Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Vision and Image Understanding 100 (2005) 3–40

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

# Attending to visual motion

John K. Tsotsos [a,b,*], Yueju Liu [a,b], Julio C. Martinez-Trujillo [c], Marc Pomplun [d], Evgueni Simine [a,b], Kunhao Zhou [a,b]

[a] *Department of Computer Science and Engineering, York University, Toronto, Canada*
[b] *Centre for Vision Research, York University, Toronto, Ont., Canada M3J 1P3*
[c] *Department of Physiology, McGill University, Montreal, Canada*
[d] *Department of Computer Science, University of Massachusetts, Boston, MA, USA*

Received 6 February 2004; accepted 5 October 2004
Available online 26 July 2005

## Abstract

Visual motion analysis has focused on decomposing image sequences into their component features. There has been little success at re-combining those features into moving objects. Here, a novel model of attentive visual motion processing is presented that addresses both decomposition of the signal into constituent features as well as the re-combination, or binding, of those features into wholes. A new feed-forward motion-processing pyramid is presented motivated by the neurobiology of primate motion processes. On this structure the Selective Tuning (ST) model for visual attention is demonstrated. There are three main contributions: (1) a new feed-forward motion processing hierarchy, the first to include a multi-level decomposition with local spatial derivatives of velocity; (2) examples of how ST operates on this hierarchy to attend to motion and to localize and label motion patterns; and (3) a new solution to the feature binding problem sufficient for grouping motion features into coherent object motion. Binding is accomplished using a top-down selection mechanism that does not depend on a single location-based saliency representation.

*Keywords:* Attention; Visual motion analysis; Feature binding; Selective tuning; Affine motion

[*] Corresponding author. Fax: +1 416 736 5857.
*E-mail address:* tsotsos@cs.yorku.ca (J.K. Tsotsos).
*URL:* http://www.cs.yorku.ca/~tsotsos (J.K. Tsotsos).

1077-3142/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2004.10.011

## 1. Introduction

The Selective Tuning model is a proposal for the explanation at the computational and behavioral levels of visual attention in humans and primates. Key characteristics of the model, all previously detailed in [1,2] include: (1) a top-down coarse-to-fine winner-take-all (WTA) selection process, (2) a unique WTA formulation with provable convergence properties, (3) a WTA that is based on region rather than point selection, (4) a task-relevant inhibitory bias mechanism, (5) selective inhibition in both spatial and feature dimensions for elimination of signal interference that leads to a suppressive surround for attended items, and (6) a task-specific executive controller. These characteristics lead to an extensive set of biological predictions many of which have now been supported by experiment. The bulk of the paper will focus on attention to visual motion. Past work will be summarized showing how this is not a well-studied issue. A new model of motion processing is presented and it is demonstrated how ST operates on this representation, with no changes to its previously described definition. In this way three points are made: first, that the weaknesses of previous demonstrations of ST have been remedied; second, that the original statement of ST has generality for a wide variety of visual processing representations; and third, examples of how feature binding can be solved using ST for complex motion patterns.

It had been suggested that previous demonstrations of the Selective Tuning model were neither biologically plausible nor very useful. In order to demonstrate that ST can indeed operate with realistic representations, the motion domain is chosen because enough is known about motion processing to enable a reasonable attempt at defining the feed-forward pyramid. Moreover, the effort is unique because it seems that no past model has presented a motion hierarchy plus attention to motion [3–12].

The layout of the remainder of this presentation is as follows. The next section will detail the feed-forward motion-processing network. Earlier versions of this network appear in [13,14]. Following this, an overview of ST is provided because this structure is imposed upon the feed-forward network. ST has been detailed several times in the past; here only a brief presentation is given and the reader is referred to [1,15,2,16–18] for further details. Section 4 will show several examples of the operation of the entire network including feed-forward and feedback components as well as a new solution to the feature-binding problem. A concluding discussion rounds out the paper.

## 2. Feed-forward motion processes

The motion representations and processes that are modeled are informed by current knowledge of motion analysis in the primate cortex. Although the literature is large on the topic, selected experimental observations are used here in order to simplify the models. It is generally accepted that motion processing in the monkey cortex goes through a series of stages, with neural representations in areas V1, MT, MST, and 7a each providing input for the next [19]. Each of the areas specializes in particular kinds of motions, that is, contains populations of neurons specialized for certain motion features, generally from simple to more complex and with smaller to larger receptive fields higher up in the hierarchy. These different neural properties will be outlined throughout this section, with one sub-section devoted to each area.

The model aims to explain how a hierarchical feed-forward network consisting of multiple neural populations in the cortical areas V1, MT, MST, and 7a of primates detects and represents different kinds of motion patterns. At best, it is a first-order motion model with much elaboration left for future work. Indeed, some previous motion models cited earlier offer better sophistication at one or another level of processing; however, none cover all these levels and incorporate selective attention processes.

### *2.1. The feed-forward motion pyramid*

The first component of the motion analysis process is a feed-forward (data-driven) one. The goal is to define a set of processing stages for areas V1, MT, MST, and 7a, corresponding to the areas of the motion processing hierarchy in macaque monkey [19], that conform to the basic properties observed in neural populations in those areas [20–29]. A very brief characterization of the processing levels follows:

- Cells in *striate area V1* are selective for a particular local speed and direction of motion in at least three main speed ranges.
- Cells in *area MT* are of two kinds. One kind is tuned for a particular local speed and direction of movement, similar to direction and speed selective cells in V1 but with larger receptive fields. The second kind is selective for a particular angle between local movement direction and spatial velocity gradient.
- Cells in *area MST* are tuned to complex motion patterns: expand or approach, contract or recede, clockwise or counter-clockwise rotation, and combinations of these, and translation but with even larger receptive fields.
- Cells in *area 7a* code four different types of patterns: translation and spiral motion as in MST, full field rotation (regardless of direction), and radial motion (expansion or contraction), within the largest receptive fields.

There is no claim that these are necessarily the only neural populations within each area in primates; these are simply the only ones modeled here.

The model includes neurons in the areas V1 (referred to as $\mathrm{VF}_{\alpha,v,i}$ and $\mathrm{VI}_{\alpha,v,i}$), MT (referred to as $\mathrm{MT}_{\alpha,v,i}$ and $\mathrm{MG}_{\alpha,\delta,v,i}$), MST (referred to as $\mathrm{ST}_{\alpha,v,i}$ and $\mathrm{SS}_{\delta,v,i}$), and 7a (referred to as $\mathrm{AT}_{\alpha,v,i}$, $\mathrm{AS}_{\delta,v,i}$, $\mathrm{ART}_{v,i}$, and $\mathrm{ARD}_{v,i}$). A number of parameters and numerical constraints have been set with guidance from the literature available or by reasonable estimations otherwise.

Fig. 1 depicts the full motion hierarchy. This figure emphasizes the scale of the search problem faced by the visual system: to determine which responses within each of these representations belong to the same event. Each layer is now described in turn.

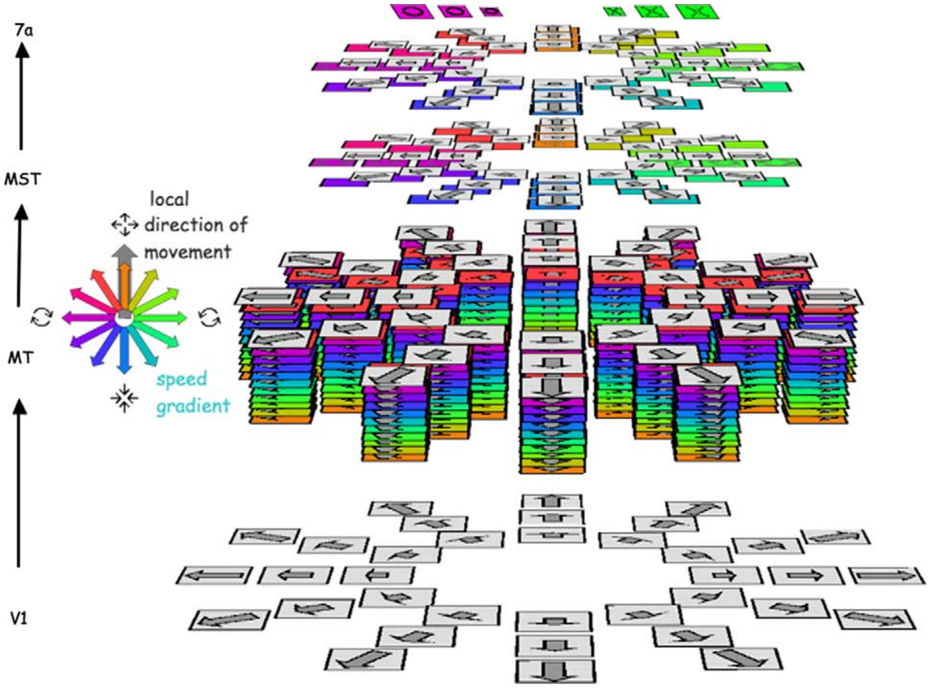

Fig. 1. The full motion hierarchy. This shows the set of neural selectivities that comprise the entire pyramidal hierarchy covering visual areas V1, MT, MST, and 7a. Each rectangle represents a single type of selectivity applied over the full image at that level of the pyramid. Large grey arrows represent selectivity for direction. Coloured rectangles in area MT represent particular angles between motion and speed gradient. The three rectangles at each direction represent the three speed selectivity ranges in the model. In this way, each single 'sheet' may be considered an expanded view of the 'hypercolumns' in a visual area. In area V1, for example, the neurons that integrate direction and speed selectivity are represented by the single sheet of grey rectangles. In area MT, there are 13 sheets, the top one representing direction and speed selectivity while the remaining 12 represent the 12 directions of velocity gradient relative to the 12 motion directions. The wheel of coloured arrows represents the colour coding within area MT for speed gradient with respect to local motion, in this case the larger grey arrow pointing upwards. This codes the angle between local motion and speed gradient. MST units respond to patterns of motion—contract, recede, and rotate. The 7a layers represent translational motion, spiral motion, both as in area MST, plus radial and rotation without direction in the topmost set of six rectangles.

