spatiotemporal filtering. Thus, the optic flow can be computed from a sequence of cortical images by using the V1-MT feed-forward architecture originally designed in the Cartesian domain.

Motion estimation

Motion estimation — i.e., the computation of the cortical optic flow — is achieved by using a V1-MT feed-forward architecture, derived from the model proposed by Simoncelli and Heeger (1998). The V1-MT feed-forward architecture, for the computation of optic flow in the Cartesian domain, has been described by Solari, Chessa, Medathati, and Kornprobst (2015). There a set of rules is presented for designing a discrete log-polar mapping that allows a direct application in the log-polar domain of the algorithms, based on spatial multiscale and multiorientation filtering, originally developed for the Cartesian domain without ad hoc modifications. The design rules are summarized as follows:

(a) The aspect ratio $\gamma$ of the log-polar pixel has to be close to 1. This allows the extraction of image features in the cortical domain by applying the same local operators (e.g., filtering) employed in the Cartesian domain.

(b) The spatial support of the local operators has to be small on the cortical domain.

(c) The mapping of a vector field of image features has to be expressed in terms of general-coordinate transformation.

With these proposed rules, the distortion due to the log-polar transformation has a negligible effect on the RFs of the neural model cells. Thus we can exploit the advantages of log-polar mapping without the drawbacks (i.e., ad hoc modifications of the model) related to the distortions of the filters in the cortical domain. In particular, we adopt specific constraints on the aspect ratio of the log-polar pixel (i.e., $\gamma = 1$) and on the spatial support of the filter in order to obtain undistorted RFs, since the first stage of the proposed feed-forward architecture is based on spatiotemporal filtering. Thus, the optic flow can be computed from a sequence of cortical images by using the V1-MT feed-forward architecture originally designed in the Cartesian domain.

Such an architecture is a three-step feed-forward model: Step 1 corresponds to the V1 simple and complex cells, Step 2 corresponds to the MT pattern cells, and Step 3 corresponds to a decoding stage to obtain the optic flow from the MT population response.

Step 1: V1 (motion-energy estimation and normalization)

In the V1 layer of the model, two populations of neurons are involved in the processing of information, namely V1 direction-selective simple cells and complex cells. Simple cells are characterized by the preferred (spatial) orientation $\theta$ of their contrast sensitivity in the spatial domain and their preferred velocity $v^x$ in the direction orthogonal to their contrast orientation, often referred to as component speed.

The RFs of the V1 simple cells are classically modeled using band-pass filters in the spatiotemporal domain $(x, y, t)$. In order to achieve low computational complexity, the spatiotemporal filters $g(x, y, \theta, f_s, f_t)$ are decomposed into separable filters in space $h(x, y, \theta, f_s)$ and time $p(t, f_t)$. The spatial component of the RF is described by Gabor filters,

$$h(x, y, \theta, f_s) = B e^{\left(-\frac{(x^2+y^2)}{2\sigma^2}\right)} e^{2\pi(f_s \cos(\theta)x + f_t \sin(\theta)y)}, \quad (6)$$

and the temporal component by an exponential decay function

$$p(t, f_t) = e^{(-\tau)} e^{2\pi(f_t \tau)}, \quad (7)$$

where $f_s$ and $f_t$ are the spatial and temporal peak frequencies, related to the preferred velocity by $v^x = f_s / f_t$, and $\sigma$ and $\tau$ define the spatial and temporal scales, respectively.
The parameter values chosen in the current model implementation are \( f_s = 0.25 \) c/pixel, \( f_i = [0, 0.10, 0.15, 0.23] \) c/frame (population of simple cells tuned to different preferred velocities), \( \sigma = 2.27 \) pixels, and \( \tau = 2.5 \) frames. The filters’ spatial orientations were chosen to be \( \theta = \pi i / 8 \), \( i = 0, 1, \ldots, 7 \).

We adopt a multiscale, multi-orientation decomposition as in Freeman and Simoncelli (2011). In particular, we use a bank of filters of different spatial orientations, and we adopt a specific multiscale approach. The multiscale decomposition is based on a standard pyramidal approach (Burt & Adelson, 1983), which can be considered as a vertical multiscale—i.e., the variation of the filter size at a single location—and a horizontal multiscale, where the log-polar spatial sampling produces a multiscale for different spatial locations—i.e., the variation of the filter size across the visual field (Bonmassar & Schwartz, 1997).

In order to process the cortically transformed images, it is necessary to characterize the filters, defined in the Cartesian domain, with respect to the cortical domain—i.e., to map the filters into the cortical domain—thus obtaining \( g(x(\xi, \eta), y(\xi, \eta), \theta, f_s, f_i) \). As a consequence of the nonlinearity of the log-polar mapping, the mapped filters are distorted; Solari et al. (2012) have shown that under specific conditions such distortions can be kept to a minimum. This happens when the spatial support of the RFs is sufficiently small and the aspect ratio \( \gamma \) of the log-polar pixel is equal to 1 (see Figure 3, top). Under these assumptions, it is possible to work directly in the cortical domain by considering spatiotemporal filters sampled in log-polar coordinates \( g(\xi, \eta, \theta, f_s, f_i) \)—see Figure 3 (bottom).

Given the response of the simple cells layer, the V1 complex cells are described as a combination of the quadrature pair of simple cells by using the motion-energy formulation (Adelson & Bergen, 1985), followed by a divisive normalization (Heeger, 1992) denoted \( E^v(\xi, \eta, t, \theta, v') \). A key property of V1 cells is their tuning to the spatial orientation and velocity of a stimulus, which arises from spatiotemporal-frequency selectivity for motion in a direction perpendicular to the contrast of the underlying pattern (Adelson & Movshon, 1982).

**Step 2: MT pattern-cell response**

MT neurons exhibit velocity tuning irrespective of the contrast orientation. This is believed to be achieved through pooling of different responses from V1 layers, in both spatial and orientation domains, followed by a nonlinearity. In particular, in the proposed model we perform the following processing:

The output of the V1 afferent cells is spatially pooled through a Gaussian kernel.

The previous output is pooled by MT linear weights, which give rise to the MT tuning to speed direction \( d \); these weights are defined as \( \cos(d - \theta) \) where \( d \in [0, 2\pi] \).

The output of the MT orientation pooling is then fed into an exponential function which describes the static nonlinearity.

The responses of an MT pattern cell tuned to the speed \( v^c \) and to direction of speed \( d \) are denoted \( E^{MT}(\xi, \eta, t, d, v^c) \).

**Step 3: Decoding**

In this step, optic flow is estimated by decoding the population responses of the MT neurons. Indeed, a unique velocity vector cannot be recovered from the activity of a single velocity-tuned MT neuron, as multiple scenarios could evoke the same activity, but a unique vector can be recovered from the activity of a population. Here we adopt a linear combination approach to decode the MT population response, as described previously (Pouget, Zhang, Deneve, & Latham, 1998; Rad & Paninski, 2011).

