Biol Cybern (2017) 111:185–206
DOI 10.1007/s00422-017-0715-0

ORIGINAL ARTICLE

Biological Cybernetics

An insect-inspired model for visual binding I: learning objects and their characteristics

Brandon D. Northcutt[1] · **Jonathan P. Dyhr**[2] · **Charles M. Higgins**[3]

Received: 2 April 2016 / Accepted: 27 February 2017 / Published online: 16 March 2017

Abstract Visual binding is the process of associating the responses of visual interneurons in different visual submodalities all of which are responding to the same object in the visual field. Recently identified neuropils in the insect brain termed *optic glomeruli* reside just downstream of the optic lobes and have an internal organization that could support visual binding. Working from anatomical similarities between optic and olfactory glomeruli, we have developed a model of visual binding based on common temporal fluctuations among signals of independent visual submodalities. Here we describe and demonstrate a neural network model capable both of refining selectivity of visual information in a given visual submodality, and of associating visual signals produced by different objects in the visual field by developing inhibitory neural synaptic weights representing the visual scene. We also show that this model is consistent with initial physiological data from optic glomeruli. Further, we discuss how this neural network model may be implemented in optic glomeruli at a neuronal level.

✉ Brandon D. Northcutt
brandon@northcutt.net

Jonathan P. Dyhr
jonathan.dyhr@northwestu.edu

Charles M. Higgins
higgins@neurobio.arizona.edu

[1] Department of Electrical and Computer Engineering, University of Arizona, 1230 E. Speedway Blvd., Tucson, AZ 85721, USA

[2] Department of Biology, Northwest University, 5520 108th Ave. N.E., Kirkland, WA 98033, USA

[3] Departments of Neuroscience and Electrical/Computer Engineering, University of Arizona, 1040 E. 4th St., Tucson, AZ 85721, USA

Keywords Vision · Neural networks · Biomimetic · Visual binding · Neuromorphic · Image understanding

1 Introduction

Visual binding refers to the process of grouping neuronal responses produced by one object while differentiating them from responses produced by others (von der Malsburg 1999). This process has been long studied and modeled with reference to vertebrate brains, but it is currently unknown whether insects make use of visual binding, and if so what neuronal mechanisms may be used. The presence of recently identified structures termed optic glomeruli (Strausfeld and Okamura 2007) in the insect brain suggest one method by which rudimentary visual binding may be performed. These structures have been identified in the lateral protocerebra of both flies (Strausfeld and Okamura 2007) and bees (Paulk et al. 2009), and it is probable that they are present in many other insect species. Optic glomeruli receive retinotopic input from the visual system, and these signals are likely to consist of visual "submodalities," including motion, orientation and color (Okamura and Strausfeld 2007; Mu et al. 2012). Output of the optic glomeruli are far fewer than their inputs, and this reduction suggests optic glomeruli are involved in processing visual information into higher-level representations—possibly coding for features and/or objects (Okamura and Strausfeld 2007; Strausfeld and Okamura 2007; Strausfeld et al. 2007; Mu et al. 2012).

Detailed anatomical studies of optic glomeruli have been carried out (Strausfeld and Okamura 2007) but their physiology is still under active investigation, and only a very limited set of experiments have been conducted (Okamura and Strausfeld 2007; Mu et al. 2012). Initial electrophysiological experiments have shown optic glomeruli to receive

broadly orientation-tuned inputs from the optic lobes, and that neurons projecting from optic glomeruli have a narrower orientation tuning, presumably due to computations within the glomeruli. Perhaps the best route to model these structures is to leverage their anatomical similarity to antennal lobe olfactory glomeruli (Strausfeld and Okamura 2007), which are well mapped and modeled in flies (Jefferis 2005), honeybees (Linster and Smith 1997), locusts (Bazhenov et al. 2001), and moths (Hildebrand 1996).

In the antennal lobes, all olfactory receptor neurons expressing a given receptor type converge to the same glomerulus (Jefferis 2005). Each glomerulus serves to process incoming information from olfactory receptor neurons using local inhibitory interneurons, and to provide processed information via projection neurons to higher-level neural circuits in the mushroom bodies and the lateral protocerebrum (Ng et al. 2002). Local interneurons are thought to get synaptic input from only one glomerulus (Fonta et al. 1993). In models of the antennal lobe (Linster and Masson 1996; Bazhenov et al. 2001), olfactory receptor neurons excite both local interneurons and projection neurons, and local interneurons inhibit other local interneurons, projection neurons, and the receptor neurons themselves.

Given the apparent anatomical homology between optic and olfactory glomeruli, what would be the most likely correspondence of elements between their neuronal circuits? Columnar neurons observed projecting from the lobula complex to optic glomeruli would undoubtedly take the place of olfactory receptor neurons. Recent studies (Okamura and Strausfeld 2007; Mu et al. 2012) have described neurons which might well be morphologically homologous to antennal lobe local inhibitory interneurons. Projections from optic glomeruli to higher brain areas likely correspond to olfactory projection neurons. It is reasonable to assume that similar interconnections may exist between these populations of neurons to those known for the antennal lobe.

What visual inputs might optic glomeruli receive? A number of visual submodalities are available from the lobula complex, including coarsely retinotopic motion, orientation, and likely color information. However, there are only 27 optic glomeruli in the large blowfly *Calliphora*, many fewer than the number of retinotopic visual sampling points, even when compared to the eye of the tiny fruit fly *Drosophila* that has only 900 ommatidia. Perhaps in rough correspondence to the number of optic glomeruli, there are 23 types of columnar neurons projecting from the lobula complex to the optic glomeruli (Okamura and Strausfeld 2007). From this information, we conclude that visual information is spatially integrated before processing by optic glomeruli.

The functional significance of insect antennal lobe olfactory glomeruli is still a subject of debate. These structures may provide a degree of concentration invariance, provide a spatial code for complex odor mixtures, and perhaps even synchronize firing of projection neurons to make a temporal code (Heisenberg 2003). Models of the antennal lobe have demonstrated short-term memory (Linster and Masson 1996), synchronization of output neurons (Bazhenov et al. 2001), overshadowing, blocking, and unblocking (Linster and Smith 1997). Strong similarities exist between insect antennal lobe olfactory glomeruli and the vertebrate olfactory bulb, the most crucial being that in both structures like-typed olfactory receptor signals converge into glomerular regions (Hildebrand 1996). In fact, a number of existing models may apply to both vertebrate and insect systems. The common theme behind all of these possible functions seems to be that olfactory glomeruli encode the *identity* of the odor, but abstract away the details such as spatial concentration and the detailed time course of receptor responses.

It has been hypothesized (Hopfield 1991) that the olfactory bulb may be solving the olfactory binding problem; that is, the olfactory bulb may be able to use information about the fluctuation of individual receptor responses to bind together those responses that encode a single scent. Hopfield proposed a recurrent neural network for modeling vertebrate olfactory glomeruli. Olfactory glomeruli are presumed to group similar chemical features together into an "odor space" where unique odors, composed of chemical mixtures having unique structures, are identified based on their unique patterns of glomerular activation (Hildebrand and Shepherd 1997). Hopfield's model utilized a Hebbian-style learning rule to separate time-varying components of unknown scent mixtures, thus solving an olfactory version of the well-studied *blind source separation* problem, in which the goal is to separate out the contributions of individual "sources" only given an unknown ("blind") linear mixture of those sources. Blind source separation is an area well addressed in the neural network literature (Herault and Jutten 1986; Cichocki et al. 1997) and is discussed in detail in our companion paper (Northcutt and Higgins 2016).

If optic glomeruli are homologous to olfactory glomeruli, what might their function be, translated into visual terms? If they encode the identity of what is seen, abstracting away the details—in particular, the spatial location of visual features—they might be encoding for visual features corresponding to a given object without regard to where it is in the visual field, and thus addressing the visual binding problem.

We have developed a model of optic glomeruli which extends the work of Hopfield (1991) and Herault and Jutten (1986), thus relating optic glomeruli to previous work on olfactory processing and blind source separation. This model, described below, uses first-stage recurrent inhibitory neural networks to model the sensory refinement observed in fly optic glomeruli (Okamura and Strausfeld 2007) by sharpening the selectivity of very broadly tuned inputs. We demonstrate below how this sensory refinement network can be used to improve visual information coding

in the orientation, color and motion visual submodalities. The outputs of these first-stage networks are then provided to a second-stage recurrent inhibitory neural network layer to demonstrate rudimentary visual binding. Each of these recurrent inhibitory networks may correspond to an optic glomerulus.

2 The visual binding network

Since visual information is almost certainly spatially integrated before projecting to optic glomeruli, but the exact pattern of this integration is unknown, in our initial model of this neuronal circuit we spatially integrated all visual submodalities across the entire visual field. This leads to an initial model with far less "glomeruli" than observed in the fly brain, but which (as will be shortly shown) has properties that make it worthy of deeper investigation.

The input to our model consists of a two-dimensional Cartesian spatial array of visual sampling points, each of which has red, green, and blue (RGB) photoreceptors. Although a strict model of insect compound eye color vision would be based on a hexagonal array of green, blue, and ultraviolet (UV) photoreceptors (Snyder 1979), we use a standard RGB image for simplicity of human visualization and computer representation, and without loss of generality, since neither the spatial sampling pattern nor the particular spectral content of the input image are integral to the model.

As diagrammed in Fig. 1, this spatial array of photoreceptors is processed to produce local measures of three visual submodalities: motion, orientation, and color. This processing results in two-dimensional (2D) "feature images" indicating local image motion in four cardinal directions, orientation at three different angles, and each of the three colors. Details of each of these computations are given below.

Each of these 10 local 2D "feature images" was then spatially summed and group-normalized so that different submodalities were comparable in magnitude, resulting in 10 wide-field scalar signals which became input to the model. We refer to these inputs analytically as

$$\boldsymbol{i}(t) = \left[i_1(t)\ i_2(t)\ \ldots\ i_{10}(t)\right]^{\mathsf{T}} \tag{1}$$

This 10-element column vector represents motion, orientation, and color across the entire visual scene without regard to spatial position, and was provided as input to the three first-stage networks of Fig. 2, which refined the selectivity of visual information in each submodality. For future reference, it will be convenient to define subsets of these inputs

$$\boldsymbol{i}_{\mathrm{M}}(t) = \left[i_1(t)\ i_2(t)\ i_3(t)\ i_4(t)\right]^{\mathsf{T}} \tag{2}$$

$$\boldsymbol{i}_{\mathrm{O}}(t) = \left[i_5(t)\ i_6(t)\ i_7(t)\right]^{\mathsf{T}} \tag{3}$$

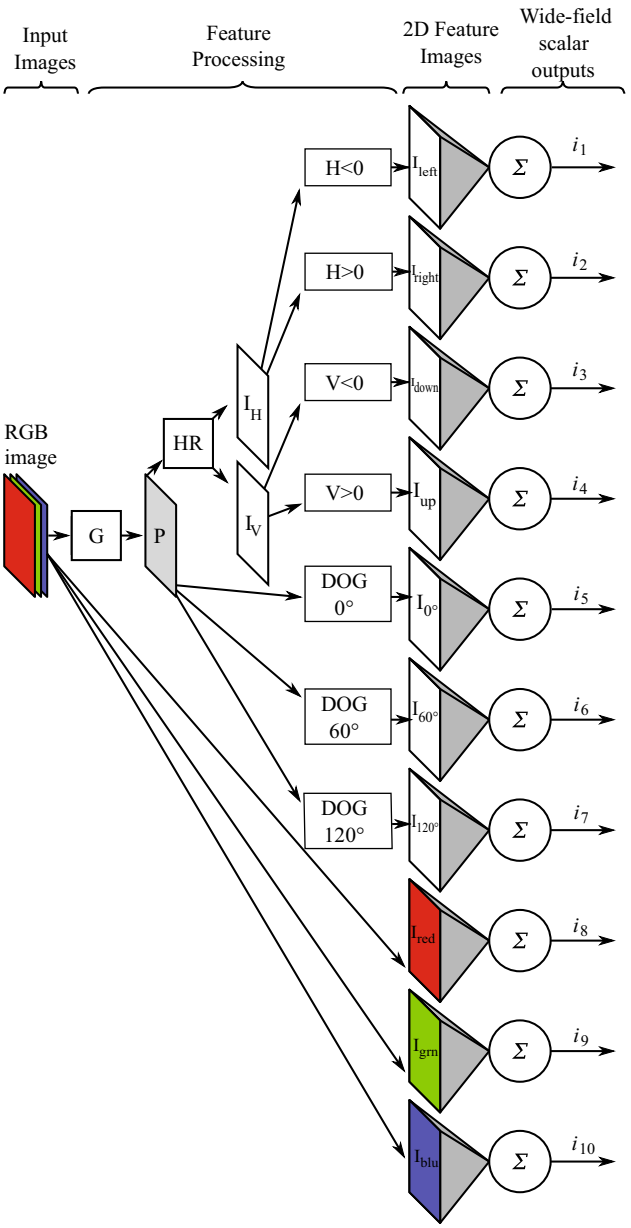

Fig. 1 Computational diagram of visual inputs to binding model. See text for details. The input RGB image was converted to grayscale (G) for motion and orientation processing. Local image motion was computed using the Hassenstein–Reichardt model (HR) in both the *horizontal* ($\boldsymbol{I}_H$) and *vertical* ($\boldsymbol{I}_V$) *directions*, and then separated into four strictly positive 2D "feature images" indicating *upward*, *downward*, *leftward*, and *rightward* motion. Similarly, 2D orientation feature images were computed from the grayscale image using three difference-of-Gaussian filters oriented at 0°, 60° and 120°. Finally, each of the *red*, *green*, and *blue* color planes was taken as a feature image, for a total of 10. Each feature image was then spatially summed and group-normalized (Σ, see text), resulting in 10 scalars which became input to the neural network model (color figure online)

$$\boldsymbol{i}_{\mathrm{C}}(t) = \left[i_8(t)\ i_9(t)\ i_{10}(t)\right]^{\mathsf{T}} \tag{4}$$

corresponding respectively to the inputs to the first-stage motion, orientation, and color networks shown in Fig. 2. Sim-