### 2.2. Area V1

Area V1 receives visual input as a temporal sequence of images. Spatiotemporal filters are used to model the selectivity of V1 neurons for speed and direction of local motion (see [21]). Our first attempt at this employed the spatiotemporal filter approach of Heeger [30]; however, for the number and resolution of images in a sequence, the output of these filters was too noisy for the subsequent velocity gradient computation in area MT. Consequently, another computational mechanism for V1 was defined that generates a more appropriate input to the MT neurons. This mechanism is presented in the following paragraphs.

The functionality of layer V1 is realized by two types of artificial neurons, namely those performing spatiotemporal filtering, referred to as $\mathrm{VF}_{\alpha,v,i}$, and those integrating local filter unit activations, referred to as $\mathrm{VI}_{\alpha,v,j}$. The filter units have spatiotemporal RFs that provide access to the intensity values of the $T$ images in the most recent sub-sequence. In the present evaluation of the model, $T = 5$. The intensity value at position $p$ in the image taken at time $t$ is $I(p, t)$, where $t = 1$ indicates the first and $t = T$ indicates the most recent image in the sequence.

In this input space, the RF of a neuron $\mathrm{VF}_{\alpha,v,i}$ is oriented in such a way that local motion at its position $i$ in direction $\alpha$ and with speed $s_v$ would induce constant intensity across the RF. V1 consists of neurons of three distinct speed selectivity types (following [21]): type 1 (low speed), type 2 (medium speed), and type 3 (high speed). In the model, these neurons are implemented with three different preferred speeds, which were set to $s_1 = 0.5$ pixels/frame, $s_2 = 1$ pixel/frame, and $s_3 = 2$ pixels/frame. To limit the computational complexity of the model, only 12 different preferred directions were realized ($\alpha = 0°, 30°, \ldots, 330°$), although it is known that a wider range of preferred directions exist in area V1 of macaques [21]. For a preferred direction $\alpha$ and a preferred speed $s_v$ (in pixels per frame), the spatial offset $\Omega(\alpha, s_v, t)$ of the line of constant intensity in the image taken at time $t$ can be computed as follows:

$$\Omega(\alpha, s_v, t) = s_v\left(t - \frac{T+1}{2}\right)\begin{pmatrix}\cos\alpha\\ \sin\alpha\end{pmatrix}. \tag{1}$$

Since in most cases this function will not yield integer values, up to four inputs per image are used to estimate actual intensity in defining the line of constant intensity. This can be visualized for a sequence of one-dimensional images (see Fig. 2), where two inputs per image can be necessary.

Here, the darker a pixel, the larger its weight in the computation of the input from its image (because the line is passing through it more centrally). A neuron $\mathrm{VF}_{\alpha,v,i}$

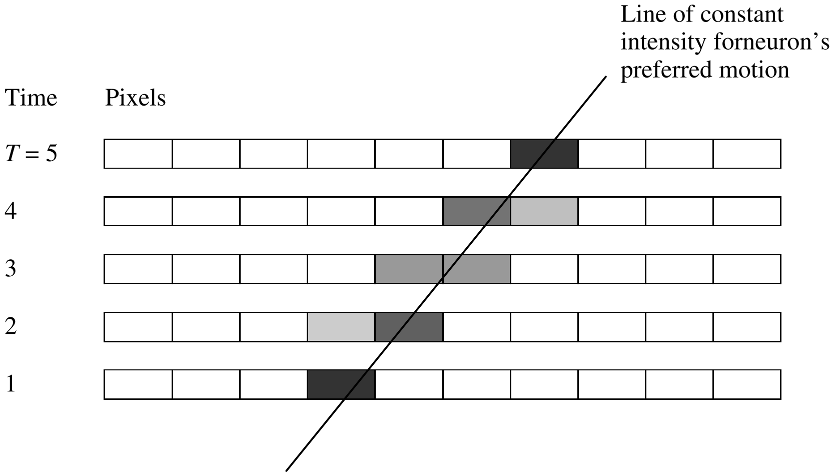

Fig. 2. Illustration of spatiotemporal energy computation.

receives input from each temporal layer and computes an intensity constancy value IC that decreases with increasing standard deviation of the intensity across the layers. In the following quantitative description the real-valued function $\Omega(\alpha, s_v, t)$ is used:

$$\mathrm{IC} = M_{\mathrm{VF}} - \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(I(p(i) + \Omega(\alpha, s_v, t), t) - \frac{\sum_{t=1}^{T}(I(p(i) + \Omega(\alpha, s_v, t), t))}{T}\right)^2}, \quad (2)$$

where $M_{\mathrm{VF}}$ is the maximum activation of the neuron, and $p(i)$ is the central RF position of neuron $\mathrm{VF}_{\alpha,v,i}$ in the visual input. With intensities ranging from 0 to 255, the value of $M_{\mathrm{VF}}$ is set to be 128 so that IC assumes only non-negative values. In the current implementation of the model, in order to compensate for noise in the visual input, each filter neuron's RF consists of multiple ($N = 20$) linear arrangements of inputs. The preferred directions $\alpha_n$ and speeds $s_n$ for these arrangements deviate slightly from the preferred motion $(\alpha, s)$ of the filter neuron. The following probability functions describe the statistical distribution of the variables $\alpha_n$ and $s_n$:

$$p(\alpha_n) = \frac{1}{30^\circ \cdot \sqrt{2\pi}} \exp\left\{-\frac{(\alpha_n - \alpha)^2}{2 \cdot (30^\circ)^2}\right\} \quad \text{and}$$

$$p(s_n) = \frac{1}{0.25s \cdot \sqrt{2\pi}} \exp\left\{-\frac{(s_n - s)^2}{2 \cdot (0.25s)^2}\right\} \quad \text{for } n = 1, \ldots, N. \quad (3)$$

Every filter neuron obtains one value $\mathrm{IC}_n$ for each linear arrangement of inputs, and the activation of a neuron $\mathrm{VF}_{\alpha,v,i}$ is then given by

$$\mathrm{VF}_{\alpha,v,i} = \max_{i \leqslant n \leqslant N} \mathrm{IC}_n. \quad (4)$$

Obviously, the filter neurons reach a state of high activation if a motion of their preferred orientation and velocity is present in their RF. However, maximum activation is also induced if there is no motion at all in a region of homogeneous intensity in the image sequence. Therefore, the function of the integration units $\mathrm{VI}_{\alpha,v,j}$ is not only to reduce the noise of the raw filter unit activations, but also to eliminate such "pseudo motion" detected by the filter cells. This is achieved by implementing lateral inhibition between units $\mathrm{VI}_{\alpha,v,j}$ with identical positions and speed selectivity, but different preferred directions of motion:

$$\mathrm{VI}_{\alpha,v,j} = \frac{1}{|R(j)|}\sum_{i \in R(j)} \mathrm{VF}_{\alpha,v,i} - \frac{1}{(n_\alpha - 1) \cdot |R(j)|}\sum_{\beta, \beta \neq \alpha}\sum_{i \in R(j)} \mathrm{VF}_{\beta,v,i}, \quad (5)$$

where $R(j)$ is the set of neurons whose outputs converge onto the integration unit $j$, and $n_\alpha$ denotes the number of implemented preferred directions of motion; in the present model, $n_\alpha = 12$. In the above formula, $\alpha$ and $\beta$ assume only these directions.

To achieve the V1 computation, the model uses one hypercolumn of $\mathrm{VF}_{\alpha,v,i}$ neurons for each pixel in the visual field. Each hypercolumn comprises one neuron of each type—because there are three different preferred speeds and 12 different preferred directions of motion, there are 36 units in each hypercolumn. Furthermore, the model employs 64 × 64 evenly distributed $\mathrm{VI}_{\alpha,v,j}$ hypercolumns (also 36 units per hypercolumn) that receive input from local filter units. In the present implementation the size of the input images are 256 × 256 pixels and integration units with RFs covering 8 × 8 neighboring filter units are used, thereby creating substantial overlap of RFs. The 64 × 64 hypercolumns of integration units provide the input for the model's MT neurons.

Figs. 3A–H show the filter outputs for all of the layers of the hierarchy for an input image sequence using a purely motion-defined object, that is, a pattern of random elements moving within a static random field (Gaussian noise). The motion is a counter-clockwise rotating square. The darker the pixel-value in the representation the stronger is the response; white means zero response.