In particular, we linearly decode the MT representation of the speeds \( v^c \) and the speed directions \( d \); it is possible to first decode the MT responses \( E^{MT}(\xi, \eta, t, d, v^c) \) along each speed direction \( d \) to compute the speed, then to apply the intersection of constraints on such estimated velocities.

The estimates \( v^d \) of the speed along each speed direction \( d \) can be obtained as follows (i.e., linear combination):

\[
v^d(\xi, \eta, t, d) = \sum_{v^c} v^c E^{MT}(\xi, \eta, t, d, v^c). \tag{8}
\]

The estimate of the full cortical velocity is then

\[
v^c = \frac{2}{Q} \sum_{d_i=d} v^d(\xi, \eta, t, d_i) \cos d_i \tag{9}\]

\[
v^c = \frac{2}{Q} \sum_{d_i=d} v^d(\xi, \eta, t, d_i) \sin d_i.
\]

It is worth noting that an expansion in the Cartesian domain is represented by a constant flow along the \( \xi \) axis in the cortical domain, but if the FRM is shifted with respect to the center of the visual field (i.e., we have an expansion with a constant translation), the corresponding cortical flow has nonlinear components, as shown in Figure 4. Moreover, the cortical flow generated by a shift of the FRM toward the right is notably different from the cortical flow corresponding to a shift of the FRM toward the left.
Motion interpretation

Given the sensitivity of MSTd neurons for elementary flow components (EFCs) alone or their combination with translation components (G. Orban et al., 1992), here we consider a population of cells which model the tuning to EFCs (see Figure 5).

In particular, considering the Cartesian (i.e., retinal) space, we have four classes of deformation gradients: one stretching and one shearing, for each cardinal direction; and two translation components. The deformation and translation templates can be combined to obtain deformation subspaces representing EFCs such as expansion, shear, and rotation. It is worth noting that the expansion and rotation subspaces represented in Figure 5 are slices that can be obtained from the spiral space model presented by Xu et al. (2014).

In the Cartesian domain, the EFCs can be described in terms of affine description (for details, see Chessa, Solari, & Sabatini, 2013), and Solari et al. (2014) show that the affine description in the Cartesian domain can be recovered from the cortical affine description. Moreover, that article derives the relationships among the first-order description of the cortical flow and the estimation of the 3-D rigid-body motions. Thus, it is possible to recover a description of the EFCs by working directly on the cortical optic flow.

The cortical motion field can be described as linear deformations by a first-order Taylor decomposition, around each cortical image point: $\mathbf{v} = \mathbf{v} + \mathbf{T}[\zeta, \eta]^\dagger$, where $\mathbf{T}$ is the tensor composed of the partial derivatives of the cortical motion field. By describing the tensor through its dyadic components, we obtain

$$\mathbf{v} = \mathbf{a}^\zeta \mathbf{v}_\zeta + \mathbf{a}^\eta \mathbf{v}_\eta + \mathbf{d}_\zeta^y \frac{\partial v_z}{\partial \xi} |_{(\zeta_0, \eta_0)} + \mathbf{d}_\eta^y \frac{\partial v_z}{\partial \eta} |_{(\zeta_0, \eta_0)} + \mathbf{d}_\zeta^x \frac{\partial v_y}{\partial \xi} |_{(\zeta_0, \eta_0)} + \mathbf{d}_\eta^x \frac{\partial v_y}{\partial \eta} |_{(\zeta_0, \eta_0)},$$

where $\mathbf{a}^\zeta: (\zeta, \eta) \mapsto (1, 0)$ and $\mathbf{a}^\eta: (\zeta, \eta) \mapsto (0, 1)$ are pure translations and $\mathbf{d}_\zeta^y: (\zeta, \eta) \mapsto (\zeta, 0)$, $\mathbf{d}_\eta^y: (\zeta, \eta) \mapsto (\eta, 0)$, $\mathbf{d}_\zeta^x$
Figure 4. (top) Optic flows representing expansion (or divergence) in the Cartesian domain, with different shifts of the FRM, and (bottom) corresponding cortical optic flows. Small (±4°) variations of the FRM location produce high nonlinearities in the cortical flows.

\( \{\xi, \eta\} \rightarrow (0, \xi) \) and \( d_n^{\eta}: (\xi, \eta) \rightarrow (0, \eta) \) represent the cardinal deformations (gradients)—i.e., the basis of the linear deformation space—in the cortical domain. Figure 6 shows the cardinal deformations: the stretching and shearing gradients along each directions, and the two translation components.

The sensitivity to such deformations can be modeled as a population of cells whose response is computed through an adaptive template matching on the cortical optic flow. From the responses of such a population we compute the first-order (affine) description (Koenderink, 1986) of the cortical optic flow:

Figure 5. Two deformation subspaces (Chessa et al., 2013), representing an expansion (left) and a rotation (right), obtained from the combination of deformation gradients and translation components.
where $\tilde{c}_i$ are constants and $v_{\xi}$ and $v_{\eta}$ are the components of the cortical optic flow. The parameter vector $[\tilde{c}_1, \tilde{c}_2, ..., \tilde{c}_6]$ describes a specific configuration of cortical optic flow in a local patch.

Motion interpretation (the estimation of the 3-D orientation of the surfaces, of the time to collision, and of the FRM; Xiao, Marcar, & Raiguel, 1997) is performed with respect to the Cartesian coordinates. Indeed, the interaction with the real world occurs in Cartesian coordinates, and thus we need the relationships among the cortical descriptors and the Cartesian ones in the area of interest (for details, see Appendix and Solari et al., 2014) in order to perform motion interpretation by using the cortical first-order descriptors, as it happens in the visual cortex.

Solari et al. (2014) derived the mathematical formulations that allow us to recover time to contact and the surface-orientation information from the cortical optic flow. Here we derive the mathematical formulation that can be employed to estimate the FRM from the cortical optic flow. Following the spiral space model, we compute the location of the FRM by detecting the area where the optic flow can be locally described by a divergence (i.e., expansion) component and a null translational (i.e., laminar motion) component. In particular, considering the cortical first-order description, the FRM is computed as follows. To detect a null translational component we have to locate the retinal region where the Cartesian affine coefficients $c_1$ and $c_4$ (see Appendix, Equation 17) are null. We can reformulate this rule in the cortical domain by using the relationships described in Equation 20, yielding

$$
\tilde{c}_2^2 + \tilde{c}_5^2 < \frac{T_1}{\rho_0 d^2},
$$

where $T_1$ denotes a threshold value. It is worth noting that the right side of Equation 12 is not a constant—i.e., it decreases as a function of $\xi$.