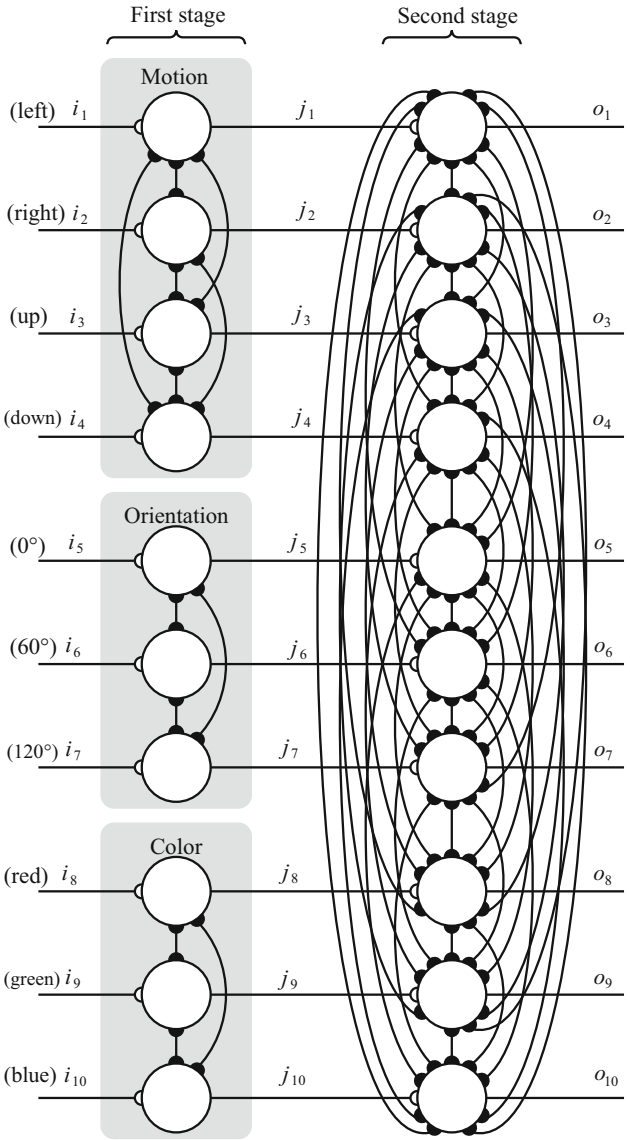

Fig. 2 Diagram of the two-stage neural network model of visual binding. *Large circles* represent units in the neural network. *Unshaded half-circles* at connections indicate excitation, and *filled half-circles* indicate inhibition. The system input consisted of a vector of 10 time-varying inputs $\boldsymbol{i}(t)$ representing spatially summed motion, orientation, and color information. Three first-stage recurrent inhibitory networks refined the selectivity of each visual submodality separately, producing signals $\boldsymbol{j}(t)$ which were then input to an identically organized second-stage network, resulting in outputs $\boldsymbol{o}(t)$

ilarly, we refer to the outputs of the first-stage networks as

$$\boldsymbol{j}(t) = \left[j_1(t)\ j_2(t)\ \ldots\ j_{10}(t)\right]^{\mathsf{T}} \tag{5}$$

where it will again be convenient to define subsets for each submodality $\boldsymbol{j}_{\mathrm{M}}(t)$, $\boldsymbol{j}_{\mathrm{O}}(t)$, and $\boldsymbol{j}_{\mathrm{C}}(t)$ with the same indices as in (2)–(4).

The full set of first-stage output signals $\boldsymbol{j}(t)$ comprised the input to the larger second-stage neural network shown in Fig. 2. The set of outputs from second-stage neurons will be referred to as

$$\boldsymbol{o}(t) = \left[o_1(t)\ o_2(t)\ \ldots\ o_{10}(t)\right]^{\mathsf{T}} \tag{6}$$

which represent the signals projecting from optic glomerulus processing to the central brain.

2.1 Processing of visual inputs

Inputs to the model were sequences of RGB images, each of which had to be converted to grayscale to model biological achromatic motion and orientation processing. We chose the simplest possible algorithm for this conversion by taking the average of the red, green, and blue color values for each individual pixel.

Details of motion, orientation, and color processing are given below.

2.1.1 Motion

Hassenstein and Reichardt (1956) proposed a cybernetic model of the insect optomotor response, which has since been elaborated (van Santen and Sperling 1985) to become the best-accepted model of insect small-field motion detection (Borst and Egelhaaf 1989), and which is mathematically equivalent to models of primate motion detection (Adelson and Bergen 1985). We used a simple version of the elaborated Reichardt detector (ERD) to emulate retinotopic motion processing in the insect compound eye.

Despite the roughly hexagonal organization of the compound eye (which may also be viewed as a distorted rectangular lattice), retinotopic motion computing circuits are organized along the "vertical" and "horizontal" axes of the lattice (Stavenga 1979).

Referring to Fig. 3, horizontal and vertical motion feature images $I_{\mathrm{H}}(x, y, t)$ and $I_{\mathrm{V}}(x, y, t)$ were calculated from the grayscale input image $P(x, y, t)$ at each time t as

$$\begin{aligned} I_{\mathrm{H}}(x, y) &= P_{\mathrm{H}}(x+1, y) \cdot P_{\mathrm{HL}}(x, y) \\ &\quad - P_{\mathrm{H}}(x, y) \cdot P_{\mathrm{HL}}(x+1, y) \end{aligned} \tag{7}$$

$$\begin{aligned} I_{\mathrm{V}}(x, y) &= P_{\mathrm{H}}(x, y+1) \cdot P_{\mathrm{HL}}(x, y) \\ &\quad - P_{\mathrm{H}}(x, y) \cdot P_{\mathrm{HL}}(x, y+1) \end{aligned} \tag{8}$$

where $P_{\mathrm{H}}(x, y, t)$ was $P(x, y, t)$ after being processed at each point (x, y) by a first-order temporal high-pass filter with a time constant of 0.5 s, the intent of which was simply to remove any sustained component of the input signal. $P_{\mathrm{HL}}(x, y, t)$ was $P_{\mathrm{H}}(x, y, t)$ after being further processed at each point (x, y) by a first-order temporal low-pass filter

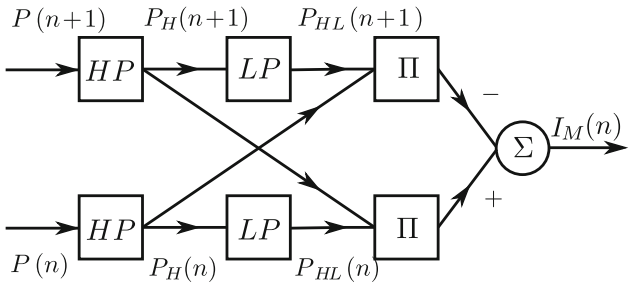

Fig. 3 Computational diagram of one elaborated Reichardt detector (ERD) unit at spatial position n. 2D arrays of such units were computed in both *horizontal* and *vertical* orientations to compose motion feature images. Each ERD required the grayscale photoreceptor input $P(n)$ along with a neighboring input $P(n+1)$, and passed these signals through a set of high-pass (HP) and low-pass (LP) temporal filters as shown. After the final multiplication (Π) and difference (Σ), the magnitude and sign of the output signal $I_M(n)$ reflect the speed and direction of visual motion along the orientation from pixel n to pixel $n+1$

with a time constant of 50 ms, used to introduce phase delay. After cross-multiplication and subtraction, these computations provide signed feature images $\boldsymbol{I}_{\mathrm{H}}$ and $\boldsymbol{I}_{\mathrm{V}}$ representing the spatiotemporal "motion energy" (Adelson and Bergen 1985) at each pixel in both horizontal and vertical directions.

To compute the four motion feature images, we eliminated negative signs by computing outputs representing each of the four cardinal directions separately

$$I_{\mathrm{left}}(x,y) = \mathrm{pos}(-I_{\mathrm{H}}(x,y)) \tag{9}$$

$$I_{\mathrm{right}}(x,y) = \mathrm{pos}(I_{\mathrm{H}}(x,y)) \tag{10}$$

$$I_{\mathrm{down}}(x,y) = \mathrm{pos}(-I_{\mathrm{V}}(x,y)) \tag{11}$$

$$I_{\mathrm{up}}(x,y) = \mathrm{pos}(I_{\mathrm{V}}(x,y)) \tag{12}$$

where

$$\mathrm{pos}(s) = \begin{cases} s & s \geq 0 \\ 0 & s < 0 \end{cases} \tag{13}$$

The four scalar motion signals $\hat{i}_1(t)$, $\hat{i}_2(t)$, $\hat{i}_3(t)$, and $\hat{i}_4(t)$, comprising the vector $\hat{\boldsymbol{i}}_{\mathrm{M}}(t)$, were computed by spatial sums over all x and y of the four motion feature images of (9)–(12) above, and respectively provide wide-field scalar measurements of global image motion in the leftward, rightward, downward and upward directions. The hat notation is used to denote "raw" input signals prior to adaptive group normalization (explained in Sect. 2.1.4).

2.1.2 Orientation

Cells that respond preferentially to orientation of visual stimuli have been observed in a plethora of organisms, including felines (Hubel and Wiesel 1959), primates (Hubel and Wiesel 1968) and honeybees (Srinivasan et al. 1994). "Center-surround" orientation selectivity has been mathematically modeled in numerous ways, including Gabor wavelets (Adelson and Bergen 1985) and by using a difference-of-Gaussians (DoG) function (Rodieck 1965).

The leading model of orientation selectivity in insects supports a direct neuronal implementation of the DoG model (Rivera-Alvidrez et al. 2011). This model, based on both electrophysiological and neuroanatomical evidence, makes use of spatial spreading of photoreceptor inputs by two distinct types of amacrine cells that results in two Gaussian-blurred versions of the input image. Subtraction of these two blurred images can produce a literal difference of Gaussians.

In contrast to visual motion, which is computed along two axes of the compound eye and thus four directions, behavioral data on orientation selectivity in honeybees (Yang and Maddess 1997; Srinivasan et al. 1994) suggests that insects are maximally sensitive to three orientations, which may seem more natural given the hexagonal shape of the compound eye.

For these reasons, we have chosen to model orientation selectivity with DoG functions at three orientations: $\theta_{s1} = 0°, \theta_{s2} = 60°$ and $\theta_{s3} = 120°$. The shape of these functions was chosen to approximate electrophysiological data on narrowing of orientation selectivity by optic glomeruli (Strausfeld et al. 2007).

DoG filter kernels $G(x, y, \theta_s)$ with orientation preference θ_s were computed as

$$x_r(x,y,\theta_s) = -x \cdot \sin(\theta_s) - y \cdot \cos(\theta_s) \tag{14}$$

$$y_r(x,y,\theta_s) = x \cdot \cos(\theta_s) - y \cdot \sin(\theta_s) \tag{15}$$

$$G(x_r, y_r, \theta_s) = \frac{e^{-\left(x_r^2/(2\sigma_{x1}^2)+y_r^2/(2\sigma_{y1}^2)\right)}}{2\pi\,\sigma_{x1}\,\sigma_{y1}} - \frac{e^{-\left(x_r^2/(2\sigma_{x2}^2)+2\,y_r^2/(2\sigma_{y2}^2)\right)}}{2\pi\,\sigma_{x2}\,\sigma_{y2}} \tag{16}$$

in which (14) and (15) serve to rotate the coordinate system to the desired angle θ_s, and in (16), σ_{x1} and σ_{y1} are constants dictating the x and y size and shape of the "center" Gaussian, just as σ_{x2} and σ_{y2} do for the "surround" Gaussian. The kernel $\boldsymbol{G}(\theta_s)$ is formulated to have zero spatial sum and therefore reject the mean spatial intensity. In our simulations, we used $\sigma_{x1} = 19$, $\sigma_{y1} = 6$, $\sigma_{x2} = 22$, and $\sigma_{y2} = 9$, all in units of pixels. For convenience in referring to visual stimuli later, we have adopted the angular convention that a bar with 0° orientation had its long axis perfectly vertical.

At each time t, 2D spatial convolution of the dynamic grayscale image $\boldsymbol{P}(t)$ with each of the three static filter kernels $\boldsymbol{G}(\theta_{s1})$, $\boldsymbol{G}(\theta_{s2})$, and $\boldsymbol{G}(\theta_{s3})$ produced three orientation feature images $\boldsymbol{I}_{0°}(t)$, $\boldsymbol{I}_{60°}(t)$, and $\boldsymbol{I}_{120°}(\mathrm{t})$. Each of the three kernels was computed at full image resolution and convolution was accomplished by multiplication in the frequency

domain. Spatial sums over all x and y of the absolute value (so that both signs of contrast are represented) of each of these three feature images respectively produced scalar orientation features $\hat{i}_5(t)$, $\hat{i}_6(t)$, and $\hat{i}_7(t)$, which together comprise the vector $\hat{\boldsymbol{i}}_{\mathrm{O}}(t)$.

2.1.3 Color

A multitude of organisms, including flies, honeybees, and humans, have trichromatic visual systems (Land and Nilsson 2002). As mentioned earlier, despite the well-known spectral shift between human and insect photoreceptor tunings, for convenience of human visualization and internal representation we have made use of the three colors commonly used in computer image formats: red, green and blue (RGB). If input images were provided instead with "color planes" of green, blue, and UV, as if viewed by fly photoreceptors (Snyder 1979), the model would produce similar results to what is shown here for RGB images.