### 2.3. Area MT

One group of cells in MT is tuned for a particular local speed and direction of movement, similar to V1 cells [20,27]. Another sub-population of MT neurons is selective for a particular angle between the local direction of movement and the speed gradient [25,31]. Here, MT has been designed with two different types of neurons: cells with selectivity identical to V1 neurons but larger RFs (detectors of translational motion) and cells selective for the angle between the direction of motion and the velocity gradient (detectors of velocity gradients). MT is implemented as a 30 × 30 array of hypercolumns, 468 neurons each (36 for translation as in V1 and 432 for gradient detection—three speeds, 12 directions, 12 direction/gradient angles). Each MT cell receives input from a 4 × 4 field of V1 neurons with the same direction and speed tuning.

For a translation neuron $i$ with preferred direction of motion $\alpha$ and speed selectivity of type $v$, its activation is given by:

$$\mathrm{MT}_{\alpha,v,i} = \sum_{j\in R(i)} G_{i,j} V_{\alpha,v,j}, \tag{6}$$

where

$$G_{i,j} = \frac{2}{\pi|R(i)|}\exp\left\{\frac{2(x_j^2+y_j^2)}{|R(i)|}\right\}. \tag{7}$$

In Eq. (7) and the following equations, $R(i)$ stands for the set of units that constitute the RF of neuron $i.G_{i,j}$ denotes the value of a two-dimensional (2D) Gaussian function at the position $(x_j, y_j)$ of neuron $j$ in the RF of neuron $i$. Here, $G_{i,j}$ represents the connection weights, and $x_j$ and $y_j$ are measured in unit spaces relative to the center of the RF. As can be seen from Eq. (7), the Gaussian functions were centered on their corresponding RF, and their standard deviation was always chosen to be half the length of the square-shaped RF. For example, since MT neurons have RFs of size 4 × 4, the peak of the Gaussian function is between the second and third RF neurons horizontally and vertically, and its standard deviation is 2.

The activation of a velocity gradient detector is computed as the product of the activation of V1 cells feeding into its RF with the same speed and direction tuning,

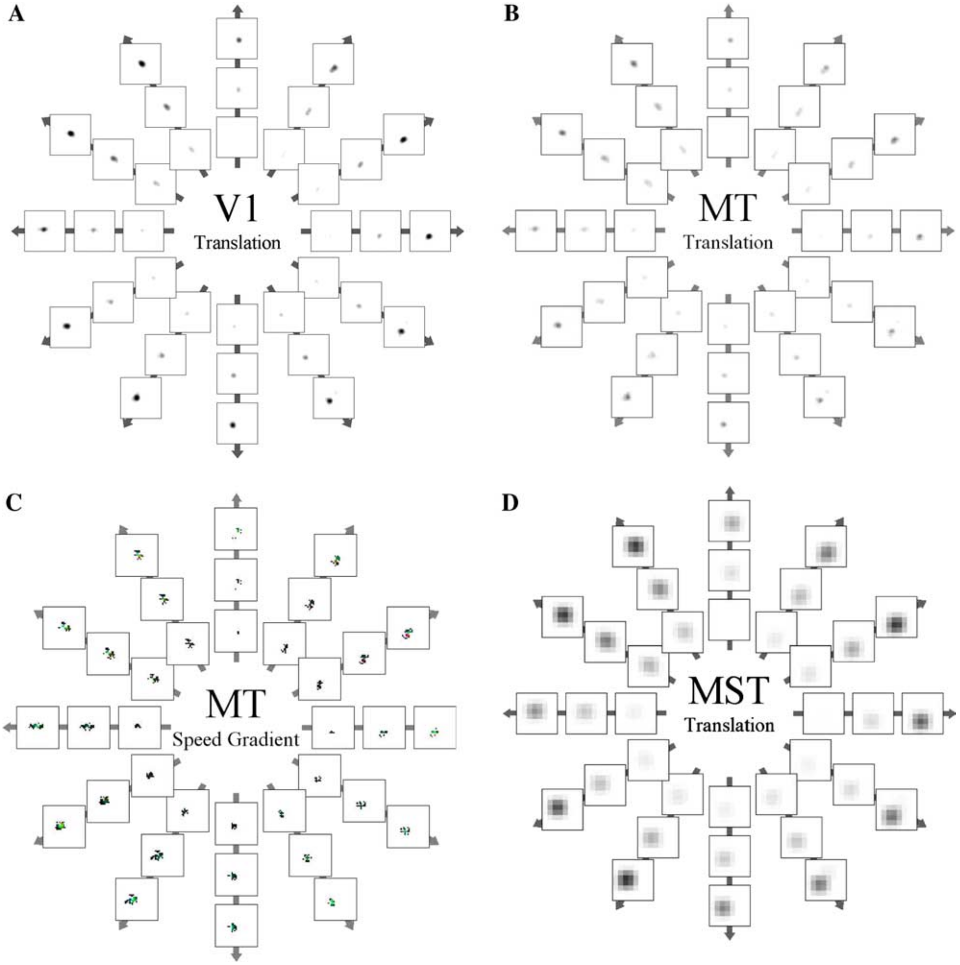

Fig. 3. Output of area computations. This series of figures gives in detail all of the filter outputs throughout the full motion hierarchy. The input is a square of random noise rotating counter-clockwise in place on a background of random noise. (A) The overall output of area V1, that is, the output of the integrative units. (B) Translation output in area MT. (C) Speed gradient output in area MT. This is a summary representation of the 12 different speed gradients at each local speed and direction. Each coloured dot is the maximum value across the 12 representations. This is not used for any decision process; it is only for a simpler visualization. (D) Translation output in area MST. (E) Generalized spiral output in area MST. (F) Translation output in area 7a. (G) Generalized spiral output in area 7a. (H) Rotation and radial output in area 7a.

and the gradient response. The gradient is determined by oriented RFs (example: RF for detecting upward gradient). For a velocity gradient neuron $i$ with preferred direction of motion $\alpha$, preferred angle $\delta$ between motion and gradient speed selectivity of type $v$:

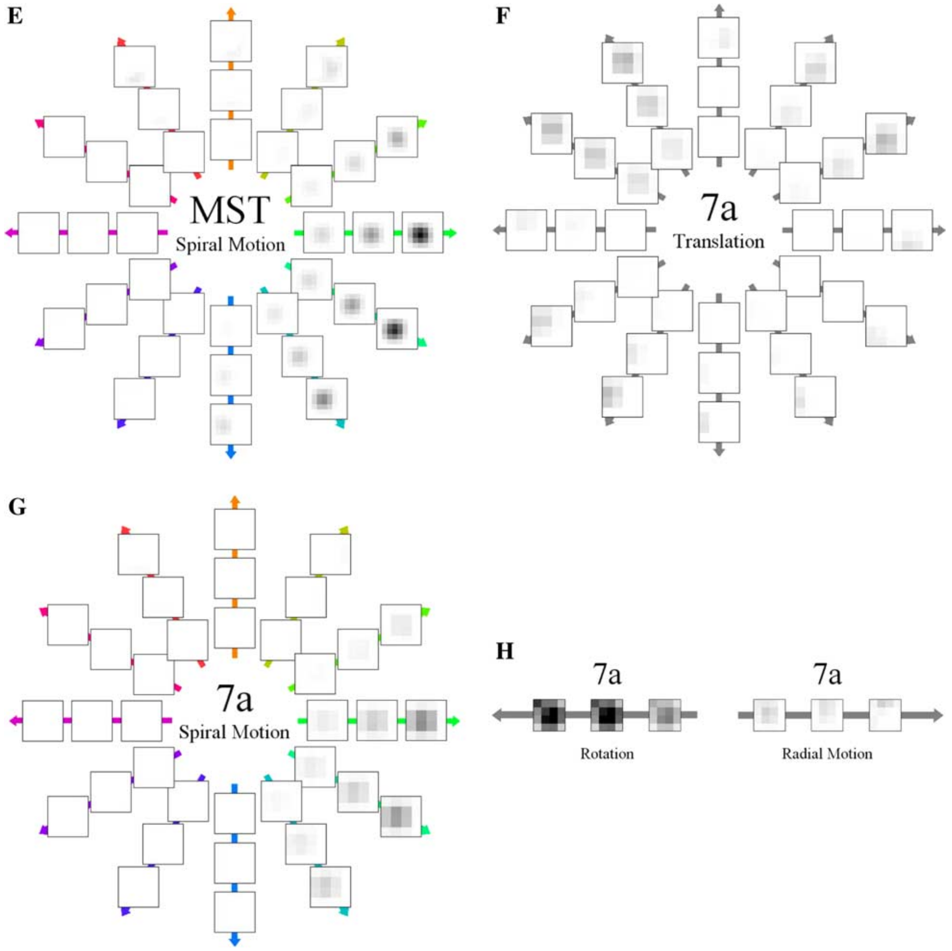

Fig 3. (*continued*)

$$\mathrm{MG}_{\alpha,\delta,v,i} = \begin{cases} 0 & \text{if } [\Delta(\alpha,\alpha+\delta,2,i) \\ & -\Delta(\alpha,\alpha+\delta,1,i) < \Theta] \wedge [\Delta(\alpha,\alpha+\delta,3,i) \\ & -\Delta(\alpha,\alpha+\delta,2,i) < \Theta], \\ \left(\sum\limits_{j\in R(i)} G_{i,j} V_{a,v,j}\right) \sum\limits_{k=1}^{3} c_k \Delta(\alpha,\alpha+\delta,k,i) & \text{otherwise.} \end{cases} \tag{8}$$