All cortical regions which satisfy Equation 12 will correspond to retinal locations with negligible translational optic flow. Next, we must locate the retinal locations where the optic flow is expanding or contracting and can thus be described as a nonnull divergence—i.e., the Cartesian affine coefficients must obey the rule $|c_2 + c_6| > 0$. Using the relationships described in Equation 20, this rule can be reformulated with affine cortical coefficients as

$$
\tilde{c}_2 + \tilde{c}_6
$$
\[ c_1^2 + c_2^2 + c_3^2 > \frac{T_2}{\ln(a^2) \ln(a) + 1} = T_3, \]  
where \( T_2 \) (\( T_3 \)) denotes a threshold value and \( a > 1 \).

**Model implementation**

A preliminary implementation of the proposed model has been implemented in C++ and is capable of real-time performance (Solari et al., 2014). The C++ code for the retino-cortical mapping module has been made publicly available by the authors on OpenCV2, and the MATLAB code of the motion estimation module is available on ModelDB.3

In the current experiments, we employed a MATLAB implementation of the proposed model, with nine frames fed to the model each trial. Because the model does not process color information in any way, RGB image sequences were fed to the model in grayscale. The retino-cortical mapping module uses as inputs images of 1920×1080 pixels and produces cortical images of 214×150 pixels, thus achieving a compression ratio of CR = 64.6. Then the cortical images are processed by the motion estimation module: The size of the spatiotemporal V1 RFs is 11×11×5—i.e., the spatial support is 11×11 pixels and the temporal support is five frames—near the fovea, the considered spatial orientations are 16, the number of spatial scales is six, and the V1 tuning velocities are nine between −0.9 and 0.9 pixels/frame. Finally, the motion interpretation (MST) module processes the cortical optic flows by using larger RFs (48×48 pixels near the fovea).

---

### FRM estimation as a function of eccentricity

The model described thus far implements known aspects of the first stages of the dorsal cortical pathway for motion processing, and is potentially able to estimate the FRM in image sequences containing motion. We thus tested the model’s performance at FRM estimation and compared its performance to that of human observers. We were interested in whether the proposed model would capture the changes in human performance at motion estimation across the visual field. If the model were to show similar patterns of performance to human observers, we could take that as good evidence that the model implementation is consistent with the neural computations performed by the visual system regarding motion estimation through exploitation of a logic similar to that employed by Freeman and Simoncelli (2011). If the model were unable to perform the motion-estimation task, or if the pattern of performance differed between model and human observers, we could conclude that the implemented aspects of known neural motion processing are insufficient to fully describe how the human visual system extracts motion information from the cortical visual representation of the world.

Broadly, it is well established that motion sensitivity covaries with spatial resolution in the peripheral visual field (McKee & Nakayama, 1984; Wright & Johnston, 1983). More specifically, Bex and Falkenberg (2006) have shown that the accuracy with which human observers can estimate the FRM in random-dot stimuli decreases as a function of retinal eccentricity. We proceeded to replicate the findings from Bex and Falkenberg (2006) in human observers with more naturalistic dead-leaves stimuli (Bordenave et al., 2006; Lee et al., 2001), tested the model on the same task and stimuli, and compared the model’s performance to that of the human observers.

### Experiment 1: FRM estimation in human observers

**Methods**

All methods were approved by the Internal Review Board of Northeastern University and adhered to the tenets of the Declaration of Helsinki.

**Participants**

Two authors (GM and PJB) and one naïve observer (WH) were recruited to participate in Experiment 1. All three subjects were men; they were 27, 49, and 31 years of age, respectively, with normal or corrected-to-normal vision in the test eye. All subjects provided written informed consent.

**Stimuli**

An example of generated stimuli can be seen in Movie 1 (Figure 7). Stimuli were expanding or contracting dead-leaves patterns (Bordenave et al., 2006; Lee et al., 2001) generated using Psychophysics Toolbox Version 3 (Brainard, 1997; Pelli, 1997) running on MATLAB (Mathworks, Natick, MA). Dead leaves were constructed from a set of 2,000 limited-lifetime ellipses, each drawn with a random orientation, aspect ratio, and RGB color, randomly placed within a circular region with diameter equal to half the monitor length. FRM sensitivity is relatively invariant to element lifetime (Warren, Blackwell, Kurtz, Hatsopoulos, & Kalish, 1991); thus, to keep observers from tracking single elements over time, the lifetime of the ellipses was fixed at five movie frames.
The limited lifetime also prevented large density changes that might otherwise occur as ellipses cluster in the center of the image with contracting motion. In the first movie frame the ellipses were assigned a random starting location and a random age between one and five movie frames, which ensured that not all elements would expire simultaneously. The location of each ellipse was updated at every movie frame. The direction of motion dotDir of each ellipse was computed as

$$\text{dotDir} = \text{atan2} \left( \frac{y_{\text{dot}} - y_{\text{FRM}}}{x_{\text{dot}} - x_{\text{FRM}}} \right) + \text{MotionDir},$$

where \(\text{atan2}(y, x)\) is the four-quadrant inverse tangent function, \((x_{\text{dot}}, y_{\text{dot}})\) and \((x_{\text{FRM}}, y_{\text{FRM}})\) are the \((x, y)\) coordinates of the ellipse and FRM, respectively, and \(\text{MotionDir} = 0\) rad for expanding motion and \(\pi\) rad for contracting motion.

The updated \((x, y)\) position of each ellipse was then computed as

$$x_{\text{dot}} = x_{\text{dot}} + \cos(\text{dotDir}) \times \text{dotDist} \times \text{dotSpeed},$$  \(y_{\text{dot}} = y_{\text{dot}} + \sin(\text{dotDir}) \times \text{dotDist} \times \text{dotSpeed},$$

where \(\text{dotDist}\) is the distance of the ellipse from the FRM and \(\text{dotSpeed}\) is the speed of an ellipse 1 pixel away from the FRM, which was set to be 0.1 pixel/frame. These computations generated expanding or contracting motion with a realistic speed gradient (cf. Movie 1, Figure 7). Elements that fell outside the circular stimulus region or whose lifetime exceeded five movie frames were randomly repositioned within the stimulus and assigned an age of zero.

**Apparatus**

FRM stimuli were presented on a Dell P2815Q monitor with a resolution of 1920 x 1080 pixels run at 60 Hz from an AMD Radeon HD 7000 graphics-processing unit. Subjects were positioned 50 cm from the monitor, which subtended 64° x 38° of visual angle. Observers were positioned on a chin and forehead rest, with their nondominant eye occluded, so as to view the stimuli monocularly and reduce motion- and depth-cue conflicts. Fixation compliance was monitored using a Tobii EyeX eye tracker, a low-cost eye tracker previously validated for research purposes (Gibaldi, Vanegas, Bex, & Maiello, 2016).

**Design**

To replicate the findings of Bex and Falkenberg (2006), we measured the accuracy of motion integration by asking human observers to position the mouse cursor at the perceived location of the FRM presented at different locations into the peripheral visual field. Each observer completed 150 trials. In random order, the FRM appeared at 4° in one third of trials, 8° in one third, and 12° in one third. To minimize the buildup of motion aftereffects, the direction of motion (expansion or contraction) was randomly assigned each trial. In postprocessing, we removed a total of 19% trials in which the eye-tracker data signaled that the observers had not maintained steady fixation within 3° of the central fixation target.