Since color is explicitly represented in the image, the color "feature images" $\boldsymbol{I}_{\mathrm{red}}(t)$, $\boldsymbol{I}_{\mathrm{green}}(t)$, and $\boldsymbol{I}_{\mathrm{blue}}(t)$ were taken simply as the red, green, and blue "color planes" of the image (that is, the 2D array of pixels of a given color). Spatial sums over all x and y produced three scalar color features $\hat{i}_8(t)$, $\hat{i}_9(t)$, and $\hat{i}_{10}(t)$, which together comprise the vector $\hat{\boldsymbol{i}}_{\mathrm{C}}(t)$.

2.1.4 Adaptive group normalization

Due to the vast differences in the algorithms presented above for computing motion, orientation, and color inputs, the "raw" features $\hat{\boldsymbol{i}}_{\mathrm{M}}(t)$, $\hat{\boldsymbol{i}}_{\mathrm{O}}(t)$, and $\hat{\boldsymbol{i}}_{\mathrm{C}}(t)$ differ by orders of magnitude. In order to make these signals comparable to one another, and to simultaneously account for the dynamic nature of visual imagery, each of these raw input vectors was normalized by a scalar adaptive factor computed as the maximum value of any element of the vector in the recent past.

Specifically, at each time t each of the three vectors of inputs to the first-stage network was computed as

$$\boldsymbol{i}(t) = \hat{\boldsymbol{i}}(t)/\mathrm{M}(t) \tag{17}$$

where $\mathrm{M}(t)$ was a scalar group normalization factor computed as the maximum value of any element of vector $\hat{\boldsymbol{i}}(t)$ over the prior 2 s. If the normalization factor $\mathrm{M}(t)$ was zero, indicating that all components of any given input vector were zero in the recent past, $\boldsymbol{i}(t)$ was set to zero.

This operation, repeated independently for vectors representing each of the three visual submodalities, provided input vectors $\boldsymbol{i}_{\mathrm{M}}(t)$, $\boldsymbol{i}_{\mathrm{O}}(t)$, and $\boldsymbol{i}_{\mathrm{C}}(t)$, all elements of which remained comparable in magnitude even as the image changed, with each group of signals sustaining a maximum value of approximately unity. Despite the simplicity of this technique, it can be viewed as implementing a form of adaptation quite similar to that seen at multiple levels in biological vision systems.

2.2 Network temporal evolution

The neural network model shown in Fig. 2 employs two stages of processing. The first stage incorporates three independent networks which refine inputs $\boldsymbol{i}_{\mathrm{M}}(t)$, $\boldsymbol{i}_{\mathrm{O}}(t)$, and $\boldsymbol{i}_{\mathrm{C}}(t)$ from each of the visual submodalities into intermediate outputs $\boldsymbol{j}_{\mathrm{M}}(t)$, $\boldsymbol{j}_{\mathrm{O}}(t)$, and $\boldsymbol{j}_{\mathrm{C}}(t)$. The second stage uses a fourth larger network to combine all outputs $\boldsymbol{j}(t)$ from the first-stage networks and learn an internal representation of common temporal fluctuations within this group of inputs, resulting in a vector of outputs $\boldsymbol{o}(t)$.

We have chosen to use a two-stage network not only because optic glomeruli have been observed to refine the representation in one specific submodality (specifically, orientation: see Strausfeld et al. 2007), but because the sensory refinement from the first stage greatly improves learning of the second stage (see Results).

Despite the apparently dissimilar purposes of the two stages in our network, all four neural networks employed have identical structure and differ only in the number of inputs and thus neurons used, and in parameters of the learning rule. For each, we have used a fully connected recurrent inhibitory network which learns by changing a weight matrix which represents the inhibition between each pair of neurons. In this section we describe all networks generically, providing the time evolution equations and learning rule for a network of N neurons with a generalized column vector of inputs $\boldsymbol{i}(t)$, an $N \times N$ inhibitory weight matrix $\boldsymbol{W}$, and the corresponding column vector of outputs $\boldsymbol{o}(t)$.

All inputs to each of the four networks were processed through a high-pass filter, since neurons rarely pass on information about unchanging signals. The outputs of each network were also processed through a high-pass filter as part of the learning rule. In this section we use the compact notation $i'(t)$ to represent a first-order high-pass-filtered version of the signal $i(t)$, and $o'(t)$ to represent a first-order high-pass-filtered version of the signal $o(t)$. The time constant of the high-pass filter used on network inputs, the purpose of which is to prevent long-term sustained inputs (such as the color of a static background) from ever entering the network, was $\tau_{\mathrm{HI}} = 1.0\mathrm{s}$. The time constant of the high-pass filter used on network outputs in the learning rule described below was $\tau_{\mathrm{HO}} = 0.5\mathrm{s}$.

Our network was inspired by the seminal work of Herault and Jutten (1986) on blind source separation, and by that of Hopfield (1991), who modeled olfactory perception using temporal fluctuations in the mammalian olfactory bulb, but is distinct from both prior networks as detailed below.

2.2.1 Temporal dynamics

Early vision in the insect optic lobes is dominated by cells that represent signals with graded potentials (Arnett 1972) rather than with trains of action potentials as is the case in mammals (Albrecht and Geisler 1991). Like the optic lobes from which they take their inputs, the optic glomeruli we model here are comprised primarily of graded potential neurons, but also contain a number of spiking interneurons (Mu et al. 2012), with their detailed interconnection pattern yet unknown.

As in a multitude of prior neural network models including the ones which inspired the present network (Herault and Jutten 1986; Hopfield 1991; Anderson 1995), we allow the outputs of individual neurons to be either positive or negative, primarily for reasons of analytical tractability. Despite this oversimplification of the electrical responses of real neurons, as has been long argued for prior networks, neurons in our network may be considered an approximate model of either graded potential neurons, or (to a lesser extent) of spiking neurons. In the case of graded potential neurons, network outputs may be reasonably considered to model a scaled version of the neuronal potential relative to its resting potential; in this case, negative network outputs simply indicate a neuronal response that is inhibited with respect to rest. In the case of a spiking neuron with a nonzero spontaneous firing rate, network outputs may be considered to model the time-averaged neuronal firing rate relative to the spontaneous rate. However, since the spontaneous firing rates of neurons vary widely and may be very small, the spiking neuron approximation is less accurate than for graded potential neurons. Since optic glomeruli are primarily comprised of graded potential neurons, this network provides a reasonable compromise between modeling accuracy and analytical tractability.

By a similar line of reasoning, the "weighted sum" temporal evolution rule common to decades of neural networks—a variation of which is described below for our model—may be justified as an approximate model of neuronal interactions. Direct input from presynaptic graded potential cells in insects leads to similarly shaped postsynaptic potentials (Douglass and Strausfeld 2005), with both excitation and inhibition relative to the presynaptic resting potential being passed through some synaptic weight to postsynaptic neurons. The response of a graded potential neuron with multiple presynaptic connections may be modeled as a sum of the presynaptic inputs relative to resting, with each input weighted by the strength of the corresponding synapse. For a spiking neuron, over some limited range of integrated postsynaptic currents the average firing rate is proportional to the total current input (Koch 1999). Averaged over time, a train of action potentials from multiple presynaptic neurons may be reasonably modeled as providing a postsynaptic current input proportional to the firing rate of each presynaptic neuron weighted by the strength of the corresponding synaptic interconnection.

Given the justifications above, we can model the activation $o_n(t)$ of neuron n as

$$o_n(t) = i'_n(t) - \sum_{k=1}^{N} W_{n,k} \cdot o_k(t - \tau_i) \tag{18}$$

where $i'_n(t)$ represents a high-pass-filtered excitatory input, $W_{n,k}$ represents the strength of the inhibitory synaptic pathway from neuron k to neuron n, and $o_k(t)$ is the activation of a different neuron k in the network. Inhibition between biological neurons may be accomplished directly, or indirectly through an inhibitory interneuron, but in either case, it inevitably results in a finite delay, which we represent as a single lumped delay τ_i. This equation may be written in matrix form as

$$\boldsymbol{o}(t) = \boldsymbol{i}'(t) - \boldsymbol{W} \cdot \boldsymbol{o}(t - \tau_i) \tag{19}$$

thus expressing the current activation of each neuron as a sum of the corresponding high-pass-filtered input with a weighted sum of the delayed inhibitory activation of all other neurons (as described in the next section, diagonal elements of $\boldsymbol{W}$ were constrained to be zero to avoid self-inhibition). Since biophysical details of the inhibition within optic glomeruli are not yet available, the value of τ_i is unknown, but the very *existence* of this finite inhibition delay is (as we show below) crucial to the function of the model. For this reason, we have formulated the temporal dynamics of our model as

$$\boldsymbol{o}(t) = \boldsymbol{i}'(t) - \boldsymbol{W} \cdot \boldsymbol{o}(t - \Delta t) \tag{20}$$

where Δt is the simulation time step of 10 ms. The use of Δt as the inhibition delay τ_i provides the smallest finite delay possible in our model. This equation for temporal dynamics was used in all simulations.

In the case when the simulation time step Δt is much smaller than the time course of changes in the high-pass-filtered inputs $i'_n(t)$, (20) may be approximated as

$$\boldsymbol{o}(t) = \boldsymbol{i}'(t) - \boldsymbol{W} \cdot \boldsymbol{o}(t) \tag{21}$$

Equation (21)–apart from the high-pass filtering of the inputs—has long been a common formulation for a fully connected inhibitory neural network used in blind source separation (Herault and Jutten 1986; Jutten and Herault 1991; Cichocki et al. 1997). However, while (21) is linear and well suited for theoretical analysis, it is not a realistic model of any physical system because the outputs have *absolutely no time-dependence* on their own history or that of any other signal. In fact, directly from this equation the outputs $\boldsymbol{o}(t)$ can be computed instantaneously as

$$\boldsymbol{o}(t) = [\boldsymbol{I} + \boldsymbol{W}]^{-1} \cdot \boldsymbol{i}'(t) \tag{22}$$

(where $\boldsymbol{I}$ is the identity matrix) so long as $[\boldsymbol{I} + \boldsymbol{W}]$ is not singular. Thus if the input $\boldsymbol{i}'(t)$ changes radically in a femtosecond, so will the output, meaning that the network has no true "dynamics," but rather computes an instantaneous function of the inputs. This cannot be true for any realistic neuronal model. Further, since the outputs can be computed instantaneously without any history dependence, a network described by (21) can be singular and thus impossible to evaluate, but cannot be temporally unstable.

The seemingly minor difference between Eq. (20) and (21) has significant consequences to the dynamics of the network, despite the fact that the time scale of changes to network inputs and outputs is typically much larger than the simulation time step Δt, making the approximation of (21) quite reasonable. Unlike the approximate equation, the recurrent network of (20) contains closed loops through which a signal could pass over time, growing larger with each pass if any 'loop gain' were greater than one, thus leading to the possibility of temporal instability under certain conditions of the inhibitory weight matrix $\boldsymbol{W}$.

The stability of systems of equations such as (20) has long been studied in the theory of linear control systems (Trentelman et al. 2012), and the condition for temporal stability is most simply stated by requiring that the magnitude of all eigenvalues of the weight matrix $\boldsymbol{W}$ be strictly less than unity. This condition is equivalent to guaranteeing that the loop gain around all loops in the network is less than one. The closer the magnitude of the eigenvalues of $\boldsymbol{W}$ are to unity, the more the system is prone to oscillation in response to high temporal frequency inputs.

For these reasons, we only use the approximation of (21) when required to make theoretical analysis tractable (see Northcutt and Higgins 2016), while (20) is used in all simulations.

To distinguish between the specific weight matrices of our four networks, the generic symbol $\boldsymbol{W}$ used above will be replaced for the first-stage motion, orientation, and color networks, respectively, with $\boldsymbol{M}$ (4×4), $\boldsymbol{O}$ (3×3), and $\boldsymbol{C}$ (3×3), and for the second-stage network with $\boldsymbol{T}$ (10×10).

2.2.2 Learning rule

Given the fully connected inhibitory structure of these networks, the function of the model is largely dictated by the learning rule implemented. The learning rule described below is common to all four networks in our model and serves to detect common temporal fluctuations in a set of input signals. In the case of the first stage, this has the effect of refining the representation of each visual submodality by developing lateral inhibition between elements which are simultaneously activated. For the second stage, this same learning rule develops inhibitory associations between inputs from the first stage which come to represent the characteristics of distinct objects in the visual scene.

The learning rule for each of our four networks, used to generate the inhibitory weight matrices generically described as $\boldsymbol{W}$ based on common temporal fluctuations of the network inputs, is a modified version of the learning rule of Cichocki et al. (1997), which itself is a refinement of Hebb's venerable learning rule (Hebb 1949). Hebbian learning, first modeled as an increase in synaptic strength when the average firing rate of pre- and postsynaptic neurons was simultaneously large, is now associated with the biological phenomena of long-term potentiation and depression (Markram et al. 1997; Bi and Poo 1998; Song et al. 2000). These phenomena—which intriguingly were modeled by Gerstner et al. (1996) before the seminal biological results were published—describe how synaptic efficacy increases or decreases depending on the relative timing of pre- and postsynaptic neuronal firing. Since our model does not explicitly incorporate spiking neurons, using a learning rule based on this spike-timing-dependent plasticity (STDP) is not possible. However, Gerstner and Kistler (2002) have shown that when pre- and postsynaptic spikes are generated from independent Poisson processes, very similar results to STDP may be obtained from a learning rule based on average firing rate. Such a rule is used in our networks and described below and is chosen because it provides very well-developed spatially asymmetric Hebbian learning and also because it fits well into the existing theoretical framework for blind source separation. With this being said, as noted earlier, spiking neurons *are present* in optic glomeruli—although their connection pattern is yet unknown and thus not yet modeled—and STDP may well be the underlying biological basis for the learning modeled here.