$\Theta$ is a threshold, $c_k$ are coefficients for the linear reconstruction of absolute speed from the activation of the three types of speed-selective neurons, and $\Delta(\alpha,\beta,v,i)$ specifies the activation increase in direction $\beta$ in the receptive field of neuron $i$ for speed selectivity type $v$ and direction $\alpha$:

$$\Delta(\alpha, \beta, v, i) = \sum_{j \in R(i)} G_{i,j} O_{\beta,j} V_{\alpha,v,j}. \tag{9}$$

$O_{\beta,j}$ is an orientation-specific configuration of weights that leads to maximum activation if the inputs increase in direction $\beta$. The individual weights are set to either $\frac{1}{|R_i|}$ or $-\frac{1}{|R_i|}$ in order to make the range of the neuron's activation values independent of its RF size. For example, a $4 \times 4$ RF whose preferred direction of motion is rightward ($\beta = 0$) has the following configuration $O_{\beta,j}$:

| | | | |
|---|---|---|---|
| $-\frac{1}{16}$ | $-\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ |
| $-\frac{1}{16}$ | $-\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ |
| $-\frac{1}{16}$ | $-\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ |
| $-\frac{1}{16}$ | $-\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ |

If in a region of the input image we consistently find the same angle $\delta$ between motion and speed gradient across all directions of motion, this signifies a particular motion pattern. An angle $\delta$ of 0° indicates expansion, 90°, indicates clockwise rotation, 180° indicates contraction, and 270° indicates counter-clockwise rotation. Any type of spiral motion can be represented this way; for example, an angle $\delta$ of 30° stands for expansion and a smaller proportion of clockwise rotation. This is the same coding as used in [7,25]. The color-coding for the angles used is shown in Fig. 1. The output of area MT computations for this input image sequence is shown in Figs. 3B and C.

### 2.4. Area MST

Cells in MST have larger receptive fields than MT cells and seem tuned to complex motion patterns: expand or approach, contract or recede, and rotation [26]. Two types of neurons are modeled: translation (as in V1) and spiral motion (clockwise and counter-clockwise rotation, expansion, contraction, and combinations). The reason translational motion is included here (as in area 7a) is so that a full pyramid of translation at all scales is included. MST is implemented as a $5 \times 5$ array of hypercolumns, 72 neurons each (36 for translation as in MT and 12 types of pattern selectivity for motion patterns with three speeds each). Each MST cell receives input from a $15 \times 15$ field of MT neurons that have the same tuning as the MST cell.

The activation of a translation neuron $i$ with preferred direction of motion $\alpha$ and speed selectivity of type $v$ is:

$$\mathrm{ST}_{\alpha,v,i} = \sum_{j \in R(i)} G_{i,j} \mathrm{MT}_{\alpha,v,j}. \tag{10}$$

Response for a spiral neuron $i$ selective for a pattern $\delta$ (angle between the direction and the speed gradient of motion as described in Section 2.3 and of speed selectivity type $v$ is:

$$\mathrm{SS}_{\delta,v,i} = \sum_{j \in R(i)} \sum_{\alpha} G_{i,j} \mathrm{MG}_{\alpha,\delta,v,j}. \tag{11}$$

As for area MT, a direction to speed gradient angle of 0° indicates expansion, 90° indicates clockwise rotation, 180° indicates contraction, and 270° indicates counter-clockwise rotation and other angles represent combinations of motion types. The output of area MST neurons is shown in Figs. 3D and E.

### 2.5. Area 7a

Area 7a seems to involve at least four different types of computations but with larger RFs than the other areas [23]: translation and spiral motion, as in MST, rotation (clockwise or counter-clockwise regardless of direction), and radial motion (irrespective of direction, expansion or contraction). Types AT and AS have the same properties as ST and SS, respectively, except for their RF size:

$$\mathrm{AT}_{\alpha,v,i} = \sum_{j \in R(i)} G_{i,j} \mathrm{ST}_{\alpha,v,j}, \tag{12}$$

$$\mathrm{AS}_{\delta,v,i} = \sum_{j \in R(i)} G_{i,j} \mathrm{SS}_{\delta,v,j}. \tag{13}$$

Neuron ART responds to rotation, regardless of the direction of rotation (clockwise or counter-clockwise). The activation of ART neuron $i$ with speed selectivity type $v$ is computed as follows:

$$\mathrm{ART}_{v,i} = \sum_{\delta} \sum_{j \in R(i)} G_{i,j} \mathrm{SS}_{\delta,v,j} [D(\delta - 90^\circ) + D(\delta - 270^\circ)], \tag{14}$$

where $D$ is the direction selectivity function defined as a one-dimensional Gaussian function:

$$D(\gamma) = \frac{1}{\sigma_D \sqrt{2\pi}} \mathrm{e}^{-\gamma^2/2\sigma_D^2} \quad \text{with } \sigma_D = -45^\circ. \tag{15}$$

Neuron ARD responds to radial motion, regardless of whether it is expansion or contraction. The activation of ARD neuron $i$ of speed selectivity type $v$ is given by:

$$\mathrm{ARD}_{v,i} = \sum_{\delta} \sum_{j \in R(i)} G_{i,j} \mathrm{SS}_{\delta,v,j} [D(\delta) + D(\delta - 180^\circ)]. \tag{16}$$

Area 7a is implemented as a $4 \times 4$ array of hypercolumns, 78 neurons each (36 for translation, 3 speeds of rotation, 3 speeds of radial motion, and 3 speeds times 12 patterns for spiral motions). Each 7a cell receives input from a $4 \times 4$ field of MST neurons that have the relevant tuning. The output of area 7a computations is shown in Figs. 3F–H.

### 2.6. Motion hierarchy summary

The motion pyramid is novel in that it contains two separate streams of processing, translation, and generalized spiral motions. It features a decomposition of motion into simpler components including local spatial gradients of velocity. As the previous figures demonstrate the representations that result from the several different neural populations (654 separate full field filter representations) are complex and non-trivial. Although the outputs are noisy, the effect of noise is gradually ameliorated as the signals reach higher levels of the pyramid. It is satisfying to note that even with this complexity of representation, peak responses are right where they should be in terms of correct feature detection and a search strategy (as described in the next section) can successfully find those peaks.

It is important to acknowledge a weakness of the present work that has resulted from the original motivation for the research described here. This research was motivated by the valid criticism that past demonstrations (as in [2]) used simple Gaussian pyramids for the features on which the ST mechanisms were demonstrated. This choice did not affect the demonstrations, which did indeed properly show all the characteristics of ST. However, it did make those demonstrations less useful and did allow the possibility that the ST mechanisms might not work as expected with realistic feature pyramids. We chose to address these criticisms within the visual motion domain. Thus, the task of defining motion 'neurons' along the motion pathway was addressed in a coarse, first-order fashion only and thus the filter definitions (but not the motion decompositions) are perhaps not as strong as they could be. Current research is attempting to improve the motion representation; this in no way will affect the ST demonstration and will only strengthen the motion representations.

## 3. The selective tuning model of visual attention

The modeling effort described herein must be distinguished from others in at least the following ways: it is not a neural network that learns to attend; it is not a model whose goal is to explain a particular set of quantitative observations; it is not a data fitting exercise; it is not a set of equations whose numerical simulation leads to output functions whose form seems similar to experimental data. In contrast, we are trying to show from 'first principles' what qualitative form visual processing must take and to define a theory and an accompanying computer simulation that can take as input digital images and perform at least qualitatively in the same manner as human or primate vision performs. It features a theoretical foundation of provable properties based in the theory of computational complexity [1,32,33,34]. The 'first principles' arise because vision is formulated as a search problem (given a specific input, what is the subset of neurons that best represent the content of the image?) and complexity theory is concerned with the cost of achieving solutions to such problems. This foundation suggests a specific biologically plausible architecture as well as its processing stages as will be briefly described in this article (a more detailed account can be found in [1,2]). It should be clear that these foundations were derived using the visual search problem. Considerations related to other dimensions of attention functionality remain (however note that in [2] saccades to peripheral targets are included in the model and that in Zaharescu et al. [35] added active visual search to ST).

ST is not compared here to other attention models except where particular points need to be made in order to keep the paper length under control; general comparisons have appeared elsewhere in previous ST publications as cited herein.

### *3.1. The model*

The visual processing architecture is pyramidal in structure with units within this network receiving both feed-forward and feedback connections. When a stimulus is presented to the input layer of the pyramid, it activates in a feed-forward manner all of the units within the pyramid with receptive fields (RFs) mapping to the stimulus location; the result is a diverging cone of activity within the processing pyramid. It is assumed that response strength of units in the network is a measure of goodness-of-match of the stimulus within the RF to the model that determines the selectivity of that unit.