**Procedure**

Observers were instructed to fixate a central fixation dot while stimuli were presented to their nasal visual field (in order to avoid the physiological blind spot). The stimulus, a circular dead-leaves patch generated as described earlier, subtended 35° of visual angle and was centered 17.5° to the right of fixation. At the start of each trial, the observers were presented with a fixation dot and were shown their own gaze point (estimated through the use of the eye tracker) to remind them to maintain steady central fixation. When ready, observers initiated a trial by mouse click. Within the circular stimulus, the FRM was randomly selected along 30° arcs at either 4°, 8°, or 12° from central fixation (see Figure 8a; black asterisks indicate FRM test locations). Motion stimuli were presented for 120 frames (2 s), during which time the mouse cursor was not presented. Once the stimulus had been extinguished, the mouse cursor was presented on the screen at a random position on a uniform gray background. Observers
were free to move their eyes, and their task was to move the tip of the cursor to the perceived FRM and press the mouse button.

**Experiment 2: FRM estimation in the proposed model**

To compare the performance of the human observers to the performance of the model, the same expanding and contracting dead-leaves stimuli were fed to the model. The output of the computations performed by the model on each image sequence was the estimated FRM for that sequence. We ran 150 image sequences through the model and recorded its estimated FRM for each sequence.

**Results**

Figure 8 shows the results of Experiments 1 (from the three human observers) and 2 (from the model). Human-observer data were discarded when the eye-tracker data reported incorrect fixations or when the monitor refresh rate fell below 50 Hz, which would occasionally occur due to the high computational costs of generating the dead-leaves stimuli in real time. This resulted in 19% of discarded trials from observers. Data from the model were discarded when the thresholds in Equations 12 and 13 were not reached and the model was thus unable to output a FRM estimate. This resulted in 21% of discarded trials from the model.

Figure 8 presents the error vectors between the actual FRM test locations and the perceived FRM locations in the three human observers and the estimated FRM locations from the model. Qualitatively, we note that both the model and the human observers can perform the task, although with some degree of error. Furthermore, the spread of the data across human observers and model is qualitatively similar. Figure 8c shows the average perceived location of the FRM at each eccentricity for the three observers and the model. All three observers exhibited a consistent fovea-centric bias at all three test eccentricities. Similar biases have been previously observed in FRM-estimation tasks (Johnston, White, & Cumming, 1973; Warren & Saunders, 1995), and especially with Bex and Falkenberg (2006), in spite of differences between the stimuli employed in the different studies. The model’s absolute error in FRM estimation as a function of retinal eccentricity is similar, both in magnitude and trend, to the data from human observers. Given these results, we can conclude that the theoretically based neural computations implemented in the proposed model are consistent with the computations performed by the human visual system regarding complex motion estimation.

**Equivalent-noise analysis of FRM estimation as a function of eccentricity**

Two main factors may affect the accuracy and precision with which humans estimate the FRM in complex, noisy motion stimuli. Loss of resolution in the peripheral visual field may add noise to local direction estimates on which global optic-flow computations are performed. Additionally, RF size differences throughout the visual field may affect the efficiency with which motion information present within a stimulus is integrated. Resolution loss and RF size changes throughout the visual field are built into the proposed model. Thus, we now ask whether these features of the model produce the same effects as they do in human observers.

To tease apart the contributions of internal noise and integration efficiency to the precision of FRM estimation, we employ the equivalent-noise (EN) paradigm (Barlow, 1956; Pelli, 1990), as employed by Bex and Falkenberg (2006). The EN paradigm is based on the assumption that an observer’s performance is limited by additive internal noise as well as by how efficiently the observer samples the information available from the stimulus. Assuming that the variance in the stimulus and variance in the visual system are additive, thresholds for FRM estimation can be expressed as

$$\sigma_{\text{FRM}} = \sqrt{\frac{\sigma_{\text{int}}^2 + \sigma_{\text{ext}}^2}{n}},$$  \hspace{1cm} (16)

where $\sigma_{\text{FRM}}$ is the FRM discrimination threshold, $\sigma_{\text{int}}$ is the internal noise of the system, $\sigma_{\text{ext}}$ is the external noise contained in the stimulus, and $n$ is the sampling efficiency, which relates to how well the system is able to integrate the information contained within the stimulus.

Following the logic employed in the previous set of experiments, we perform EN analysis on both human observers and the model, and compare the results. If the patterns of results in humans and in the model are compatible, this would suggest that the implemented features of the model regarding resolution loss and RF
size are exerting similar effects on the performances of humans and the model.

**Experiment 3: EN analysis in human observers**

**Methods**

All methods were approved by the Internal Review Board of Northeastern University and adhered to the tenets of the Declaration of Helsinki.

**Participants**

Six human observers were recruited to participate in Experiment 3: authors GM and PJB and four naive participants. In total, there were five men and one woman, with a mean (SD) age of 34 ± 9 years. All observers had normal or corrected-to-normal vision in the test eye. All subjects provided written informed consent.

**Stimuli**

Stimuli were the same expanding and contracting dead-leaves patches employed in the previous set of experiments. External noise was added to the stimuli as follows. Instead of selecting a single FRM position for all dead-leaves elements, the \((x, y)\) positions of the FRM of each element were selected from normal distributions with means equal to \((x_{FRM}, y_{FRM})\) and standard deviation equal to \(\sigma_{ext}\)—i.e., the external noise.

**Apparatus**

FRM stimuli were presented on an ASUS VG278HE monitor with resolution of 1920 × 1080 pixels run at 60 Hz from an NVidia Quardo 580 graphics-processing unit. Subjects were positioned 55 cm from the monitor, which subtended 57° × 34° of visual angle. Observers were positioned on a chin and forehead rest, with their nondominant eye occluded, so as to view the stimuli monocularly and reduce motion- and depth-cue conflicts. Observers were instructed to fixate a central fixation dot while the stimuli, circular dead-leaves patches generated as described earlier, subtended 34° of visual angle and were centered at fixation. As previously, fixation compliance was monitored using a Tobii EyeX eye tracker (Gibaldi et al., 2016).

**Design**

The full EN function is typically estimated by measuring observers’ thresholds (in this case, the FRM discrimination threshold \(\sigma_{FRM}\)) at varying amounts of external noise \(\sigma_{ext}\), and the observed thresholds are then fitted to Equation 16, thus obtaining estimates of the internal noise \(\sigma_{int}\) and sampling efficiency \(n\). FRM discrimination thresholds at each tested eccentricity were thus measured at five fixed levels of external noise:
The thresholds were measured via 15 randomly interleaved staircases (Wetherill & Levitt, 1965). The raw data from a minimum of 50 trials from each staircase were combined and fitted with a cumulative normal function by weighted least-squares regression (in which the data are weighted by their binomial standard deviation). FRM discrimination thresholds were estimated from the 80% correct point of the psychometric function. For each tested eccentricity, these thresholds were fitted via nonlinear least-squares regression to the EN function presented in Equation 16.