In our simulations, weight matrices $\boldsymbol{W}$ were initialized to zero so that the initial state of the system was $\boldsymbol{o}(t) = \boldsymbol{i}'(t)$, and thus before learning began, network outputs were exactly equal to the high-pass filtered inputs. Each off-diagonal element $W_{n,k}$ $(n \neq k)$ of the weight matrix was learned based on high-pass-filtered versions of network outputs $o_n(t)$ and $o_k(t)$ as

$$\frac{\mathrm{d}W_{n,k}}{\mathrm{d}t} = \gamma \cdot \mu(t) \cdot g(o_n') \cdot f(o_k') \tag{23}$$

where γ is a scalar learning rate. The learning onset function $\mu(t)$ was used to prevent sudden weight changes at the time t_{train} at which learning began

$$\mu(t) = \left(1 - e^{-(t - t_{\text{train}})/\tau_l}\right) \cdot u(t - t_{\text{train}}) \tag{24}$$

where $\tau_l = 2$ s is the time constant used to gradually activate the learning rule, and $u(t)$ is the unit step function. Weights were updated at each simulation timestep by numerical integration of (23). Diagonal elements of $\boldsymbol{W}$ were held

at zero, thus preventing self-inhibition. Any element of $\boldsymbol{W}$ that became negative from a learning rule update was set to zero to avoid unintentional excitation.

The high-pass filters used on outputs $o_n(t)$ and $o_k(t)$ caused learning of the weight matrix to be dependent on *temporal fluctuations* of the input, rather than simply on input values. This was true despite the fact that inputs were already high-pass-filtered, because the time constant $\tau_{\mathrm{HO}} = 0.5$ s of the high-pass filter used on the outputs was smaller than the one previously used on the inputs with time constant $\tau_{\mathrm{HI}} = 1.0$ s, resulting in a higher cutoff frequency that attenuated lower-frequency signals.

Key to the learning rule are the nonlinear "activation functions" $f()$ and $g()$ through which the high-pass-filtered outputs were processed before being used for learning, and without which the learning rule is symmetrically Hebbian, and may only develop symmetric weight matrices $\boldsymbol{W}$. These activation functions were used to introduce higher than second-order statistics of the filtered outputs into the learning rule, and an extremely wide variety of choices is possible (Hyvärinen and Oja 1998). We have empirically chosen $f(x) = x^3$ and $g(x) = \tanh(\pi x)$ to improve separation of signals in the present model, similar to the activation functions long used for blind source separation networks (Herault and Jutten 1986; Jutten and Herault 1991; Cichocki et al. 1997). However, in our learning rule, the positions of the expansive and compressive activation functions $f()$ and $g()$ are exchanged with one another as compared to previous work on blind source separation, with $f()$ applying to column elements k and $g()$ to row elements n.

As addressed in detail in a companion paper (Northcutt and Higgins 2016), this exchange of activation function positions has the effect of optimizing our network's learning for the "overdetermined case" (Joho et al. 2000) in which the number of hidden sources to be separated is less than the number of neurons. The overdetermined case has rarely been considered crucial in blind source separation, since in most cases the number of network inputs (for example, microphones in an auditory case) may be easily changed to match the number of hidden sources present. For this reason, the overdetermined case is less well addressed in the literature. However, given the fixed size (10 units) of our second-stage network, and the unknown number of distinct objects in the input image sequence, this is always the case for our second-stage visual binding network.

2.3 Training of first-stage networks

The purpose of the first stage of our model is to sharpen the representation of each sensory modality by learning lateral inhibition, a well-known technique for sensory refinement (Linster and Smith 1997) that has been proposed as a method by which redundant information is removed from photoreceptor signals in the fly visual system (Laughlin 1983).

Because we consider the first-stage network to represent long-term learning from visual experience rather than developing a representation of the current visual scene as in the second stage, all three first-stage networks (color, motion, and orientation) were trained simultaneously using a visual stimulus specifically designed to elicit equal response from all visual submodalities. This visual stimulus is a radially symmetric contracting pattern of concentric rings with slowly flickering overall brightness and is described mathematically at each point (x, y) and time t by

$$\Theta(t) = 2\pi \cdot f_{\mathrm{f}} \cdot t \tag{25}$$

$$\Psi(r, t) = 2\pi \cdot f_{\mathrm{R}} \cdot r + 2\pi \cdot f_{\mathrm{m}} \cdot t \tag{26}$$

$$S(r, \Theta, \Psi) = e^{-\frac{r^2}{2\sigma_S^2}} \left(\frac{1 + \sin(\Theta)}{2}\right)\left(\frac{1 + \cos(\Psi)}{2}\right) \tag{27}$$

where $r = \sqrt{x^2 + y^2}$ is the radial distance from the stimulus image center. The first term of (27) is a radial Gaussian envelope with spatial standard deviation $\sigma_S = 25$ pixels. The second term provides a temporal flicker with frequency $f_{\mathrm{f}} = 0.5$ Hz. The third term describes a pattern of contracting radial rings with spatial frequency $f_{\mathrm{R}} = 0.2$ cycles per pixel and temporal frequency $f_{\mathrm{m}} = 0.5$ Hz.

The visual stimulus of (27) was provided before training of the first-stage networks began for a time $t_{\mathrm{train},1} = 4$ s sufficient for all temporal filters and the input adaptation algorithm described in Sect. 2.1.4 to stabilize.

Unless otherwise specified below, all image sequences were presented at 100 frames per second. The learning rates used for first-stage motion, orientation, and color networks respectively were $\gamma_{\mathrm{M}} = 5$, $\gamma_{\mathrm{O}} = 5$, and $\gamma_{\mathrm{C}} = 5$. Because the visual stimulus of (27) provides identical signals to all inputs of each of the three submodalities, it functionally reduces the learning rule of (23) to a purely symmetric Hebbian rule, a situation in which all network weights will increase uniformly so long as the network continues to be trained. Therefore, to guarantee temporal stability of the final network, we continued training each first-stage network only until the magnitude of the largest eigenvalue of each weight matrix reached a value of $V_{1,\max} = 0.9$, after which the corresponding learning rate γ for that network was set to zero, terminating training. First-stage training was considered complete when all three networks had reached this state.

The second stage was not trained (γ_2 was set to zero) until all three first-stage networks had finished training, after which the weight matrices of the three first-stage networks were fixed. It would certainly be possible to train both first and second stages simultaneously, thus using a meaningful image sequence to train the first stage rather than the contrived stimulus of (27), and after a longer training period than

that shown in Results, quite similar results to those shown would be obtained. However, to most clearly demonstrate the function of each stage, we have trained each independently.

2.4 Training of second-stage networks

As with the first stage, the visual stimulus was provided before training for a time $t_{\text{train},2} = 4$ s sufficient for all temporal filters, the input adaptation algorithm, and the first-stage networks to stabilize, after which training began.

Unless otherwise specified below, the learning rate used for the second-stage network was $\gamma_2 = 0.5$. Since the second-stage network model is intended to learn continually in order to reflect changing objects in the visual scene, no condition for stopping its training was required. However, during training, we ensured network stability by limiting the maximum magnitude of any eigenvalue of the connection matrix $\boldsymbol{T}$ to $V_{2,\max} = 0.95$. If, after any update of the connection matrix, the maximum eigenvalue magnitude V exceeded $V_{2,\max}$, the matrix $\boldsymbol{T}$ was multiplied by a scalar factor $V_{2,\max}/V$, which had the effect of reducing the maximum eigenvalue to exactly $V_{2,\max}$.

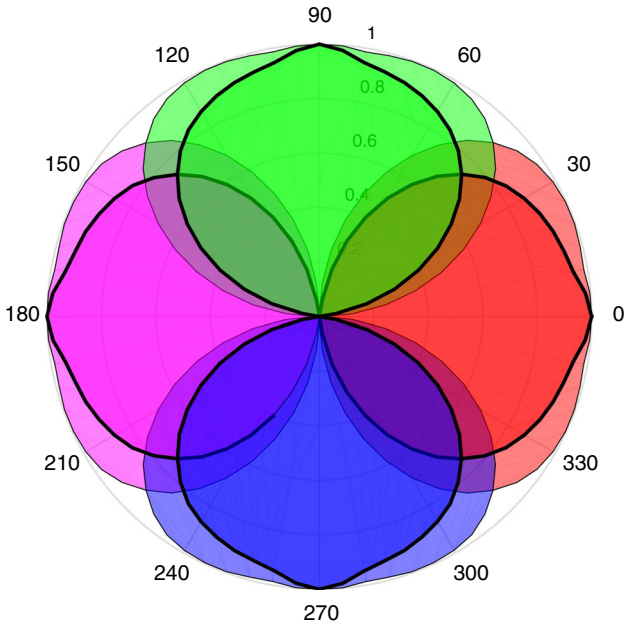

Fig. 4 First-stage motion refinement. On this polar plot, inputs $\boldsymbol{i}_{\text{M}}(t)$ are visible as *thin-outlined* nearly *circular lobes* in each of the four cardinal directions plotted against the direction of visual stimulus motion. Outputs $\boldsymbol{j}_{\text{M}}(t)$ are *outlined in bold* and are clearly narrowed in angular extent with respect to the inputs, although this effect is not pronounced

3 Results

All experiments were performed in MATLAB (The MathWorks, Natick, MA). For all but the last of the experiments shown below, the fundamental visual stimulus element was a 50×12 pixel bar on a black background. To characterize the first-stage networks, a single bar was presented in a sequence of images—each of which was 100 pixels wide by 100 pixels high—in which the direction of motion, orientation, or color varied during the experiment.

For all second-stage visual binding experiments but the last shown below, one, two, or three bars were presented in sequences of 500 pixel wide by 500 pixel high images as different parameters of the stimulus were varied as described below.

3.1 Motion refinement

The motion refinement network was trained as described in Sect. 2.3, and the resulting 4×4 connection matrix $\boldsymbol{M}$ was nearly uniform with all off-diagonal values approximately equal to 0.3.

To demonstrate the effect of the trained motion refinement network, an image sequence containing a bar moving orthogonal to its orientation was first presented to the network. A vector of four inputs $\boldsymbol{i}_M$ was computed from this input image sequence and processed through the first-stage motion network to produce refined outputs $\boldsymbol{j}_M$. Outputs were allowed time to stabilize, after which their value was recorded. The orientation of this bar was varied over all possible angles, and the results are shown in Fig. 4. Due to the operation of the HR motion detector, inputs on this polar plot appear as near-circular lobes oriented in each of the four cardinal directions. Outputs are outlined in bold and are clearly narrower in angular extent than the inputs, but this narrowing is not exaggerated due to the excellent angular separation of the inputs.

Because the motion inputs were already well separated in angle, does that mean that the first-stage motion network has little or no effect? To show that this is not the case, we presented image sequences in which the bar always moved to the right (0°), but varied in orientation from −85° (leaning to the far left) to 85° (leaning to the far right), with an orientation of 0° meaning that it moved orthogonal to its longest dimension. This stimulus demonstrates the well-known *aperture problem* (Nakayama and Silverman 1988), which arises in visual motion detection when the small spatial extent of a local motion detector makes it impossible to unambiguously resolve the global direction of an object's motion. Due to the aperture problem, an angled bar moving strictly to the right generates signals from small-field motion detectors in vertical directions as well.

Figure 5 shows the output of the motion refinement network in response to these stimuli. Note that across the entire angular extent, in cases where more than two motion inputs were simultaneously active, the weakest output is almost

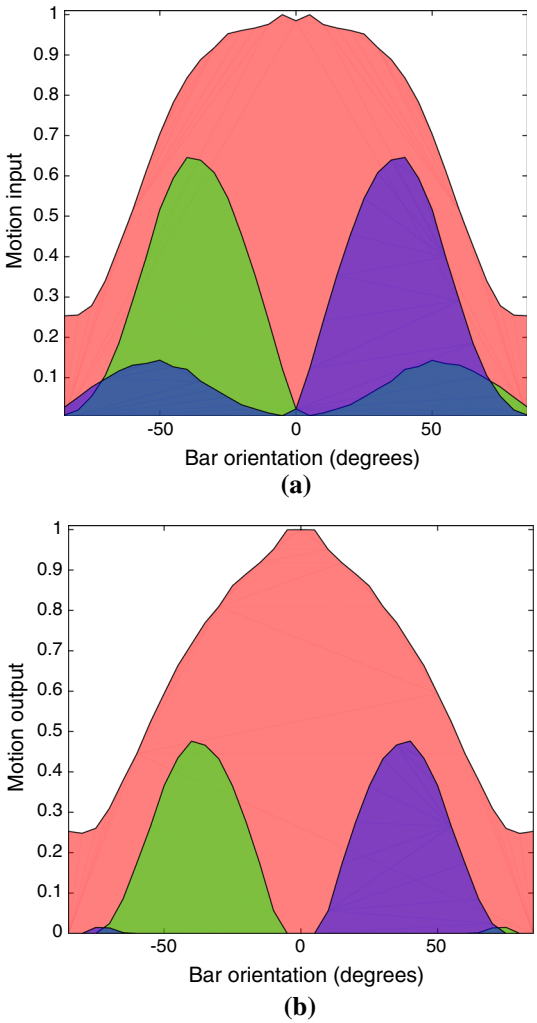

Fig. 5 First-stage motion processing in the presence of the aperture problem. **a** Motion inputs to the first-stage network as the orientation of a bar that always moved to the right (0°) was varied from −85° to 85° (plotted on the *horizontal axis*). The large *central lobe* peaking at 0° corresponds to the desired response (*rightward* motion), whereas the two smaller lobes that peak at −45° and 45° respectively correspond to motion in the *upward* and *downward* directions, and result from local motion detector responses to the vertical components of motion from all four edges of the *bar*. **b** Corresponding motion outputs from the first-stage network, showing significant reduction of the undesired *upward* and *downward* responses