Selection relies on a hierarchy of winner-take-all processes. WTA is a parallel algorithm for finding the maximum value in a set. First, a WTA process operates across the entire visual field at the top layer where it computes the global winner, i.e., the units with largest response (see Section 3.3 for details). The WTA can accept guidance to favor areas or stimulus qualities if that guidance is available but operates independently otherwise. The search process then proceeds to the lower levels by activating a hierarchy of WTA processes. The global winner activates a WTA that operates only over its direct inputs to select the strongest responding region within its RF. Next, all of the feed-forward connections in the visual pyramid that do not contribute to the winner are pruned (inhibited). As a result, the input to the higher-level unit changes and thus its output changes. This refinement of unit responses is an important consequence because one of the important goals of attention is to reduce or eliminate signal interference [1]. By the end of this refinement process, the output of the attended units at the top layer will be the same as if the attended stimulus appeared on a blank field. This strategy of finding the winners within successively smaller RF, layer by layer, in the pyramid and then pruning away irrelevant connections through inhibition is applied recursively through the pyramid. The end result is that from a globally strongest response, the cause of that largest response is localized in the sensory field at the earliest levels. The paths remaining may be considered the pass zone of the attended stimulus while the pruned paths form the inhibitory zone of an attentional beam. The WTA does not violate biological connectivity or relative timing constraints. Fig. 4 gives a pictorial representation of this attentional beam.

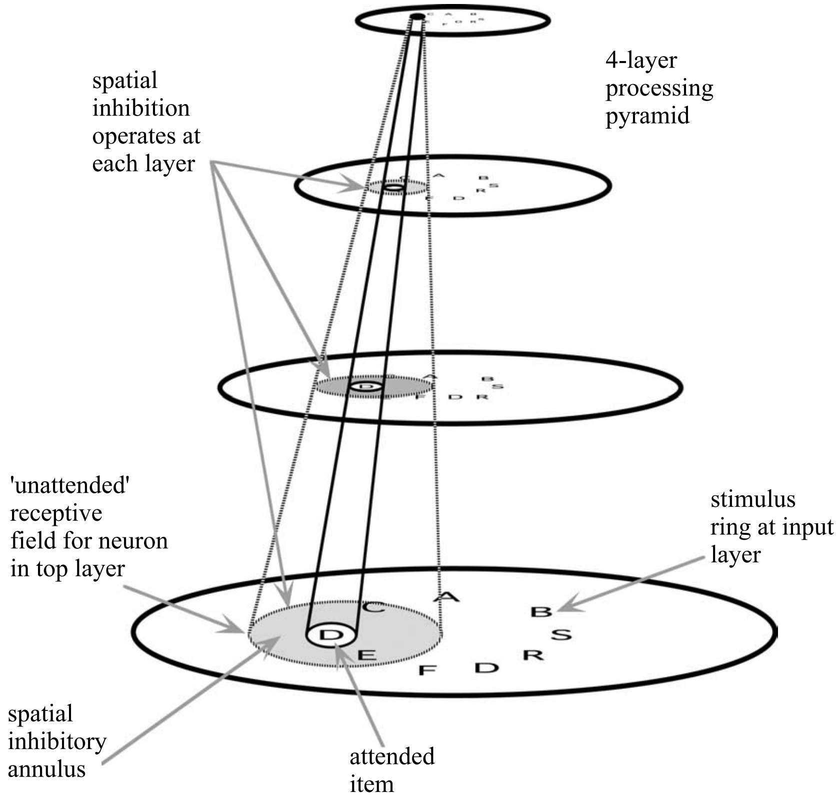

Fig. 4. Attentional beam. This shows the rationale for suppression around attended items that is a feature of ST.

An executive controller is responsible for implementing the following sequence of operations for visual search tasks:

1. Acquire target as appropriate for the task, store in working memory.
2. Apply top-down biases, inhibiting units that compute task irrelevant quantities.
3. 'See' the stimulus, activating feature pyramids in a feed-forward manner.
4. Activate top-down WTA process at top layers of feature pyramids.
5. Implement a layer-by-layer top-down search through the hierarchical WTA based on the winners in the top layer.
6. After completion, permit time for refined stimulus computation to complete a second feed-forward pass. Note that this feed-forward refinement does not begin with the completion of the lowermost WTA process; rather, it occurs simultaneously with completing WTA processes (step 5) as they proceed downwards in the hierarchy. On completion of the lowermost WTA, some additional time is required for the completion of the feed-forward refinement.
7. Extract output of top layers and place in working memory for task verification.

8. Inhibit pass zone connections to permit next most salient item to be processed.
9. Cycle through steps 4–8 as many times as required to satisfy the task.

This multi-pass process may seem to not reflect the reality of biological processes that seem very fast. However, it is not claimed that all of these steps are needed for all tasks. Several different levels of tasks may be distinguished, defined as:

*Detection*—is a particular item present in the stimulus, yes or no?
*Localization*—detection plus accurate location;
*Recognition*—localization plus accurate description of stimulus;
*Understanding*—recognition plus role of stimulus in the context of the scene.

The executive controller is responsible for the choice of task based on instruction. If detection is the task, then the winner after step 4, if it matches the target, will suffice and the remaining steps are not needed. Thus simple detection in this framework requires only a single feed-forward pass ([1], also argued by Thorpe [36]). If a localization task is required, then all steps up to 7 are required because, as argued in Section 2.2, the top-down WTA is needed to isolate the stimulus and remove the signal interference from nearby stimuli. This clearly takes more time to accomplish. If recognition is the task, then all steps, and perhaps several iterations of the procedure, are needed in order to provide a complete description. The understanding task has similar requirements, although this is not quite within the scope of the model at this point.

### *3.2. Top-down selection*

ST features a top-down selection mechanism based on a coarse-to-fine WTA hierarchy. Why is a purely feed-forward strategy not sufficient as Riesenhuber and Poggio claim [37]? There seems to be no disagreement on the need for top-down mechanisms if task/domain knowledge is considered, although few non-trivial schemes seem to exist. Biological evidence, as well as complexity arguments, suggests that the visual architecture consists of a multi-layer hierarchy with pyramidal abstraction. One task of selective attention is to find the value, location, and extent of the most 'salient' image subset within this architecture. A purely feed-forward scheme operating on such a pyramid with:

(i) Fixed size receptive fields with no overlap, is able to find the largest single stimulus input with local WTA computations for each receptive field but location is lost and stimulus extent cannot be considered.
(ii) Fixed size overlapping receptive fields, suffers from the spreading winners problem due to neural convergence, and although the largest input value can be found, the signal is blurred across the output layer, location is lost, and extent is ambiguous.
(iii) All possible RF sizes in each layer, becomes intractable due to combinatorics [1].

While case: (i) might be useful for certain computer vision detection tasks, it cannot be considered as a reasonable proposal for biological vision because it fails to localize targets. Case (iii) is not plausible as it is intractable. Case (ii) reflects a biologically realistic architecture, yet fails at the task of localizing a target. Given this reality, a purely feed-forward scheme is insufficient to describe biological vision. Only a combined bottom-up and top-down strategy can successfully determine the location and extent of a selected stimulus in a constrained pyramidal architecture.

### 3.3. WTA and saliency

The Winner-Take-All scheme within ST is defined as an iterative process that can be realized in a biologically plausible manner insofar as time to convergence and connectivity requirements are concerned. It has its roots in Koch and Ullman's model [38] (but also see [39–41]) but provides a complete redefinition with proofs of convergence and convergence properties fully described in [2]. The basis for its distinguishing characteristic comes from the fact that it implicitly creates a partitioning of the set of unit responses into bins of width determined by a task-specific parameter, $\theta$. The partitioning arises because inhibition between units is not based on the value of a single unit but rather on the absolute value of the difference between pairs of unit values. Further, this WTA process is not restricted to converging to single points as all other formulations. The winning bin of the partition, whose determination is now described, is claimed to represent the strongest responding contiguous region in the image (this is formally proved in [2]).

First, the WTA implementation uses an iterative algorithm with unit response values updated after each step until convergence is achieved. Competition depends linearly on the difference between unit strengths in the following way. Unit $A$ will inhibit unit $B$ in the competition if the response of $A$, denoted by $\rho(\mathrm{A})$ satisfies $|\rho(\mathrm{A}) - \rho(\mathrm{B})| > \theta$. Otherwise $A$ will not inhibit $B$. The overall impact of the competition on unit $B$ is the weighted sum of all inhibitory effects, each of whose magnitude is determined by $|\rho(\mathrm{A}) - \rho(\mathrm{B})|$. It has been shown [2] that this WTA is guaranteed to converge, has well-defined properties with respect to finding strongest items, and has well-defined convergence characteristics. The time to convergence, in contrast to any other iterative or relaxation-based method is specified by a simple relationship involving $\theta$ and the maximum possible value, $Z$, across all unit responses. The reason for this is that the partitioning procedure uses differences of values. All larger units will inhibit the units with the smallest responses, while no units will inhibit the largest valued units. As a result the small response units are reduced to zero very quickly while the time for the second largest units to be eliminated depends only on the values of those units and the largest units. As a result, a two-unit network is easy to characterize. The time to convergence is given by $\log_2(\frac{A-\theta}{A-B})$ where $A$ is the largest value and $B$ the second largest value. This is also quite consistent with behavioral evidence; the closer in response strength two units are, the longer it takes to distinguish them.

Second, the competition depends linearly on the topographical distance between units, i.e., the features they represent. The larger the distance between units, the greater the inhibition. This strategy will find the largest, most spatially contiguous subset within the winning bin. A spatially large and contiguous region will inhibit a contiguous region of similar response strengths but of smaller spatial extent because more units from the large region apply inhibition to the smaller region than inhibit the larger region from the smaller one. At the top layer, this is a global competition; at lower layers, it only takes place within receptive fields. In this way, the process does not require implausible connectivity lengths. For efficiency reasons, this is currently only implemented for the units in the winning bin. With respect to the weighted sums computed, in practice the weights depend strongly on the types of computations the units represent. There may also be a task-specific component included in the weights. Finally, a rectifier is needed for the whole operation to ensure that no unit values go below zero. The iterative update continues until there is only one bin of positive response values remaining and all other bins contain units whose values have fallen below $\theta$. Note that even the winning bin of positive values must be of a value greater than some threshold in order to eliminate false detections due to noise.