**Procedure**

The sequence of events from a single trial is shown in Figure 9. At the start of each trial, the observers were presented with a fixation dot; after 250 ms they were also shown a cross at either the fovea or at 4° or 8° eccentricity, to cue where the stimulus would appear in the visual field. The cue was extinguished after 500 ms, and motion stimuli were then presented for nine frames (150 ms). The FRM was located at the fovea or at 4° or 8° eccentricity, shifted into one of four quadrants of the image patch by an amount that was under the control of a three-down, one-up staircase (Wetherill & Levitt, 1965) that adjusted the shift to a level that produced 79% correct trials. The testing procedure’s goal was to estimate the minimum FRM shift necessary for observers to identify in which of the four quadrants the FRM had been presented. Once the stimulus had been extinguished, observers were required to indicate via mouse click in which of the four image quadrants centered at the eccentric testing location they had perceived the FRM. Observers were given unlimited time to respond. The following trial commenced as soon as observers provided a response.

**Figure 9. Schematic of a single trial from Experiment 3.**

Observers were required to fixate a central fixation target (blue). While maintaining steady fixation, observers were shown a green cross at either the fovea or 4° or 8° eccentricity, to cue approximately where the FRM would appear in the visual field. Fixation compliance was ensured with an eye tracker. The stimulus, a full field of expanding or contracting dead leaves, then appeared on screen for 150 ms, an interval that is too brief for a change in fixation. Following the stimulus presentation, observers were required to indicate via mouse click in which of the four image quadrants centered at the eccentric testing location they had perceived the FRM (question marks are for illustration only and were not present in the experiment).

0.25°, 0.5°, 1°, 2°, and 4°. The thresholds were measured via 15 randomly interleaved staircases (Wetherill & Levitt, 1965). The raw data from a minimum of 50 trials from each staircase were combined and fitted with a cumulative normal function by weighted least-squares regression (in which the data are weighted by their binomial standard deviation). FRM discrimination thresholds were estimated from the 80% correct point of the psychometric function. For each tested eccentricity, these thresholds were fitted via nonlinear least-squares regression to the EN function presented in Equation 16.

**Experiment 4: EN analysis in the proposed model**

As in the previous set of experiments, in order to compare the performance of the human observers to the performance of the model, the same expanding and contracting dead-leaves stimuli were fed to the model. The stimuli, testing procedure, and data processing used on the model were the same as those used with human observers in Experiment 3. Instead of a mouse click, the model output was an estimate of the FRM location within its visual field. A single trial from the model was scored as correct if the estimate fell within the correct stimulus quadrant, and incorrect otherwise. If for a single trial the model was unable to return an estimate of the FRM, the trial was repeated with a new image sequence.

EN functions at each eccentricity were estimated from seven separate model-parameter configurations. In each of the seven experimental runs, the model parameters were pseudorandomly chosen from within a sensible range. The parameter configurations employed in each run are reported in Table 1.

**Results**

Figure 10 summarizes the results of the EN analysis performed in six human observers. Figure 10a shows FRM threshold offsets as a function of external noise for all six observers. Red, green, and blue curves passing through the data are the average fitted EN curves at each eccentricity. The average estimated EN parameters are plotted as a function of eccentricity in Figures 10b (internal noise) and 10c (sampling efficiency). Error bars are 95% bootstrapped confidence intervals of the mean parameter estimates.

Figure 11 presents the results of the EN analysis performed on seven separate model-parameter config-
Performances overlaid onto the observed range of human performance. Figure 11a shows FRM threshold offsets for all seven model-parameter configurations and fitted EN curves as a function of external noise, averaged across the seven model-parameter configurations at each eccentricity. The average estimated EN parameters for the model are plotted as a function of eccentricity in Figures 11b (internal noise) and 11c (sampling efficiency). Average EN parameter estimates bounded by 95% bootstrapped confidence intervals are overlaid onto gray-shaded 95% confidence regions for the human EN parameter estimates.

In both human observers (Figure 10a) and the model (Figure 11a), thresholds vary lawfully as a function of external noise. FRM location-identification performance is better at low levels of external noise than at high. Furthermore, performance worsens with increasing eccentricity. The model’s performance is better than that of human observers for low levels of external noise, but worsens at a faster rate with increasing levels of external noise.

To statistically evaluate how internal noise and sampling efficiency vary throughout the visual field of human and model observers, EN parameter estimates were analyzed via a 2 (observer type) × 3 (eccentricity) ANOVA.

The ANOVA result for internal noise showed a significant main effect of observer type, $F(1, 33) = 61.12$, $p = 10^{-8}$, no significant main effect of eccentricity, $F(2, 33) = 1.34$, $p = 0.28$, and no significant interaction between observer type and eccentricity, $F(2, 33) = 0.39$, $p = 0.68$. As can be seen in Figures 10b and 11b, there is a trend for internal noise in both human and model observers to increase with eccentricity, but the increase is not statistically significant. The model has significantly less overall internal noise than human observers.

The ANOVA result for sampling efficiency showed a significant main effect of observer type, $F(1, 33) = 20.45$, $p = 10^{-4}$, a significant main effect of eccentricity, $F(2, 33) = 17.94$, $p = 10^{-5}$, and a significant interaction between observer type and eccentricity, $F(2, 33) = 4.39$, $p = 0.02$. Sampling efficiency in both human (Figures 10c) and model (Figures 11c) observers decreases with eccentricity. However, the model has significantly lower overall sampling efficiency than human observers. Furthermore, sampling efficiency decreases at a faster

### Table 1. Model parameters for Experiment 4.

<table>
<thead>
<tr>
<th>Run</th>
<th>$R$</th>
<th>$R_0$</th>
<th>$\chi$</th>
<th>CR</th>
<th>$W_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>130</td>
<td>9</td>
<td>33</td>
<td>35</td>
<td>14</td>
</tr>
<tr>
<td>II</td>
<td>160</td>
<td>7</td>
<td>40</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td>III</td>
<td>110</td>
<td>10</td>
<td>28</td>
<td>47</td>
<td>16</td>
</tr>
<tr>
<td>IV</td>
<td>110</td>
<td>3</td>
<td>40</td>
<td>63</td>
<td>21</td>
</tr>
<tr>
<td>V</td>
<td>160</td>
<td>3</td>
<td>47</td>
<td>29</td>
<td>15</td>
</tr>
<tr>
<td>VI</td>
<td>130</td>
<td>3</td>
<td>43</td>
<td>45</td>
<td>18</td>
</tr>
<tr>
<td>VII</td>
<td>105</td>
<td>9</td>
<td>27</td>
<td>54</td>
<td>18</td>
</tr>
</tbody>
</table>

Notes: For the seven experimental runs (I–VII) we varied the number of rings $R$ and the blind-spot radius $R_0$, thus obtaining the corresponding percentage of cortex used to overrepresent the fovea $\chi$ (see Equation 5), the compression ratio CR (see Equation 3), and the maximum RF size $W_{\text{max}}$ (see Equation 4).