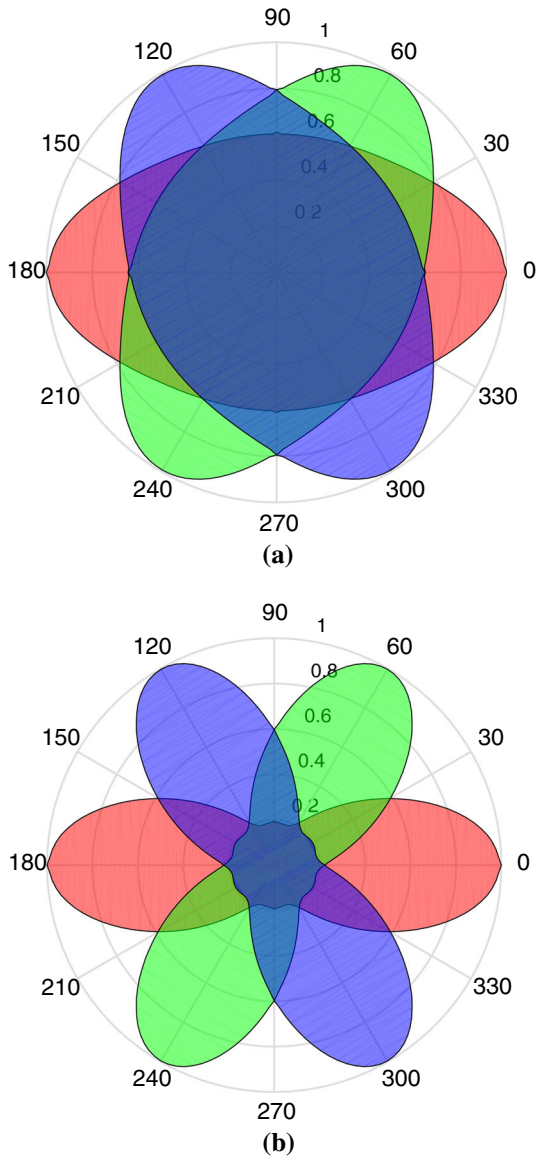

Fig. 6 First-stage orientation refinement. **a** Inputs to the orientation network *plotted* against bar angle in degrees showing three elliptical responses oriented at 0°, 60°, and 120°, directly resulting from the DoG filter of (16) with parameters given in Sect. 2.1.2 operating on a rectangular bar stimulus. **b** Outputs from the orientation refinement network, with the narrower "peanut shapes" indicating a clear reduction of angular overlap between outputs as compared to inputs. Note that we have adopted the angular convention that a bar with 0° orientation had its long axis perfectly vertical

completely suppressed. The undesired upward and downward responses are reduced in both magnitude and angular extent in the outputs relative to the inputs, resulting in a reduction of the ambiguity in the direction of bar motion.

3.2 Orientation refinement

The first-stage orientation network, which processed a vector of three inputs $\boldsymbol{i}_{\mathrm{O}}$ computed from the input image sequence to produce refined outputs $\boldsymbol{j}_{\mathrm{O}}$, was trained as described in Sect. 2.3, and the resulting 3×3 connection matrix $\boldsymbol{O}$ was nearly uniform with all off-diagonal values approximately equal to 0.45.

The orientation network was tested by presenting a centered stationary bar and recording inputs and outputs as the orientation of the bar was varied. Figure 6 shows the results of this experiment.

The elliptical shape of each of the three input orientation responses in Fig. 6a is due to the mix of the small-field

responses from the long edges at the sides of the rectangular bar with the shorter orthogonal edges at the top and bottom. Note that, since each filter is tuned for stimulus orientation rather than direction, each is equally sensitive to the angle θ_s used in (16) and to $\theta_s + 180°$. Figure 6b shows the output responses, which exhibit a distinct angular narrowing in orientation relative to the inputs due to the lateral inhibition of this network.

3.3 Color refinement

The first-stage color network, which processed an RGB vector of inputs $\boldsymbol{i}_C$ computed from the input image sequence to produce refined outputs $\boldsymbol{j}_C$, was trained as described in Sect. 2.3, and the resulting 3×3 connection matrix $\boldsymbol{C}$ was nearly uniform with all off-diagonal values approximately equal to 0.45.

The color network was tested by presenting a stationary bar which varied only in color. To demonstrate the improvement in color separation provided by this network, we varied input color using a standard HSL (hue, saturation, lightness) model of color, an alternative to RGB that is effectively a Cartesian-to-cylindrical coordinate transformation. Each HSL triplet has a unique corresponding RGB triplet, and vice versa.

We fixed the saturation of all input image colors at 0.2 (20%), intentionally making them very weak in comparison to one another, as we varied the hue and lightness of the color over their entire range as shown in Fig. 7a. Each point in this panel corresponds to an HSL triplet which was converted to RGB and then used as the color of a bar stimulus to the first-stage color network. Figure 7b shows the corresponding output colors at the position of the hue and lightness of the input. Note the marked increase in the distinction between colors: this is effectively an increase in color saturation. Figure 7c shows a cross section through the center of Fig. 7b at a lightness of 0.5. As hue is varied on the horizontal axis, the corresponding red, green, and blue input color components trade off with one another as dictated by the HSL color model. The output colors are clearly much better distinguished from one another than the inputs due to color network lateral inhibition. Note that this effect could not be achieved by simply rescaling the inputs.

3.4 Visual binding

The second and final stage of the model shown in Fig. 2 took as input the vector $\boldsymbol{j}(t)$, the combined output of the three first-stage networks, which contained ten scalar values representing refined measures of motion, orientation, and color in the input visual image sequence. After learning of the connection matrix $\boldsymbol{T}$ was complete, the second stage produced an output vector $\boldsymbol{o}(t)$ in which a small number of outputs rep-

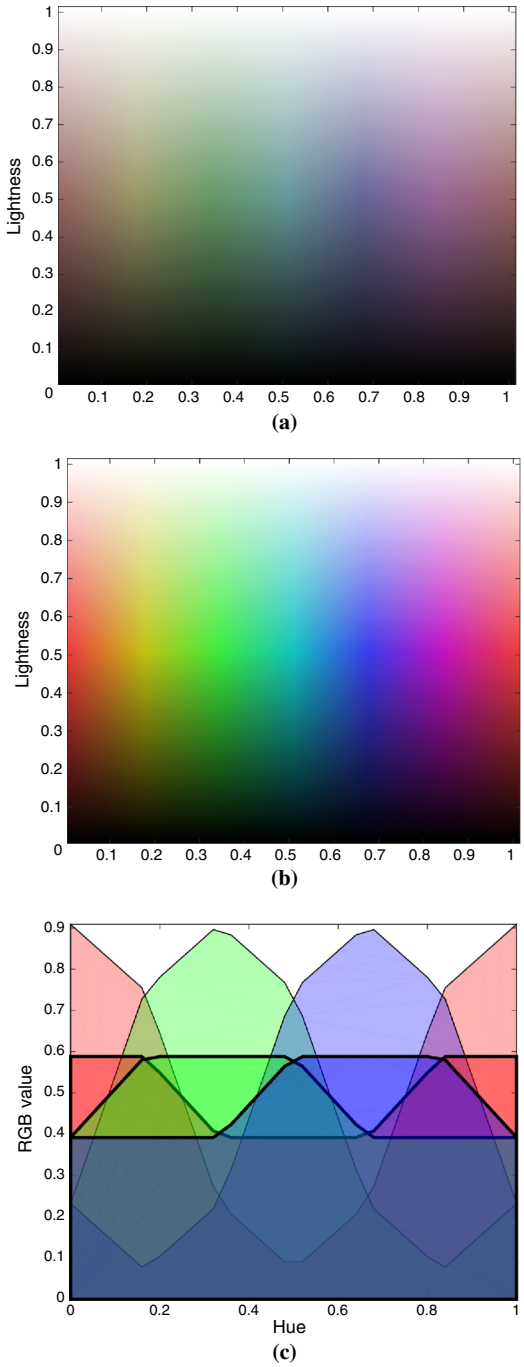

Fig. 7 First-stage color refinement. In all panels, the hue of the input is varied on the *horizontal axis*. Saturation of all colors was fixed at 0.2. The *vertical axis* of panels **a** and **b** corresponds to lightness. **a** Input colors. *Each point* in this image represents a color that was input to the network. **b** Output colors. *Each point* in this image corresponds to the output that was obtained by passing an input of that color through the color refinement network. **c** Output colors plotted as RGB values. Hue is varied on the horizontal axis with saturation fixed at 0.2 and lightness at 0.5, and the corresponding *red*, *green*, and *blue* input color components are shown (*bold lines at center*). The corresponding outputs (larger lobes in the background) show the clear increase in the difference between color responses at the output of the network (color figure online)

resenting the *unique common temporal fluctuations* found in the visual input became dominant, while all other outputs were inhibited.

Due to the fact that the second stage can process any sequence of visual images, it is simply impossible to present an exhaustive set of visual stimuli. Instead, we present below results based on sets of artificial stimuli composed of 50×12-pixel bars demonstrating the capabilities of the model with controlled variations and increasing complexity, and finish with a single demonstration of network operation using a real-world image sequence collected with a camera.

3.4.1 Response to the reference stimulus

Our reference stimulus, which we will use as a basis for comparison as we vary stimulus parameters, was composed of two bars moving on a black background. Bars moved in a direction orthogonal to their long axis, which means—due to the convention we have adopted for bar orientation—that their orientation angle and direction of motion were the same. A "red" bar (RGB = [0.75 0.1 0.1]) started near the upper left corner of the image and moved down and right at an angle of $-30°$. Simultaneously, a "green" bar (RGB = [0.1 0.75 0.1]) started near the upper right corner and moved down and left at an angle of 210°. Bars moved at a speed of 50 pixels per second. Both bars moved through the same pattern of multiplicative horizontal sinusoidal shadowing, which was used to provide predictable temporal fluctuations. This shadowing had a spatial period of 50 pixels per cycle, a mean value of 0.5, and an amplitude of 0.25. The relative phase of the temporal fluctuations generated by these two bars as they moved through the shadow was not chosen to be any particular value, but bar fluctuations were never perfectly in phase, nor precisely quadrature phase or counter-phase. So that we could use a small image resolution and still experiment with training the network over long periods of time, bars wrapped around toroidally to reenter on the opposite side as they left the image, thus creating an arbitrarily long image sequence. The results of training the second-stage network with this two-bar stimulus are detailed in Fig. 8.

Figure 8a shows the time evolution of network outputs for the first 10 s of training. Since the two bars presented were red and green, it is not surprising that the red and green outputs came to dominate all others, and by the end of the period shown had come to inhibit all other outputs. The number of outputs which are *not inhibited* corresponds to the number of objects present in the image, whereas the sinusoidal patterns revealed by the output neurons are the patterns of shadow through which the two bars moved.

Figure 8b and c shows the time evolution of inhibitory weights from columns 8 (red) and 9 (green) of the weight matrix $\boldsymbol{T}$, representing inhibition from neurons 8 and 9 to all other neurons. Note that connection weights have not precisely stabilized; rather, the temporal mean of each weight over the period of input fluctuation has come to a stable value. The other neurons to which each neuron developed inhibition are those with which that neuron had common temporal fluctuations. Thus the pattern of inhibitory weights in each column represents the visual features of each object. This is clarified in Fig. 8d and e, which respectively show the final raw and thresholded weight matrix $\boldsymbol{T}$. The fact that this weight matrix is asymmetric, showing clear patterns of column rather than row inhibition, is due to the asymmetric activation functions described in Sect. 2.2.2. Since small weights have little effect on the network output, further figures only show thresholded weight matrices.

The number of objects and their characteristics can be clearly discerned from Fig. 8e. Based on this matrix, two objects were present. The first was red, got a moderate, roughly equal response from both 0° and 120° orientation filters, and movement to the right was strongly indicated with a less prominent downward component. Referring to Fig. 6, this orientation response indicates a bar orientation either between 0° and $-60°$ or equivalently between 120° and 180°, either of which is correct. From the weight matrix, the second object was green, at an orientation between 0° and 60° (or equivalently between 180° and 240°), and moving to the left with a less prominent downward component. Owing to the direction of motion of both bars being less than 45° from horizontal, the downward component of motion from each bar was weaker in the inputs than the leftward and rightward motion components, and is thus properly represented by the weight matrix.

Although we show results with the learning rate γ_2 set to a small value of 0.5 to allow detailed scrutiny of the development of network weights, a weight matrix correctly representing the objects in the input imagery can be stably learned with values of γ_2 more than 10 times larger (data not shown). A disadvantage that accompanies the higher speed of this learning is an increase in the amplitudes of the oscillations of weights shown in Fig. 8b, which nonetheless stabilize in temporal average to the values shown.

One might reasonably question if the first-stage networks are contributing anything to the operation of the model, and so to test this question, we trained the first-stage networks only to a maximum eigenvalue of 0.1, as compared to our usual standard of 0.9 (refer to Sect. 2.3). This resulted in very weak inhibition in the first-stage connection matrices, and thus first-stage outputs $\boldsymbol{j}(t)$ were very nearly equal to inputs $\boldsymbol{i}'(t)$. Figure 9 shows the time course of second-stage network outputs in response to exactly the same stimulus used to generate the data shown in Fig. 8. Comparing Figs. 8a and 9, second-stage network learning is clearly retarded by a lack of sensory refinement in the first stage, and thus the first-stage networks do indeed provide an essential computation to the model.