The key question is how is the root of the WTA process hierarchy determined? The following is a conceptual description of this where the 'max' function used below is implemented using the iterative process just described. Let $F$ be the set of feature maps at the output layers overall, and $F^i$, $i = 1$ to $n$, be particular feature maps. Values at each $x, y$ location within map $i$ are represented by $M^i_{x,y}$. The root of the WTA computation is set by a competition at the top layers of the pyramid depending on network configuration (task biases can weight each computation). The winning value is $W$, and this is determined by:

1. If there is only a single active feature pyramid $f$,

$$W = \max_{x,y} M^f_{x,y}. \tag{17}$$

2. If $F$ contains more than one feature map, representing mutually exclusive features, then

$$W = \max_{x,y} \left(\max_{i \in F} M^i_{x,y}\right). \tag{18}$$

3. If $F$ contains more than one feature map representing features that can co-exist at each point, then there is more than one WTA process, all rooted at the same location but operating through different feature pyramids

$$W = \max_{x,y} \left(\sum_{i \in F} M^i_{x,y}\right). \tag{19}$$

4. If $F$ contains subsets representing features that are mutually exclusive (the set $A$, as in case 2 above) as well as complementary (the set $B$, as in case 3 above), the winning locations are determined by the sum of the strongest response among set $B$ (following method 3) plus the strongest response within set $A$ (using method 2). Thus, a combination of the above strategies is used. There is more than one WTA process, all rooted at the same location but operating through different feature pyramids:

$$W = \max_{x,y}\left[\sum_{b\in B} M^{b}_{x,y} + \max_{a\in A}(M^{a}_{x,y})\right]. \tag{20}$$

At the 'top' of the overall processing layers Eq. (20) applies and includes all representations. At all other layers of the hierarchy, the equation used is determined by the receptive field properties of the neurons in each representation. As a result, there is no single saliency map in this model as there is in most other models ([42,38,43] and others). Notable exceptions are the models of Hamker [44] and of Deco and Zihl [45] both of which claim no salience map. Nevertheless, the similarity ends there. The Hamker strategy considers attentional effects in V4 only, does not provide a mechanism for how attentional control signals are generated, and says nothing about the contribution to overall perception by attentional modulation in all the other visual areas. In Deco and Zihl, the Neurodynamical Model implicitly codes saliency as a distribution of modulation across the feature maps. Feature maps relevant for the task are enhanced and/or distracters are inhibited, the dynamics of the network producing winners without the need for explicit representation of salience. Selection occurs through inhibitory competition within neuronal pools. They do not consider a realistic implementation since simple saliency matrices form their input.

Although the WTA in ST was introduced in the preceding paragraphs for the top layer only, in ST there is no single WTA process necessarily, but several simultaneous WTA threads that extend through the hierarchy to all layers. Eqs. (17) and (18) lead to a single WTA thread moving from top layer through the hierarchy; Eqs. (19) and (20) lead to multiple top-down WTA threads, one for each of the representations that may co-exist. Each of those will localize their features. Saliency is a dynamic, local, distributed, and task-specific determination and one that may differ even between processing layers as required. Although it is known that feature combinations of high complexity do exist in the higher levels of cortex, the above does not assume that all possible combinations must exist. Features are encoded separately in a pre-defined set of maps and the relationships of competition or cooperation among them provide the potential for combinations. The above four types of competitions then select which combinations are to be further explored. This flexibility allows for a solution (at least in part) to the binding issues that arise for this domain.

The WTA process is implemented utilizing a top-down hierarchy of units. There are two main unit types: gating control units and gating units. Gating control units are associated with each competition in each layer and at the top, are activated by the executive controller in order to begin the WTA process. An additional network of top-down bias units can also provide task-specific bias if it is available. They communicate downwards to gating units that form the competitive gating network for each WTA within a receptive field. Whether the competition uses Eqs. (17)–(19), or (20) depends on the nature of the inputs to the receptive field. Once a particular competition converges, the gating control unit associated with that unit sends downward signals for the competition to begin at the next layer down in the pyramid. The process continues until all layers have converged.

### 3.4. The simulation

The model has been implemented and tested in several labs applying it to computer vision and robotics tasks. The current model structure is shown in Fig. 5. The executive controller and working memory, the motion pathway (V1, MT, MST, and 7a), the peripheral target area PO, the gaze WTA and gaze controller have all

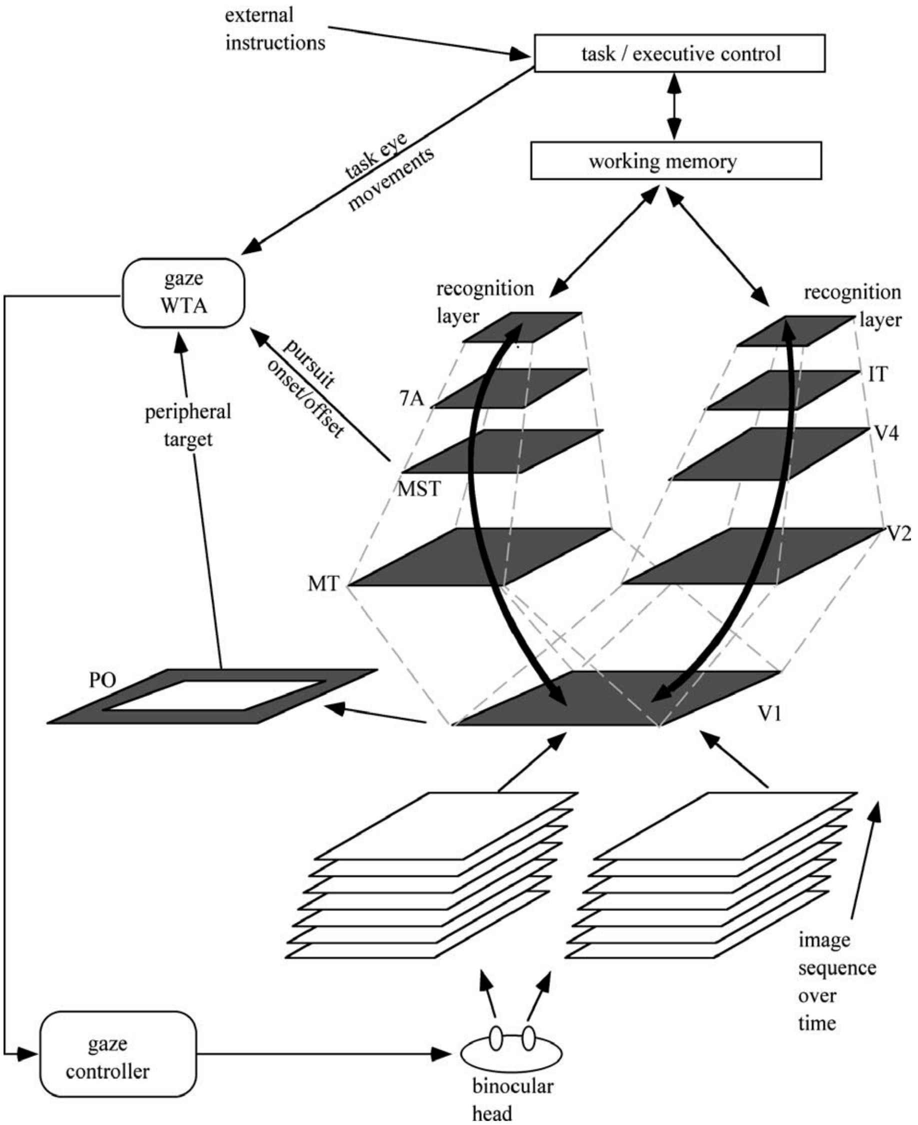

Fig. 5. ST full hierarchy. The full visual processing hierarchy on which ST operates is depicted. This paper focuses on the motion pathway—areas V1, MT, MST, and 7a. Several other components have been demonstrated previously while others are current research topics.

been implemented and examples of performance can be found in [2,14,35,46,47]. Work is currently underway to extend the implementation to the object pathway (V1, V2, V4, and IT) and to binocular stimuli as well as extensions of the executive controller and recognition layers.

### *3.5. A full hierarchy example*

Fig. 6 shows an example using a purely motion-defined object, that is, a pattern of random elements moving within a static random field (Gaussian noise). The motion is a counter-clockwise rotating square, the same as for the sequence of Fig. 3. The figure also illustrates how separate features in different locations and represented in different maps are bound together into a whole, a process that will be described in Section 4.

### *3.6. Biological and behavioral predictions*

The first description of the overall structure of the model appeared along with most of the basic predictions in 1990 [1]. These included (with support that has appeared since):

- Suppression around attended items in spatial as well as in the feature dimension [48–55].
- Attention is a top-down process; attentional guidance and control are integrated into the visual processing hierarchy, rather than being centralized in some external brain structure implying that the latency of attentional modulations *decreases* from lower to higher visual areas [56,57].
- Attentional modulation appears wherever there is many-to-one, feed-forward neural convergence, something that in 1990 had no support at all [56,58,59].
- Topographic distance between attended items and distractors affects amount of attentional modulation [51].