![Figure 10](http://arvojournals.org/) EN analysis in human observers. (a) EN functions at the fovea (red) and 4° (green) and 8° (blue) eccentricity. The data show FRM discrimination thresholds as a function of positional noise applied to each element within the stimulus. Individual data points are discrimination thresholds for all six observers. Curves are best-fitting EN functions to the averaged data across the six observers. Note that the x-axis is log scaled. (b) Internal noise and (c) sampling efficiency parameters of the estimated EN functions plotted as a function of eccentricity. Data are the average across six observers. Error bars represent 95% bootstrapped confidence intervals.
rate in the periphery of the model than in the periphery of human observers.

We further evaluated the influence of the model’s parameters on the model’s performance on the EN task. Stepwise linear regressions were calculated to assess the relationship between the model’s parameters (Table 1, predictors) and the internal noise and sampling efficiency (dependent variables), averaged across eccentricities for each parameter configuration. The stepwise regression found no significant relationship between any of the model’s parameters and the model’s estimated internal noise (no predictors were added to the constant regression model). A significant regression equation was found instead between the percentage of the model’s cortical area devoted to the fovea ($v$) and the model’s estimated sampling efficiency, $F(1, 5) = 27.6, p = 0.003, R^2 = 0.816$. As can be seen in Figure 12, the greater the portion of the model’s cortex dedicated to the fovea, the worse the sampling efficiency. This may be due to the overrepresentation of the fovea in the model’s cortex—i.e., using too many cortical processing units to represent the fovea leaves too few cortical units for the periphery. The decrease in sampling efficiency in the periphery of both model and human observers is thus sensibly attributable to RF size changes across the visual field, which are in turn driven by the lossy log-polar mapping that projects the retinal image onto primary visual areas.

In Experiments 1 and 2 we found that human observers show a systematic fovea-centric bias in FRM estimation, whereas the model does not exhibit such a bias. Conventional wisdom would associate the bias with log-polar oversampling toward the fovea; the lack of fovea-centric bias in the model was thus perplexing.

Figure 11. EN analysis on the model responses. All explanations are as in Figure 10, except for using seven separate model-parameter configurations rather than six human observers. (a) Gray shaded region represents the region of observed human performance. (b, c) Gray shaded region represent the 95% bootstrapped confidence regions from the human-observer data.

Figure 12. Sampling efficiency depends on the portion of the model’s cortical area $\gamma$ dedicated to the model’s fovea (see Table 1). Red data points are estimated sampling efficiency, averaged across eccentricities for each model-parameter configuration, plotted against the percentage of the cortical area representing the model’s fovea. Black line is best-fitting linear regression line, bounded by 95% confidence intervals of the fit (green dotted lines). Gray shaded region represents 95% confidence bounds of the estimated sampling efficiency in human observers, averaged across eccentricities for each observer.
However, having observed that the model has significantly less internal noise than human subjects, we asked whether fovea-centric bias arises from log-polar sampling in noise. We employed the model’s FRM estimates in Experiment 4 to compute its fovea-centric bias as a function of external noise. Figure 13 confirms that at low levels of noise the model exhibits little or no fovea-centric bias. However, we find that the model’s fovea-centric bias increases monotonically as a function of external noise. This suggests that the fovea-centric bias observed in our study and in others (Johnston et al., 1973; Warren & Saunders, 1995) is explained by log-polar oversampling toward the fovea in conditions of motion-detector noise.

We acquired two short (10 s) movie clips using the front camera of a Samsung S3 mobile phone. The camera was set to record video sequences at 30 frames/s with a resolution of 1920 × 1080 pixels, and thus each movie clip contained a total of 300 frames. During each movie clip’s acquisition, the experimenter manually moved the phone toward a frontal obstacle. The experimenter induced a forward as well as lateral translation in the motion of the camera. The global effect was that the direction of heading was shifted toward the left side of the image plane. The movement was produced manually in order to record video sequences containing biological motion. This introduced small vertical movements and oscillations, which have purposely not been compensated for and are visible in the original video sequence (cf. Movie 2).

The image sequences were fed to the proposed model to verify whether the model would be able to process truly natural image sequences and output sensible direction-of-heading information. The results are shown in Figure 14. The model was able to provide direction-of-heading estimates for both recorded movie clips. In the bottom left of Figure 14, red crosses represent the FRM estimates outputted by the model as the video sequences progressed, the red open circle represents the mean of the estimated FRM values, and the filled green circle represents the ground-truth FRM, computed from the geometry of the setup. These data are overlaid onto a sample movie frame. The cortical representation of the same movie frame is shown on the bottom right of Figure 14. The upper inset of the figure presents the intermediate outputs of the model to give an example of how the cortical optic flow and the affine coefficients in the cortical domain might appear when computed from real-world scenes. In particular, the cortical optic flow shows a pattern similar to the one of Figure 4, bottom left: The motion is toward the upper right on the top part and the bottom right on the bottom part, thus indicating an FRM shifted toward the left side of the image. The cortical affine description is obtained by checking the presence of the local cardinal deformations of Figure 6 on the entire cortical image. The cortical affine description shows large translational components (i.e., \( \tilde{c}_1 \) toward the right and \( \tilde{c}_4 \)) and minor deformation components (i.e., \( \tilde{c}_2, \tilde{c}_3, \tilde{c}_5, \) and \( \tilde{c}_6 \)) that make the cortical flow rotate, as happens for a diverging flow with shifted FRM (see Figure 4).

**Discussion**

We present a model of the human visual system that mimics the first processing stages of the dorsal cortical pathway for motion processing. The model was first developed by Solari et al. (2014), where it was validated.
with regards to estimation of time to contact in real movies of approaching objects. The main feature of the model is that it functions on log-polar-transformed image sequences, which can be thought of as cortical image sequences. Thus the model is able to characterize motion interpretation directly in the cortical domain. If the model is a good approximation of the human visual system, it will give us insight into how the human visual system elaborates the distorted representation of the visual world that reaches cortical visual areas.