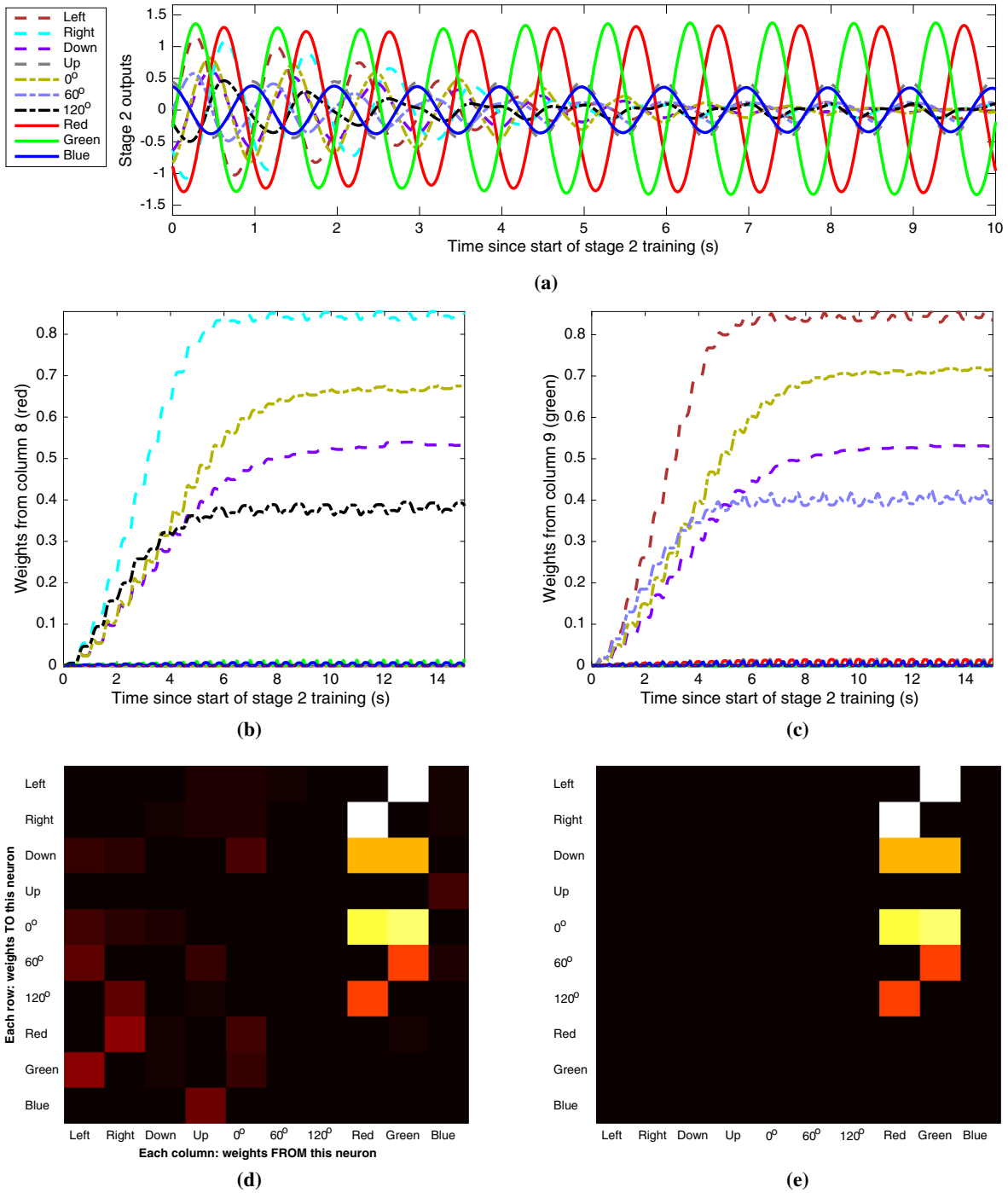

Fig. 8 Measures of the second-stage network, as it trained with a visual stimulus comprised of two bars moving through sinusoidal shadow. The legend at *top left* identifies traces throughout this and subsequent figures. **a** All ten network outputs over time. Training began at time zero. As training progressed, the *red* and *green* outputs remained largely unchanged, while all other outputs were inhibited. **b** Evolution of inhibitory weights from column 8 of the weight matrix $\boldsymbol{T}$ as the network trained, representing inhibition from the "red" neuron to all other neurons. During training, these weights grew and stabilized, learning to inhibit other neurons that had similar temporal fluctuations. **c** Evolution of inhibitory weights from column 9 of the weight matrix $\boldsymbol{T}$ as the network trained, representing inhibition from the "green" neuron to all other neurons. **d** Final state of the weight matrix $\boldsymbol{T}$ after 15 s of training. *Brighter colors* represent larger values, and *darker colors* smaller values (maximum value shown is 0.85). It is clear that the strongest weights are in columns 8 and 9. **e** The final weight matrix $\boldsymbol{T}$, after normalization to its maximum value and removal of weights less than 1/3 of the maximum. Here the patterns of inhibition are quite clear **a** Network outputs during learning **b** Inhibitory weights from neuron 8 (*red*) **c** Inhibitory weights from neuron 9 (*green*) **d** Raw weight matrix (**e**) Thresholded weight matrix (color figure online)

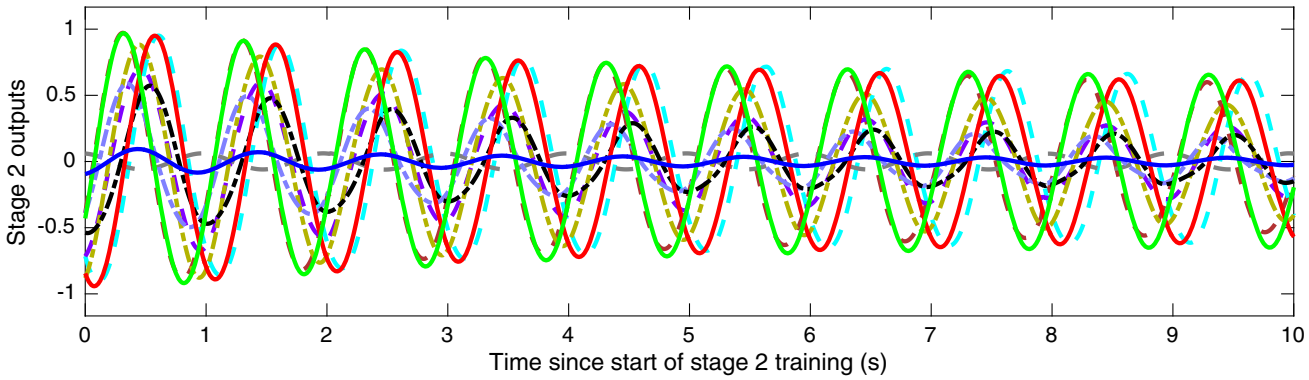

Fig. 9 Outputs of the second-stage network, as it trained with the same visual stimulus used in Fig. 8, but with the first-stage network only trained to a maximum eigenvalue of 0.1. Compared with Fig. 8a, while the network may be gradually learning the correct solution, its progress is much slowed by weak inhibition in the first stage

3.4.2 Varying the number of objects

To demonstrate that the second-stage connection matrix and the number of uninhibited outputs represent the number of unique objects in the visual input, we varied the number of bars in the stimulus of Fig. 8. Figure 10 shows a comparison of network outputs and final weight matrices with one, two, and three-bar visual stimuli.

Figure 10a and b show the results of removing the green bar from the reference stimulus, leaving only the red moving bar. The red output clearly dominates, and weights in the "red" column of the weight matrix correctly indicate an orientation between 0° and −60°, rightward motion, and a smaller component of downward motion. For comparison, Fig. 10c and d shows the corresponding data from the reference stimulus, shown in more detail in Fig. 8, and provide qualitatively the same data about the red moving bar. Figure 10e and f and shows the results of adding a blue bar (RGB = [0.75 0.1 0.1], moving directly to the left) to the reference stimulus for a total of three moving bars. Learning of this stimulus was slightly more difficult, but with no changes to parameters, in the end three distinct outputs came to dominate all others: those outputs corresponding to red, blue, and green color. The weight matrix in the red and green columns is qualitatively very similar to that for the two-bar reference stimulus, with the only significant difference being a missing representation of orientation 0° for the red bar; this visual feature was common to all three bars presented, and because the corresponding output was already inhibited by the green and blue neurons, no inhibition was learned from the red neuron. The weight matrix column corresponding to blue correctly shows a 0° orientation (equivalent to 180°) and motion to the left with no other component. Thus the number of bars in the visual stimulus is evident, along with the unique characteristics of each.

3.4.3 Varying the mechanism of fluctuations

All visual stimuli shown up to this point have used a multiplicative sinusoidal shadow pattern to generate common temporal fluctuations used to bind the characteristics of each bar together. This has made it easy to discern when the outputs have come to represent the hidden fluctuations, but one might reasonably ask if sinusoidal shadowing is required for network operation. To address this question, we have varied the method by which temporal fluctuations are generated, and the results of these experiments are shown in Fig. 11. For comparison purposes, Fig. 11a and b again shows the network outputs and final weight matrix for the reference stimulus.

Figure 11c and d shows network outputs and the weight matrix for the same pair of red and green moving bars as in the reference stimulus, but without any pattern of shadows at all. Rather, each bar oscillated in *distance* from the simulated camera (which by perspective projection changed its size in the image) at a frequency of 1 cycle per second, contracting from the reference width of 12 pixels at its initial distance to a minimum width of 9 pixels at its greatest distance, with a proportional change in length. This regular change in bar size caused a corresponding fluctuation in all visual submodalities, and the features of the moving bars are learned even more quickly by the network than while using sinusoidal shadowing. Despite some minor differences, Fig. 11d shows qualitatively the same pattern of weights as weight matrix of Fig. 11b learned from the reference stimulus. Relative to the reference stimulus, weights for this stimulus to the 60° and 120° orientations are somewhat stronger, and this is evident in Fig. 11c in the stronger inhibition of those outputs.

Figure 11e and f shows network outputs and the weight matrix for the same pair of red and green moving bars as in the reference stimulus, but in this instance the bars were overlaid with a randomly generated multiplicative shadow pattern.

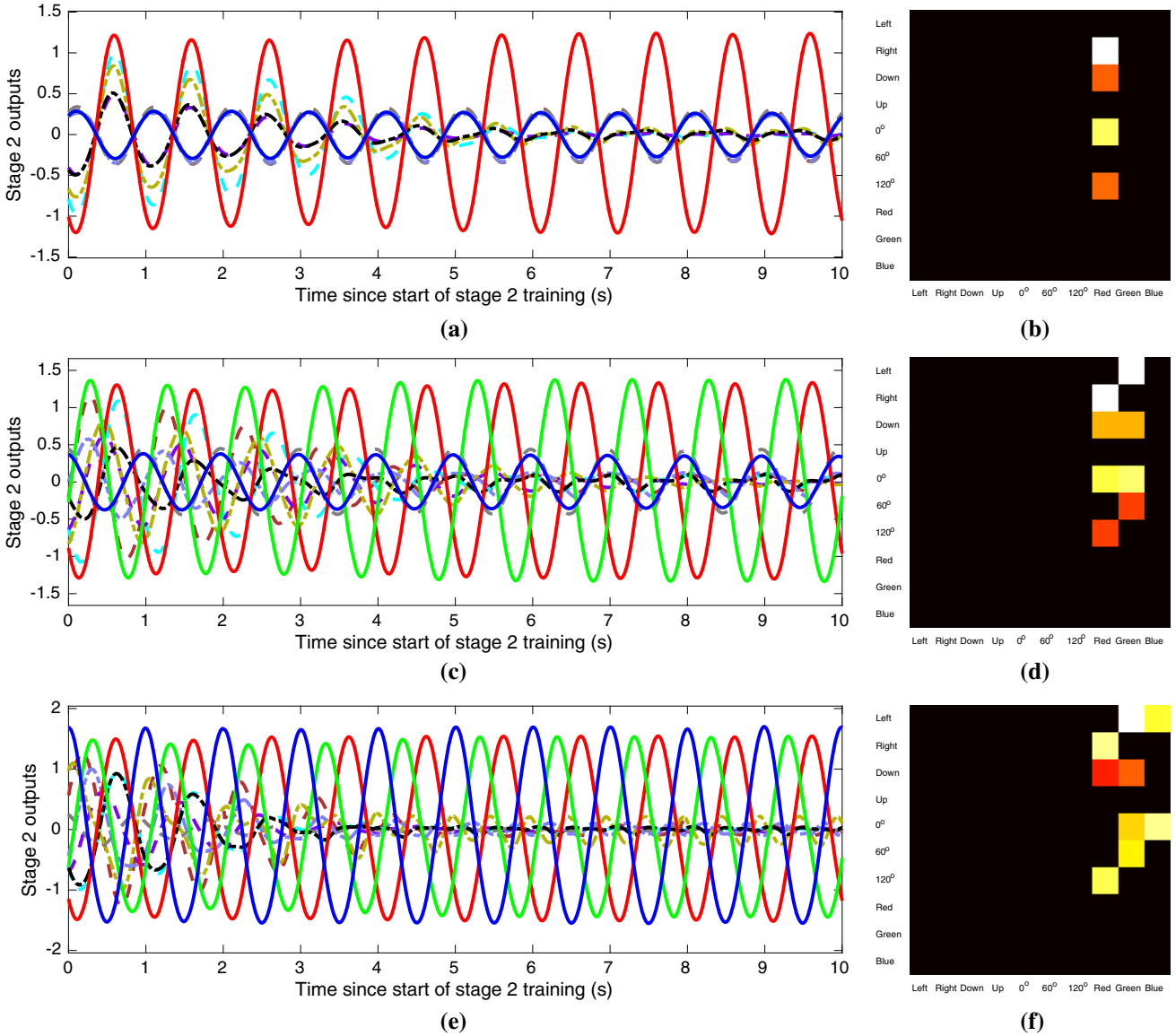

Fig. 10 Second-stage outputs and weight matrices as the number of *bars* in the visual stimulus was varied (*top to bottom*), with all other stimulus parameters held constant. The *left column* (**a**, **c**, and **e**) shows all ten network outputs as they developed over time. Refer to the *upper left corner* of Fig. 8 for a legend to identify each trace. The right column (**b**, **d**, and **f**) shows the thresholded weight matrices at the end of training. In all three cases, both network outputs and weight matrices learn to correctly represent the visual stimulus **a** Network outputs for one-bar stimulus **b** One-bar weight matrix **c** Network outputs for two-bar stimulus **d** Two-bar weight matrix **e** Network outputs for three-bar stimulus **f** Three-bar weight matrix

Prior to the beginning of the simulation, a 500×500 matrix of uniformly distributed random numbers was generated and then convolved twice with a circular 2D Gaussian spatial low-pass filter with standard deviation $\sigma_n = 6$ pixels. The resulting dappled unoriented shadow pattern was then scaled and offset so that, like the sinusoidal shadow patterns, it had a minimum value of 0.25 and a maximum of 0.75.