Additional predictions and supporting arguments can be found in [2,18]. These counter-intuitive predictions made well before any hints of experimental evidence, provide the strongest possible argument for the biological realism of the theory behind the ST model.

## 4. Using ST to attend to and localize motion patterns

Most of the computational models of primate motion perception that have been proposed concentrate on feed-forward, classical types of processing and do not address attentional issues. However, there is strong evidence that the responses of motion neurons in at least areas MT, MST, and 7a are modulated by attention [60–62]. As a result of the model's feed-forward computations, the neural responses in the high-level areas (MST and 7a) indicate the kind of motion patterns presented as an

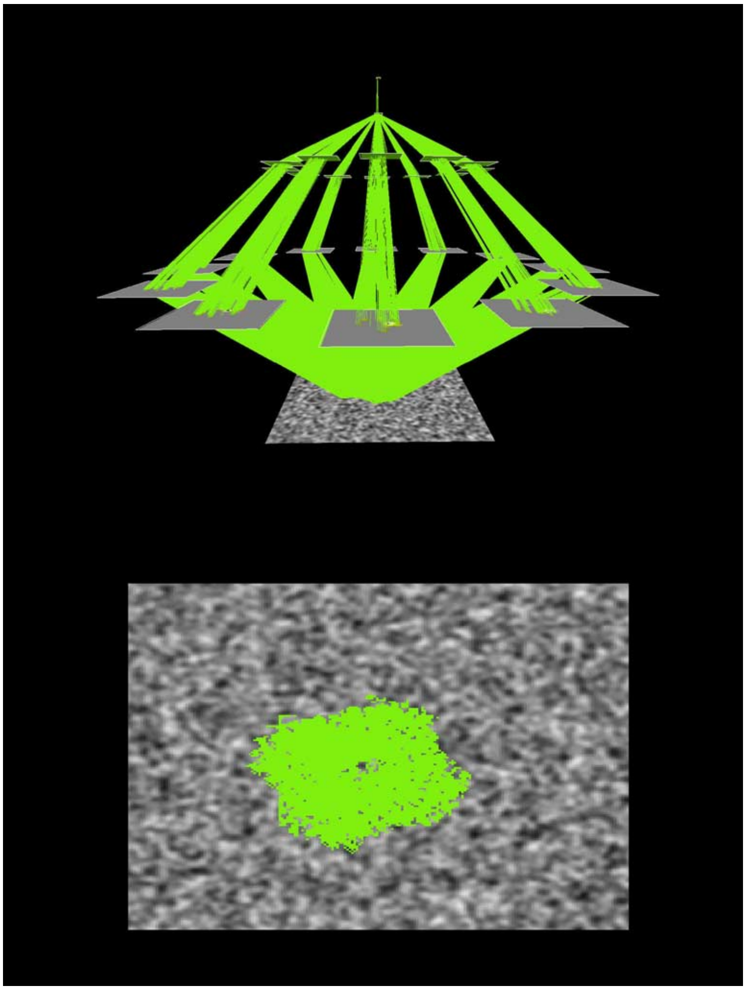

Fig. 6. Attending to a motion defined object. This shows the structure of the attention beam that localizes and labels the rotating square whose feed-forward outputs are shown in Fig. 3. The beam color is green, which signifies counter-clockwise rotation. (see the colour wheel in Fig. 1). Note also the fact that its root is in a single representation of 7a (spiral neurons), and then the beam splits to include all the components of the rotating object localizing those components in each of the MT and MST representations. The beam then reunifies at the input image, binding together the pieces into a whole. The top of the figure shows the active beam pass zone structure; the bottom of the figure shows the localization of the motion in the image. In this figure all layers of the pyramid are clearly visible, the active representations within each layer only are shown. In the figures that follow, the pyramid will be tilted into the page so that the input can be shown together with the beam structure all in one figure. However, the top layer is thus occluded from view; it is however still part of the overall beam.

input but do not localize the spatial position of the patterns. The ST model was then applied to this feed-forward pyramid, adding the required feedback connections, hierarchical WTA processes, and gating networks as originally defined. The model attends to motion, whether it exhibits a single motion or a combination of motion types, and serially focuses on each motion, sequentially, in order of response strength.

The hierarchical WTA described earlier finds the globally most active region. Then for this region, WTA processes are activated as described in Section 3. Translational and spiral motion patterns can co-exist for the same object (Eq. (18)). The remaining processing proceeds as described earlier for each of the winning patterns. The model also includes processes detecting onset and offset events (start and stop), but these are not described here (see [47]).

### *4.1. Feature binding*

A major contribution of the demonstration of how ST can operate within such a complex hierarchy is the method of grouping features (known as the binding problem in computational neuroscience [63]). It is not claimed that this particular strategy has sufficient generality to solve all possible issues within the binding problem; however it seems to solve the limited cases that occur in image sequences of simple motion patterns. As such, it is the first instance of such a solution and further work will investigate its generality.

Quoting Roskies [64], "the canonical example of binding is the one suggested by Rosenblatt [65] in which one sort of visual feature, such as an object's shape, must be correctly associated with another feature, such as its location, to provide a unified representation of that object." Such explicit association ("binding") is particularly important when more than one visual object is present, in order to avoid incorrect combinations of features belonging to different objects, otherwise known as 'illusory conjunctions' [66]. Several other examples of the varieties of binding problems in the literature appear in a special issue of *Neuron* edited by Roskies [63]. At least some authors [67,68] suggest that specialized neurons that code feature combinations (introduced as cardinal cells by Barlow [69]) may assist in binding. The solution in this paper does indeed include such cells; however, they do not suffice on their own as will be described because they alone cannot solve the localization problem.

What is demonstrated here through the use of localized saliency and WTA decision processes, is precisely what the binding problem requires: neurons in different representations that respond to different features and in different locations are selected together, the selection being in location and in feature space, and are thus bound together via the 'pass' zone(s) of the attention mechanism. Even if there is no single neuron at the top of the pyramid that represents the concept, the WTA allows for multiple threads bound through location by definition in Eqs. (17)–(20).

Part of the difficulty facing research on binding is the confusion over definitions and the wide variety of tasks included in binding discussions. For example, in Feature Integration Theory (FIT) [70] location is a feature because FIT assumes it is faithfully represented in a master map of locations. But this cannot be true; location precision changes layer to layer in any pyramid representation. In the cortex, it is not accurate in a Euclidean sense almost anywhere, although the topography seems qualitatively preserved [19]. The wiring pattern matters in order to get the right image bits to the right neurons. Thus binding needs to occur layer to layer and is not simply a problem for high-level consideration. Features from different representations with different location coding properties converge onto single cells and this seems to necessitate an active search process.

This proposal is supported by the architecture described by Felleman et al. [71] for the object recognition pathway of V1, V2, V4, and IT. They suggest that specific patterns of inter-cortical input and cortical circuitry may permit new and more complex receptive field properties for extrastriate cortical neurons. This appears true for both feed-forward as well as feedback connectivity. Projections display a complicated submodular selectivity with the modules being inter-digitating, non-overlapping, and highly intermixed. This structure necessitates a different view of how neural inputs are handled, each of these different inputs perhaps dealt with differently. The strategy presented in this paper is a step toward providing a computational framework for this architecture.

For the purposes of this argument, consider the following:

1. Location is not a feature, rather, it is the anchor that permits features to be bound together. Location is defined broadly, may be single points or groups of contiguous points, and may be differently organized in each visual area; in practice it is considered to be local coordinates within a visual area (think of an array of hypercolumns, each with its own local coordinates).
2. A grouping of features not coincident by location cannot be considered as a unitary group unless there is a unit to represent that group with a receptive field definition that takes in input from the different feature representations providing a 'template' for the group.
3. Features that compose a group may be in different locations and represented in different visual areas as long as they converge onto units that represent the group.
4. If the group is attended, then the WTA of Section 3.3 will find and attend to each of its parts regardless of their location or feature map representation.