Few studies in the literature on cortical motion processing have addressed the log-polar transform. Grossberg et al. (1999) provide both a functional explanation and quantitative simulations of experimental data of MSTd cells. To accomplish this, they analyzed MST responses to inputs of optic-flow patterns transformed into log-polar coordinates. Differently from our approach, Grossberg et al. do not mimic the entire visual pathway starting from images as inputs to the model. Thus, our work extends theirs by describing how cortical optic-flow maps can be obtained from the first stages of cortical processing, and then accounts for how this transformed cortical optic flow may be processed by areas MT and MSTd to compute FRM estimates. Furthermore, while Grossberg et al. indeed address the log-polar transform, the cortical mapping they considered is different from the one presented here. This did not allow them to verify the effects of the different parameters of the mapping to reach a cortical representation with performance similar to that of humans, as we have done in this article. Even though our model cannot be directly compared to that of Grossberg et al., since the two models require different inputs, it is interesting that both provide similar insights, such as biases toward fixation for limited depth (see our findings regarding the fovea-centric bias).

Another study in the literature that could be compared with our approach is the ViSTARS neural model proposed by Browning et al. (2009). Both the ViSTARS model and ours similarly account for the performance of a nonfoveated human observer. Browning et al. report errors of $1^\circ$–$3^\circ$ in the estimation of the FRM with several kinds of visual stimuli. This

Figure 14. FRM estimation in real-world sequences. (left) Sample frame (Movie 2) from one of the recorded sequences in the Cartesian domain, with overlapping FRM estimates for several consecutive frames (red crosses), mean of the estimates (open red circle), and ground-truth FRM (filled green circle). (right) The same frame mapped into the cortical domain, enlarged for a better visualization. The V1-MT-MST model computes FRM estimates directly from the cortical image sequence. Intermediate outputs of the model (cortical optic flow and cortical affine coefficients) are shown in the upper inset. The color maps used to show the optic flow and the affine coefficients are shown to the left of the corresponding map.
result is comparable to the error of ~1° that we observe near the fovea of both our model and human observers. The main difference between the ViSTARS model and ours is that the ViSTARS model does not consider the log-polar mapping. Thus, the ViSTARS model cannot describe how optic-flow computations and FRM estimates vary throughout the visual field. Conversely, Browning et al. investigate the effect of eye rotations on optic-flow computations, which we do not.

One important aspect of our investigation is that we aim to provide a direct comparison of our model’s output to those of human observers with respect to performance at FRM estimation. We employ the same set of naturalistic stimuli and tasks for both human observers and the model. In this way, we are able to test whether the model’s behavior is analogous to that of human subjects across the visual field. Whereas several works in the literature (e.g., Browning et al., 2009) compare their results with neurophysiological and behavioral data from the literature, none directly test their models on the same experimental procedures applied to human observers, nor do they test how varying the model’s tuning parameters affects its performance.

In the present work we describe how to mathematically locate the FRM from the computed cortical optic flow. We then proceed to compare the model’s performance on FRM estimation to that of human observers. Both the model and human observers are able to reliably estimate the FRM in expanding and contracting naturalistic dead-leaves stimuli (Bordenave et al., 2006; Lee et al., 2001). The precision with which human observers and the model can estimate the FRM location worsens in the peripheral visual field. The magnitude and trend of observed results from the model are shown to be consistent with the pattern of results observed in human observers. The theoretically based neural computations implemented in the proposed model are thus a good candidate for the computations actually performed by the human visual system regarding complex motion estimation. The main factors contributing to the changes in the precision with which humans estimate the FRM throughout the visual field are resolution loss and RF size differences. Resolution loss and RF size changes throughout the visual field are also two of the main features of the implemented model. By employing EN analysis on both human observers and the model, we show that increases in RF size across the visual field lead to decreases in sampling efficiency in the peripheral visual field of both human observers and the model. RF size is in turn modulated by retino-cortical magnification scaling. Thus, the retino-cortical transformation and the hierarchical architecture implemented in the model produce the same general effects as their biological counterparts in human observers.

However, we did find some differences between the performance of the human observers and the model. The model’s performance was better than that of human observers for low levels of external noise, and worsened at a faster rate with increasing levels of external noise. Indeed, the model had less internal noise and lower sampling efficiency than human observers. The fact that the model had lower internal noise than human observers is sensible: The model does not incorporate the various sources of internal noise that human observers are subjected to, such as the baseline level of random firing of cells, fixational eye movements, and fluctuations in level of attention. Conversely, the fact that the model has lower sampling efficiency than human observers might reflect the fact that its internal parameters were heuristically selected by the experimenters from what was believed to be a sensible range. It might be interesting in future work to develop ways of training the model with natural stimuli and verifying whether model parameters finely tuned to the spatiotemporal statistics of natural scenes provide patterns of performance more similar to those observed in human subjects. Future work should also focus on comparative simulations of the different models presented in the literature. Implementing the available models to incorporate space variance across the visual field and tuning these models to perform as closely to human observers as possible would allow researchers to determine the most parsimonious model that best accounts for human performance. This would provide further insight into the key processing stages underlying biological motion perception.

The results of our EN study differ from those reported by Bex and Falkenberg (2006). Human-observer performance was overall higher in that study, the FRM thresholds were smaller by a factor of 4, internal noise was lower by a factor of 1.5, and sampling efficiency was higher by a factor of 7. Bex and Falkenberg found that internal noise significantly increased with eccentricity, whereas we find a trend in that direction that is not statistically significant. Furthermore, we find a significant decrease in sampling efficiency in the peripheral visual field, whereas their finding of the same trend failed to reach significance. Unfortunately, individual estimates of internal noise and sampling efficiency are highly variable in both studies and are poorly constrained by the data. It is also possible that the differences between results may be driven by the different stimuli employed in the two studies. Whereas Bex and Falkenberg employed localized patches of random-dot stimuli placed at various positions in the visual field, in the current study we employed full-field naturalistic dead-leaves stimuli (Bordenave et al., 2006; Lee et al., 2001). These stimuli have the same 1/f spatial-frequency spectrum and contrast range of natural images, and are textured with
occlusions and edges at a variety of orientations. Hence, the dead leaves are better suited to test the performance of the visual system under natural viewing conditions, since they better approximate the natural stimulus range in which the visual system operates.

The localized patches employed by Bex and Falkenberg might also have been ill suited to highlight the fovea-centric bias at FRM estimation found in this and other studies (Johnston et al., 1973; Warren & Saunders, 1995). Our finding that the model’s fovea-centric bias increases with external noise suggests that this bias sensibly arises from log-polar sampling in noise.

With regard to computer vision applications, mimicking the log-polar mapping adopted by foveated mammalian visual systems might provide interesting advantages. Log-polar mapping provides a wide field of view while maintaining high spatial resolution on the region of interest and thus providing significant data reduction. This could be a desirable feature in robotic active vision systems, where the fixation point of the cameras may continuously change and a mechanism to obtain robust features by working on small images (since the cortical mapping also produces a consistent compression ratio, see Equation 3) would be useful in obtaining real-time implementations (Berton, Sandini, & Metta, 2006). In the current article (as well as Solari et al., 2014), a general approach to extracting visual features directly into the cortical domain is developed. We show how to directly use in the cortical domain standard computer vision algorithms that work in the Cartesian domain. Solari et al. (2014) proved the model to reliably estimate time to contact in automotive real-world scenes. Here we further show that direction of heading can be reliably estimated in real-world sequences acquired with a cell-phone camera. The presented model is thus a good candidate for robotics applications such as the humanoid robot iCub (Metta, Sandini, Vernon, Natale, & Nori, 2008) and could be employed to further study the cognitive processes underlying biological motion perception.