The subtle, low-amplitude random temporal fluctuations caused by the random shadowing made the binding problem more difficult to solve, and it was necessary to increase the learning rate to γ_2 to 4 from its standard value of 0.5. However, after training for the same 15-second duration used for the other stimuli, the red and green outputs had virtually suppressed all others as shown in Fig. 11e, and the network had reached a final connection matrix state, shown in Fig. 11f, which is qualitatively identical to the weights learned from the reference stimulus shown in Fig. 11b.

Taken as a whole, the results of Fig. 11 show that neither sinusoidal fluctuations nor even shadowing are required for meaningful second-stage visual binding network operation. Rather, the network learns based on temporal fluctuations of any kind that may be available.

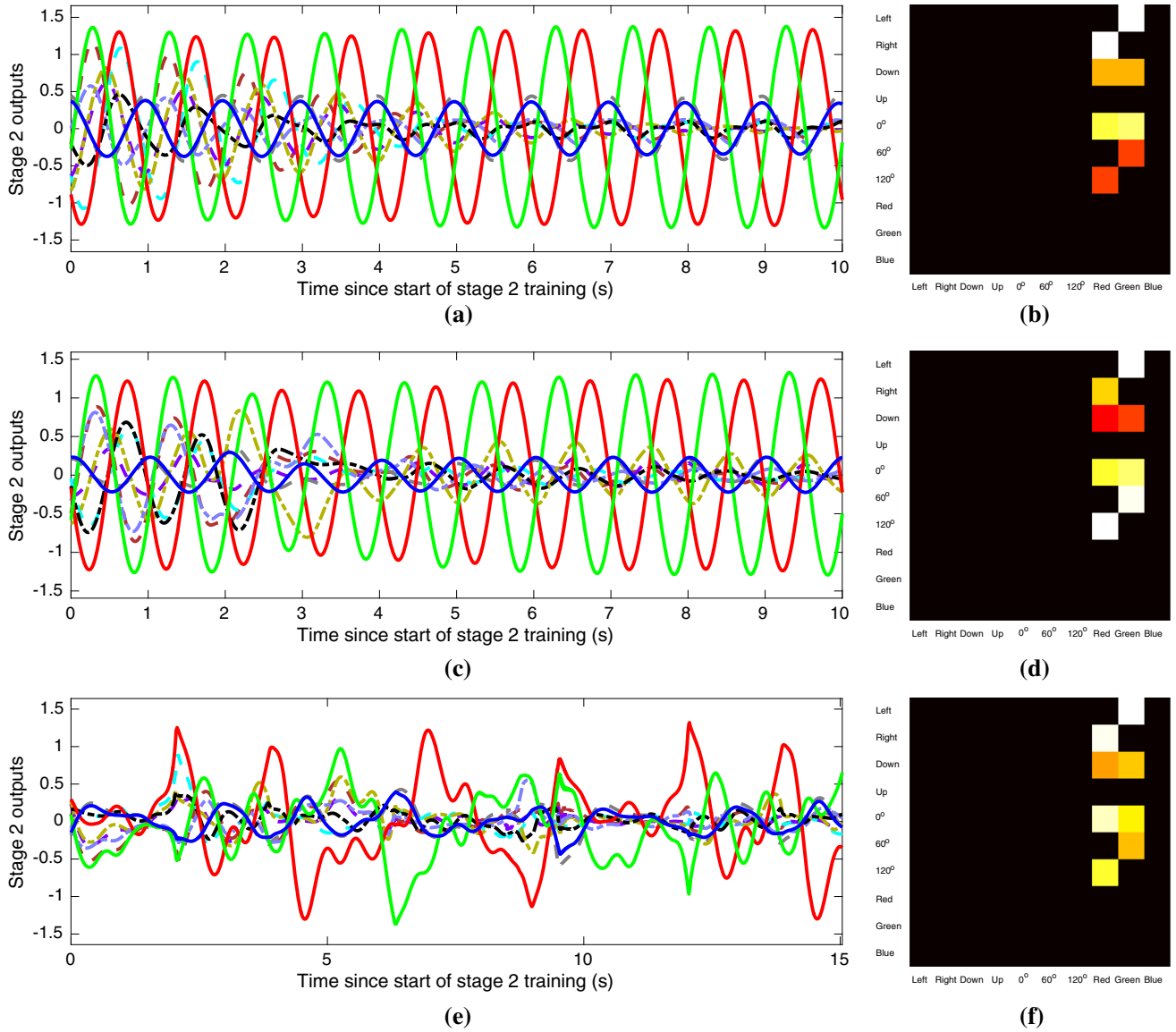

Fig. 11 Second-stage outputs and weight matrices for a two-bar visual stimulus as the manner of generating temporal fluctuations was varied, with all other stimulus parameters held constant. The *left column* of panels shows all ten network outputs as they developed over time. Refer to the *upper left corner* of Fig. 8 for a legend to identify each trace. The *right column* of panels shows the thresholded weight matrices at 15 s, the time at which training was concluded for each experiment. The *top row* (**a** and **b**) is data from the reference stimulus, which used sinusoidal shadowing. In the second row (**c** and **d**), no shadowing was used, but rather the distance of the bars from the simulated camera (and thus by perspective projection their size in the image) was varied over time. In the *bottom row* (**e** and **f**), a pattern of multiplicative random shadow was used. In this case, network outputs are shown for 15 s due to the increased complexity of the stimulus. However, in all three cases, the final weight matrix develops a very similar representation of the visual stimulus **a** Network outputs with sinusoidal shadow **b** Sine shadow weight matrix **c** Network outputs with distance variation **d** Distance variation weight matrix **e** Network outputs with random shadow **f** Random shadow weight matrix

3.4.4 Visual binding with real-world video

Given the infinite number of possible visual stimuli and the limited space of any publication, we conclude our experimental evaluation of the model by using a sequence of images captured from a video camera. Here we take the opportunity not only to show that the system works with a natural visual stimulus, but also to demonstrate yet another manner by which temporal fluctuations may be generated for use in learning by the visual binding model: appearance and disappearance of an object.

Figure 12 shows the results of the network when trained with a video of a red car passing at moderate speed horizontally from right to left through a brightly sunlit parking lot.

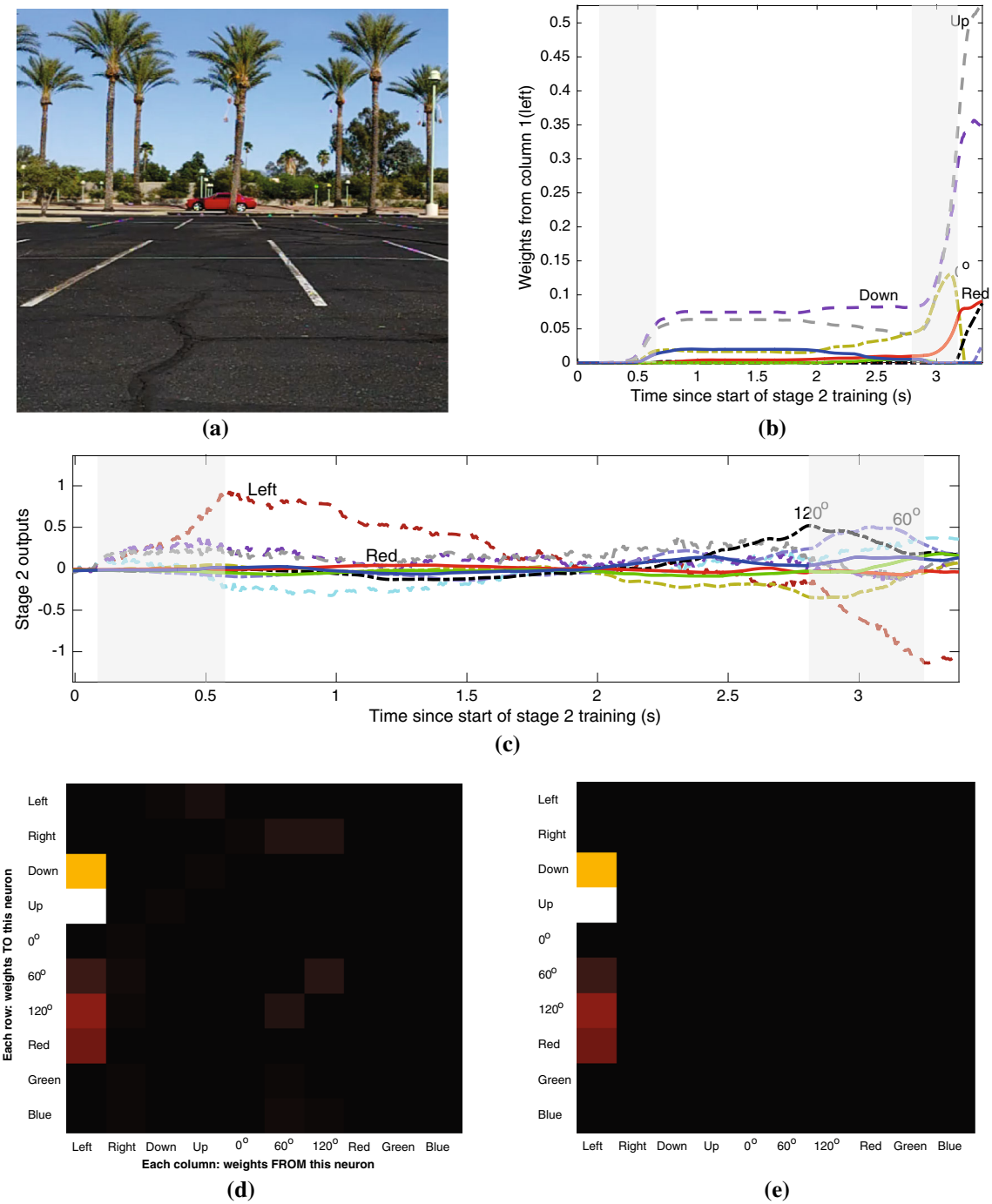

Fig. 12 Measures of the second-stage network as it trained with a 120 FPS video of a *red car* moving *right* to *left* through the scene. **a** A sample 500 × 500 video frame at 1.8 s after the beginning of second-stage training. **b** Time evolution of inhibitory weights from column 1 of the weight matrix $\boldsymbol{T}$ as the network trained, representing inhibition from the "left" neuron to all other neurons. Translucent *gray boxes* in this panel and the next indicate when the car was entering the frame from approximately 0.1–0.6 s after the start of training, and when the car was leaving the frame at approximately 2.8–3.2 s. Note that most of the changes in connection weights occurred as the car entered and left the scene. **c** Network outputs over the 3.4 s duration of training with this stimulus. A positive leftward motion output clearly dominates early in training, and becomes the largest negative output as the car leaves. **d** Final state of the weight matrix $\boldsymbol{T}$ after 3.4 s of training. *Brighter colors* represent larger values, and *darker colors* smaller values. The maximum value in this matrix is 0.53. **e** The weight matrix $\boldsymbol{T}$ after normalization to its maximum value and removal of weights less than 10% of the maximum. Only one column has nonzero weights, representing a single object **a** Sample video frame **b** Inhibitory weights from neuron 1 (*left*) **c** Network outputs during learning **d** Raw weight matrix **e** Thresholded weight matrix (color figure online)

The video was taken with 500×500 frames at 120 frames per second (FPS) to most closely match our artificially generated stimuli, all of which were at same image size but generated at 100 FPS. To better accommodate the higher temporal frequencies in this video, the input high-pass filter time constant τ_{HI} was increased from 1.0 to 1.5 s. Similarly, the output high-pass filter time constant τ_{HO} was increased from 0.5 to 0.75 s (thus maintaining the same ratio of the two time constants as used for previous experiments). The learning rate γ_2 was increased to 10 in order to learn more quickly.

Although the red car goes behind occluding palm trees as well as their shadows during the video, its appearance and disappearance in the visual scene are by far the strongest cues. Unlike our artificial stimuli, this video was not looped, but of fixed duration. The video began with 5 s of the parking lot with no movement other than that of the background (which included palm tree movement due to wind, minor camera movements, and minor overall brightness adjustments by the camera), during which time the visual binding network was allowed to adapt to the visual input. During the following 3.4 s of video, the red car passed completely across the scene, entering and leaving the scene in approximately 3 s; this was the only opportunity for the network to gather information about the object.

Figure 12a shows an example frame from this video. Note that the car is not only behind a palm tree, but also in its shadow. Figure 12b shows how network weights from the "left" column developed over time, primarily changing during appearance and disappearance of the car. Figure 12c shows all network outputs, with the "left" neuron generating the largest positive output as the car entered the scene and the largest negative output as the car left. The car was the only consistently moving object in the scene, and so its motion created a strong output. In contrast, there were a huge variety of orientations and colors already present in the background, and the car covered very few pixels relative to the image size, and thus generated weak orientation and color responses. Figure 12d and e respectively shows the raw and normalized connection weight matrices, revealing that the network has associated the leftward motion output strongly with both upward and downward motion, weakly with orientations of 60° and 120°, and weakly with the color red. The strong weights to upward and downward motion were generated primarily during exit of the car from the scene. Both upward and downward motion signals were relatively weak as the car passed through the frame, and resulted from the aperture problem. However, both signals decreased simultaneously with the strong leftward motion component, leading to their association. The nearly vertical orientation learned by the network corresponds to strong vertical components in the windows and edges of the car.

4 Discussion

We have presented a novel neural network model based on an initial hypothesis of the computations that may be performed in insect optic glomeruli (Strausfeld and Okamura 2007), a newly discovered visual processing area just beyond the optic lobes in insects. This model merges and extends prior work by Hopfield (1991) on modeling of olfactory glomeruli (which anatomically resemble optic glomeruli) and by Herault and Jutten (1986) on blind source separation. The basic function of this model is to create a non-spatial representation of objects based a wide-field mixture of their time-varying visual features. This representation implicitly allows a determination of how many objects are present in a visual image sequence, and identifies—in the form of an inhibitory connection matrix—the unique visual features of each object based on common temporal fluctuations.

The present model is organized into two stages containing four individual recurrent networks, three of which use lateral inhibition to refine inputs from a single visual submodality (motion, orientation, and color) and together comprise the first stage of visual processing, and the last of which combines refined inputs across all visual submodalities to perform visual binding.

We have demonstrated that the first-stage networks refine the representation of each submodality individually, that this refinement has some subtle side effects (in particular, we showed that refinement of visual motion provides a partial solution to the aperture problem), and that first-stage processing greatly enhances second-stage learning. The reduction in redundant information provided by each network—often interpreted as information maximization—has been proposed as a possible goal of all neural computation (Barlow 2001).

We have shown that the second-stage network is capable of learning the number of objects in an image sequence and identifying their individual characteristics using controlled artificially generated visual stimuli composed of moving bars, verified that network function is maintained as the number of bars is varied, and that network function is not dependent on any particular method of generating temporal fluctuations. Finally, we have demonstrated successful performance of the visual binding network on a sequence of real-world images.

The functional limits of this model in representing concurrently presented objects is related to existing literature on the limits of blind source separation models, and we explore these limits in detail in a companion paper (Northcutt and Higgins 2016), where we also address the consequences of our alterations to the temporal evolution equation and learning rule relative to previous work on blind source separation.

Perhaps the most interesting aspect of the current model is that the three first-stage networks, which have been characterized as performing sensory refinement, have identical temporal evolution and learning rules to the second-stage network that performs the apparently dissimilar task of visual binding. The common function of all four networks is to "orthogonalize" inputs that have significant overlap, thus reducing the ambiguity of the inputs. This computation also makes network outputs more robust to the detailed selectivity of the inputs: For example, the output of the orientation refinement network would be little changed if the input orientation filters grew moderately more or less selective.

The present model is comprised of only four networks, each of which is hypothesized to represent a single optic glomerulus. This number was arrived at by using three visual submodalities, and providing to each first-stage network a vector of inputs created by a full-field spatial sum of all local detectors for that submodality. While it is fascinating that the network can learn a high-level representation of objects in the image even after having completely thrown away all spatial information, given that optic glomeruli number more than two dozen in blowflies (Okamura and Strausfeld 2007) it is more likely that inputs to each glomerulus are not full-field spatial sums, but rather are integrated over a number of large, distinct spatial receptive fields so that not all retinotopic information is discarded. Such a model could easily incorporate dozens of glomeruli, some of which would refine wide-field inputs from different submodalities, and others of which would combine these refined inputs across submodalities to provide object-level information about each local region of the visual field to higher-level visual processing areas, maintaining a coarse retinotopy.

Our visual binding network makes use of subtractive inhibition, which makes it analytically tractable and ties it to the well-known literature on blind source separation. However, it should be noted that more biophysically realistic divisive inhibition methods have been proposed in color, orientation and motion models which have been shown to provide self-normalization of signals, improve coding efficiency, and compensate for nonlinearity of input signals (Schwartz and Simoncelli 2001; Simoncelli and Olshausen 2001). Divisive normalization has been proposed as a canonical neural computation (Carandini and Heeger 2012) and such neural circuitry could be key to adaptation and normalization. Divisive inhibition is an alternative model of inhibition that should be explored in our recurrent inhibitory networks.

Despite distinct differences in network structure and learning rules, the present model is related to many neural network models of visual binding and attention (Eckhorn et al. 1990; Engel et al. 1992; Schillen and König 1994; Itti et al. 1998), and even models of consciousness (Crick and Koch 1990; Engel and Singer 2001) in that these models all make use of temporal correlations of elementary features to solve the binding problem. Many neural network models have been proposed (Hummel and Biederman 1992; von der Malsburg 1994), which make use of temporal synchrony of neuronal firings to represent the binding of visual features. While this mechanism is unlikely to be used in the insect optic lobes where spiking neurons are relatively rare, support for the idea of neuronal spike synchrony as a representation for visual binding in mammalian brains has gathered increasing biological evidence in recent years (Martin and von der Heydt 2015).

The notion that there may exist a canonical neuronal circuit which is repeated across many sensory modalities is an attractive one, and seems quite plausible in the context of the present model. Given the strong anatomical resemblance between olfactory and optic glomeruli, and the close relationship of our model of optic glomeruli to models of olfaction (Hopfield 1991)—and more generally to blind source separation (Herault and Jutten 1986)—the recurrent inhibitory neural network, which exhibits lateral inhibition in its simplest form, could well be one such canonical neuronal circuit. It has been demonstrated in a large number of sensory modalities that this type of network is useful in sensory refinement, and the present work extends prior work on olfactory visual binding to include vision as well. Whether such a neuronal circuit is used in similar ways in other sensory modalities remains to be seen, but the present results definitively indicate that neuronal organizations based around the "simple" recurrent inhibitory network, in the presence of appropriate learning rules, can give rise to surprisingly high-level implicit representations of sensory information.

Acknowledgements The authors would like to thank the Air Force Office of Scientific Research for early support of this project with Grant Number FA9550-07-1-0165, and the Air Force Research Laboratories for supporting this research to maturity with STTR Phase I Award Number FA8651-13-M-0085 and Phase II Award Number FA8651-14-C-0108, both in collaboration with Spectral Imaging Laboratory (Pasadena, CA). We would also like to thank the reviewers, whose input greatly enhanced this manuscript.

References

Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. J Opt Soc Am A 2:284–299

Albrecht DG, Geisler WS (1991) Motion selectivity and the contrast-response function of simple cells in the visual cortex. Vis Neurosci 7(6):531–546

Anderson JA (1995) An introduction to neural networks. MIT Press, Cambridge

Arnett DW (1972) Spatial and temporal integration properties of units in first optic ganglion of dipterans. J Neurophysiol 35(4):429–444

Barlow HB (2001) Redundancy reduction revisited. Netw Comp Neural 12(3):241–253

Bazhenov M, Stopfer M, Rabinovich M, Abarbanel HD, Sejnowski TJ, Laurent G (2001) Model of cellular and network mechanisms

for odor-evoked temporal patterning in the locust antennal lobe. Neuron 30(2):569–581
Bi GQ, Poo MM (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci 18(24):10,464–10,472
Borst A, Egelhaaf M (1989) Principles of visual motion detection. Trends Neurosci 12(8):297–306
Carandini M, Heeger DJ (2012) Normalization as a canonical neural computation. Nat Rev Neurosci 13(1):51–62
Cichocki A, Bogner RE, Moszczyński L, Pope K (1997) Modified Herault–Jutten algorithms for blind separation of sources. Digit Signal Process 7(2):80–93
Crick F, Koch C (1990) Towards a neurobiological theory of consciousness. Semin Neurosci 2:263–275
Douglass JK, Strausfeld NJ (2005) Sign-conserving amacrine neurons in the fly's external plexiform layer. Vis Neurosci 22(03):345–358
Eckhorn R, Reitboeck HJ, Arndt M, Dicke P (1990) Feature linking via synchronization among distributed assemblies: simulations of results from cat visual cortex. Neural Comput 2(3):293–307
Engel AK, Singer W (2001) Temporal binding and the neural correlates of sensory awareness. Trends Cognit Sci 5(1):16–25
Engel AK, König P, Kreiter AK, Schillen TB, Singer W (1992) Temporal coding in the visual cortex: new vistas on integration in the nervous system. Trends Neurosci 15(6):218–226
Fonta C, Sun XJ, Masson C (1993) Morphology and spatial distribution of bee antennal lobe interneurones responsive to odours. Chem Senses 18(2):101–119
Gerstner W, Kistler WM (2002) Mathematical formulations of Hebbian learning. Biol Cybern 87(5–6):404–415
Gerstner W, Kempter R, van Hemmen JL, Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. Nature 383(6595):76
Hassenstein B, Reichardt W (1956) Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers Chlorophanus. Z Naturforsch B 11(9–10):513–524
Hebb DO (1949) The organization of behavior: a neuropsychological theory. Wiley, New York
Heisenberg M (2003) Mushroom body memoir: From maps to models. Nat Rev Neurosci 4(4):266–275
Herault J, Jutten C (1986) Space or time adaptive signal processing by neural network models. In: AIP Conference Proceedings, Snowbird, vol 151. American Institute of Physics, pp 206–211
Hildebrand JG (1996) Olfactory control of behavior in moths: central processing of odor information and the functional significance of olfactory glomeruli. J Comp Physiol A 178(1):5–19
Hildebrand JG, Shepherd GM (1997) Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. Annu Rev Neurosci 20(1):595–631
Hopfield JJ (1991) Olfactory computation and object perception. Proc Natl Acad Sci USA 88(15):6462–6466
Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148(3):574–591
Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 195(1):215–243
Hummel JE, Biederman I (1992) Dynamic binding in a neural network for shape recognition. Psychol Rev 99(3):480–517
Hyvärinen A, Oja E (1998) Independent component analysis by general nonlinear Hebbian-like learning rules. Signal Process 64(3):301–313
Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal 11:1254–1259
Jefferis GSXE (2005) Insect olfaction: a map of smell in the brain. Curr Biol 15(17):R668–R670
Joho M, Mathis H, Lambert RH (2000) Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture. In: Proc. International Conference on Independent Component Analysis and Blind Signal Separation. Helsinki, pp 81–86
Jutten C, Herault J (1991) Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. Signal Process 24(1):1–10
Koch C (1999) Biophysics of computation: information processing in single neurons. Oxford University Press, New York
Land MF, Nilsson DE (2002) Animal eyes. Oxford University Press, New York
Laughlin S (1983) Matching coding to scenes to enhance efficiency. Proceedings of an International Symposium Organized by The Rank Prize Funds, Springer Series in Information Sciences. Springer, London, pp 42–52
Linster C, Masson C (1996) A neural model of olfactory sensory memory in the honeybee's antennal lobe. Neural Comput 8(1):94–114
Linster C, Smith BH (1997) A computational model of the response of honey bee antennal lobe circuitry to odor mixtures: overshadowing, blocking and unblocking can arise from lateral inhibition. Behav Brain Res 87(1):1–14
von der Malsburg C (1994) The correlation theory of brain function. Springer, New York
von der Malsburg C (1999) The what and why of binding: the modeler's perspective. Neuron 24(1):95–104
Markram H, Lübke J, Frotscher M, Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275(5297):213–215
Martin AB, von der Heydt R (2015) Spike synchrony reveals emergence of proto-objects in visual cortex. J Neurosci 35(17):6860–6870
Mu L, Ito K, Bacon JP, Strausfeld NJ (2012) Optic glomeruli and their inputs in Drosophila share an organizational ground pattern with the antennal lobes. J Neurosci 32(18):6061–6071
Nakayama K, Silverman GH (1988) The aperture problem—I. Perception of nonrigidity and motion direction in translating sinusoidal lines. Vis Res 28(6):739–746
Ng M, Roorda RD, Lima SQ, Zemelman BV, Morcillo P, Miesenböck G (2002) Transmission of olfactory information between three populations of neurons in the antennal lobe of the fly. Neuron 36(3):463–474
Northcutt BD, Higgins CM (2017) An insect-inspired model for visual binding II: Functional analysis and visual attention. Biol Cybern. doi:10.1007/s00422-017-0716-z
Okamura JY, Strausfeld NJ (2007) Visual system of calliphorid flies: motion- and orientation-sensitive visual interneurons supplying dorsal optic glomeruli. J Comp Neurol 500(1):189–208
Paulk AC, Dacks AM, Phillips-Portillo J, Fellous JM, Gronenberg W (2009) Visual processing in the central bee brain. J Neurosci 29(32):9987–9999
Rivera-Alvidrez Z, Lin I, Higgins CM (2011) A neuronally based model of contrast gain adaptation in fly motion vision. Vis Neurosci 28(5):419–431
Rodieck RW (1965) Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vis Res 5(12):583–601
van Santen JPH, Sperling G (1985) Elaborated Reichardt detectors. J Opt Soc Am A 2(5):300–320
Schillen TB, König P (1994) Binding by temporal structure in multiple feature domains of an oscillatory neuronal network. Biol Cybern 70(5):397–405
Schwartz O, Simoncelli EP (2001) Natural signal statistics and sensory gain control. Nat Neurosci 4(8):819–825
Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. Annu Rev Neurosci 24(1):1193–1216
Snyder AW (1979) Physics of vision in compound eyes. In: Autrum H (ed) Handbook of sensory physiology, vol VII/6A. Springer, Berlin, pp 225–313 (chap 5)

Song S, Miller KD, Abbott LF (2000) Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. Nat Neurosci 3(9):919–926

Srinivasan MV, Zhang SW, Witney K (1994) Visual discrimination of pattern orientation by honeybees: performance and implications for 'cortical' processing. Philos Trans R Soc B 343(1304):199–210

Stavenga DG (1979) Pseudopupils of compound eyes. In: Autrum H (ed) Handbook of sensory physiology, vol VII/6A. Springer, Berlin, pp 357–439 (chap 7)

Strausfeld NJ, Okamura JY (2007) Visual system of calliphorid flies: organization of optic glomeruli and their lobula complex efferents. J Comp Neurol 500(1):166–188

Strausfeld NJ, Sinakevitch I, Okamura JY (2007) Organization of local interneurons in optic glomeruli of the dipterous visual system and comparisons with the antennal lobes. Dev Neurobiol 67(10):1267–1288

Trentelman H, Stoorvogel AA, Hautus M (2012) Control theory for linear systems. Springer Science & Business Media, Berlin

Yang EC, Maddess T (1997) Orientation-sensitive neurons in the brain of the honey bee (Apis mellifera). J Insect Physiol 43(4):329–336