Keywords: focus-of-radial-motion estimation, space-variant processing, dead-leaves stimuli, equivalent-noise analysis, V1-MT-MST neural model

Acknowledgments

Commercial relationships: none.
Guido Maiello.
Email: guido.maiello.13@ucl.ac.uk.
Address: UCL Institute of Ophthalmology, University College of London, London, UK.

Footnotes

1 Using thresholds, rather than searching for null translational and nonnull divergent flow components, allows us to handle the variability present in the visual stimulus in real-world situations. In the simulations carried out in this article we have chosen $T_1 = 0.8$ and $T_2 = 0.6$.

2 The package is available on OpenCV version 2.4.X (opencv.org/downloads.html). Once the package is downloaded, the source code is available in the [contrib] folder.

3 http://senselab.med.yale.edu/ModelDB/ShowModel.cshtml?model=181035.

References


Appendix

Relationships between the log-polar and Cartesian affine description

By considering an affine local description of the Cartesian optic flow

\[
\begin{bmatrix}
  v_x \\
  v_y
\end{bmatrix} = \begin{bmatrix}
  c_1 \\
  c_4
\end{bmatrix} + \begin{bmatrix}
  c_2 & c_3 \\
  c_5 & c_6
\end{bmatrix} \begin{bmatrix}
  x \\
  y
\end{bmatrix},
\]  
(17)

the inverse of the curvilinear-coordinate transformation (Chan Man Fong, Kee, & Kaloni, 1997)

\[
\begin{bmatrix}
  v_x \\
  v_y
\end{bmatrix} = \frac{1}{\rho_0 a^2 \ln(a)} \begin{bmatrix}
  \cos \eta & \sin \eta \\
  -\sin \eta & \cos \eta
\end{bmatrix} \begin{bmatrix}
  v_x \\
  v_y
\end{bmatrix},
\]  
(18)

and the inverse of the log-polar transformation (Equation 1), we obtain the cortical representation of an affine Cartesian optic flow:

\[
\begin{bmatrix}
  v_x \\
  v_y
\end{bmatrix} = \frac{1}{\ln(a)} \begin{bmatrix}
  a^{-\zeta} (c_1 \cos \eta + c_4 \sin \eta) + (c_5 + c_6) \frac{\sin 2\eta}{2} + c_3 \cos^2 \eta + c_6 \sin^2 \eta \\
  a^{-\zeta} \frac{\rho_0}{\ln(a)} (c_1 \sin \eta + c_4 \cos \eta) + (c_5 - c_6) \frac{\sin 2\eta}{2} + c_3 \cos^2 \eta - c_6 \sin^2 \eta
\end{bmatrix},
\]  
(19)
To consider the first-order description of \((v_n, v_g)\), we compute the Taylor expansion of Equation 19 at \((\xi_0, \eta_0)\).

The constant terms \(\tilde{c}_1\) and \(\tilde{c}_4\) are the values of the cortical flow (Equation 19) computed in \((\xi_0, \eta_0)\), and the terms \(\tilde{c}_2, \tilde{c}_3, \tilde{c}_5,\) and \(\tilde{c}_6\) are the partial derivatives:

\[
\begin{align*}
\tilde{c}_1 &= v_\xi(\xi_0, \eta_0) \\
\tilde{c}_2 &= \frac{\partial v_\xi}{\partial \xi} = -\frac{a^{-\xi}}{\rho_0}(c_1 \cos \eta + c_4 \sin \eta) \\
\tilde{c}_3 &= \frac{\partial v_\xi}{\partial \eta} = -\frac{a^{-\xi}}{\rho_0 \ln(a)}(c_1 \sin \eta - c_4 \cos \eta) \\
&\quad + \frac{1}{\ln(a)}((c_6 - c_2)\sin 2\eta + (c_3 + c_5)\cos 2\eta) \\
\tilde{c}_4 &= v_\eta(\xi_0, \eta_0) \\
\tilde{c}_5 &= \frac{\partial v_\eta}{\partial \xi} = \frac{a^{-\xi}}{\rho_0}(c_1 \sin \eta - c_4 \cos \eta) \\
\tilde{c}_6 &= \frac{\partial v_\eta}{\partial \eta} = -\frac{a^{-\xi}}{\rho_0 \ln(a)}(c_1 \cos \eta + c_4 \sin \eta) \\
&\quad + \frac{1}{\ln(a)}((c_6 - c_2)\cos 2\eta - (c_3 + c_5)\sin 2\eta).
\end{align*}
\]

In this way, we obtain six relationships that locally relate the affine coefficients \([\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_6]\) computed on the cortical optic flow to the affine coefficients of the corresponding Cartesian optic flow \([c_1, c_2, \ldots, c_6]\) (Solari et al., 2014). We can solve the resulting system of equations, thus obtaining

\[
\begin{align*}
c_1 &= \rho_0 a^\xi(-\tilde{c}_2 \cos \eta + \tilde{c}_5 \sin \eta) \\
c_2 &= \frac{1}{2}(2\tilde{c}_1 \ln(a) + \tilde{c}_2(1 + \cos 2\eta) - \tilde{c}_3 \ln(a) \sin 2\eta) \\
&\quad + -\tilde{c}_5 \sin 2\eta + \tilde{c}_6 \ln(a)(1 - \cos 2\eta)) \\
c_3 &= \frac{1}{2}(\tilde{c}_2 \sin 2\eta + \tilde{c}_3 \ln(a)(1 + \cos 2\eta) - 2\tilde{c}_4 \ln(a) \\
&\quad - \tilde{c}_5(1 - \cos 2\eta) - \tilde{c}_6 \ln(a) \sin 2\eta) \\
c_4 &= \rho_0 a^\xi(\tilde{c}_2 \sin \eta + \tilde{c}_5 \cos \eta) \\
c_5 &= \frac{1}{2}(\tilde{c}_2 \sin 2\eta - \tilde{c}_3 \ln(a)(1 - \cos 2\eta) + 2\tilde{c}_4 \ln(a) \\
&\quad + \tilde{c}_5(1 + \cos 2\eta) - \tilde{c}_6 \sin 2\eta) \\
c_6 &= \frac{1}{2}(2\tilde{c}_1 \ln(a) + \tilde{c}_2(1 - \cos 2\eta) + \tilde{c}_3 \ln(a) \sin 2\eta \\
&\quad + \tilde{c}_5 \sin 2\eta + \tilde{c}_6 \ln(a)(1 + \cos 2\eta)).
\end{align*}
\]