This is a solution to the aspect of binding that attends to groups and finds parts of groups. It applies equally well for object recognition: faces are good examples of a grouping of features. In the demonstrations below, the groups are motion patterns. There are several components to this solution. The first has to do with the particular representations chosen for motion patterns. Our representation is hierarchical with each layer being defined using components from the previous. Note how a constant speed-rotating object exhibits constant velocity gradient across location with respect to local motion. A neuron higher in the hierarchy then can be selective to regions that are homogeneous for this value and this is an easy selectivity to define and implement. As shown in Eq. (11), a motion pattern detector in layer MST simply sums responses of the corresponding MT units that feed it. An example is in order using the figure sequence in Fig. 3. In layer MT, neurons sensitive to local motion gradients respond as shown (Fig. 3C). Across all directions in the representation, one sees that the object has been deconstructed—'cut into pie pieces'—one for each local motion direction. That is, the tuning properties of the neurons have decomposed the flow field into distinct areas of constant velocity gradient. Note that these have also been partitioned depending on speed. Then, at the MST layer, the neurons whose selectivity is for rotation within this particular speed band will receive input from these MT representations (and not from the others). The MST neuron whose receptive field is best centered on the object will fire strongest if it receives sufficient stimulation, which in this case means that it sees all pieces of the pie. This best responding neuron can now be considered as having grouped the pie pieces and re-assembled the pie, that is, to have bound together the representations at the MT layer which otherwise are neither co-incident by location nor feature type. This is the feed-forward part of this process—an implicit binding action. If the task of the system were to simply detect the presence of a particular motion pattern, this representation would suffice as long as the top-level global WTA selects this region. However, if the system's task is to localize or recognize, then the job is not complete. As is clear in the figure, there are many MST neurons that respond. The feedback process of top-down attention selects the best of these responses, and actively sub-selects the particular regions of MT neurons that correspond to that best firing, and thus best fitting the pattern selectivity of the neuron. The unique aspect here is that the receptive field of the MST neuron is defined by a spatial region as well as a subset of features computed within that spatial region, each feature contributing a component across that spatial region (as specified by Eq. (11); however, it is easily seen that if spatial distributions different from uniform motion are of interest, variations of this equation can be set up to permit any spatial combination of local gradients). There are 468 feature maps in the MT layer that feed the 72 MST layer units and these can be organized for a huge variety of distinct motions. Translation and spiral motions can co-exist for the same spatial object, and thus there is a WTA thread for translation and another one for spiral motion. The overall peak then uses the strategy described in Section 3.3 to grow the full region corresponding to the moving object. At layer 7a, the same is seen. Full-field motion takes precedence over object motion in these representations. Thus, full-field rotations, full-field contractions, and spiral motions cannot co-exist. They can co-exist with translation however. All feature maps are not complementary nor do they all play equivalent roles in the WTA process. This shows the need for a more flexible view on saliency and WTA computations than has been previously shown in other models (all other models use the definition and structure first presented by Koch and Ullman [38]). No other model currently includes such a distributed definition of saliency.

What if a more complex binding problem is considered, one where multiple motion patterns appear in an image sequence? As can be seen in Fig. 7, the two moving objects (this time in an image sequence of a real scene) activate many representations within the hierarchy and both can be seen within each of several representations. The rectangular object is approaching the camera while the circular one is rotating counter-clockwise. This is again a classic binding problem. Although the representation at the output layer (7a) is more complex due to multiple stimuli, the WTA is still able to choose a peak, and use that location information as well as the tuning properties (i.e., feed-forward connectivity) of the winning units to sub-select the correct components of the winning pattern. The full sequence is

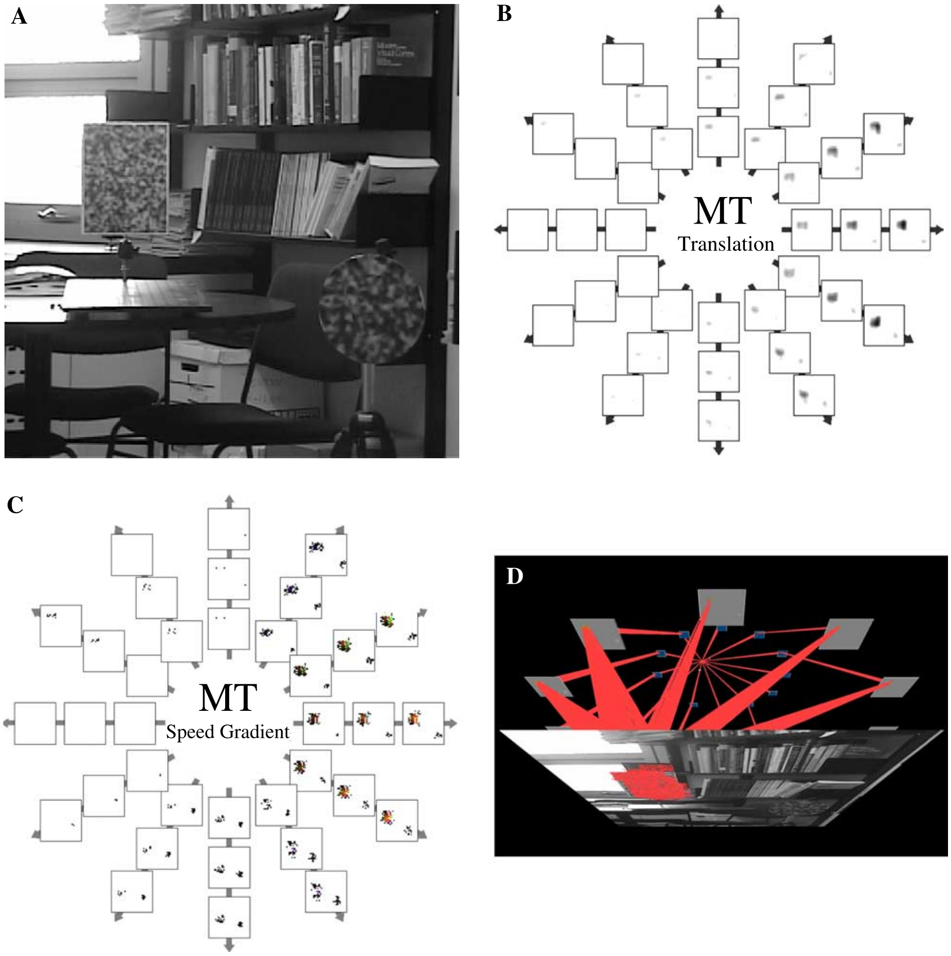

Fig. 7. Grouping across location for motion pattern detection. (A) In this example, real images are used of two textured objects against a cluttered background, the rectangle is approaching the camera while the circle is rotating counter-clockwise. (B and C) The output of area MT in the first feed-forward pass in summary form as described for Fig. 3C. (D) Localization of the first winning area in red, the colour signifying 'approach' (E) Inhibition of return on the attended pathways. (F–G) The output of area MT in the second feed-forward pass, after the inhibition of return is applied, in summary form as described for Fig. 3C. (H) The second attentional fixation, in green, the colour for counter-clockwise rotation. The inhibition of return seems to not be perfect, that is, not all pixels due to the approaching object are eliminated and so some additional responses remain. The reason for this is that the IOR is set using the attended location. The object however continues to move and since it is approaching, its appearance is beyond that original attended and thus inhibited location. Future work would enable inhibition not only of location but of the attended object.

shown in the figure in order to also show the sequential nature of the selection process that attends to each of the patterns (two passes through the algorithm in Section 3.1).

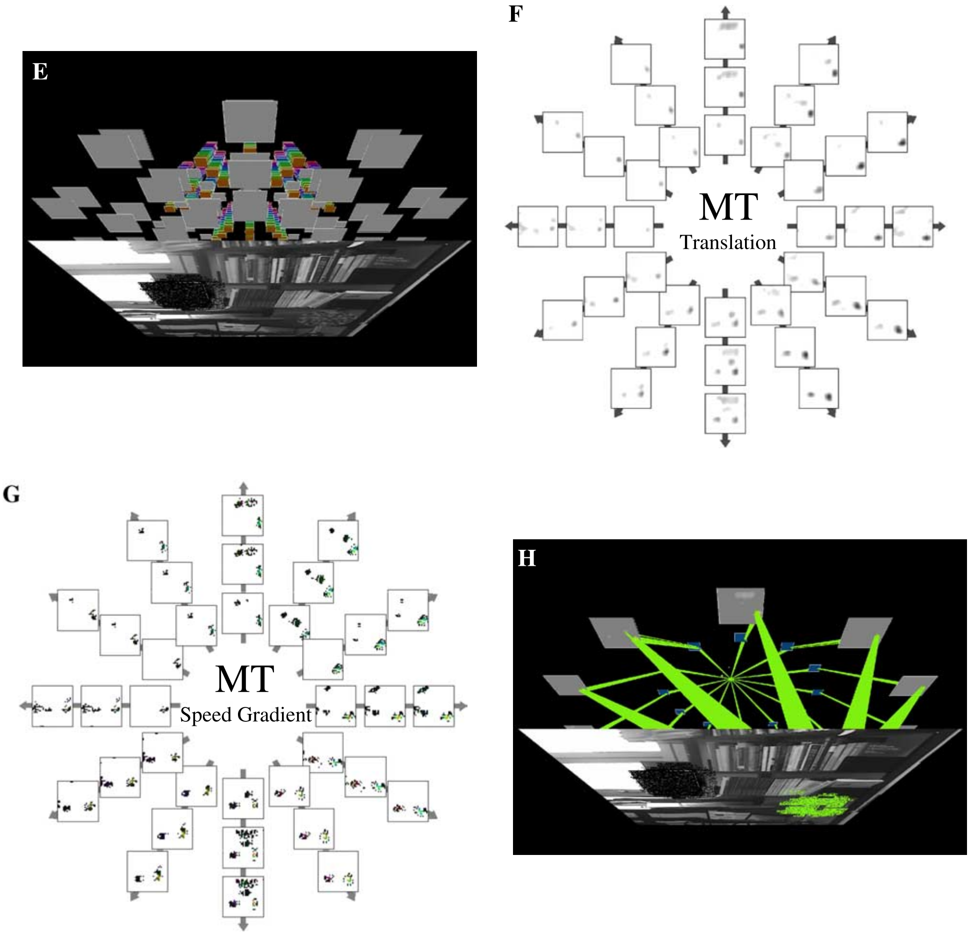

Fig 7. (*continued*)

Finally, an even more complex binding example would be one where objects are not spatially separated but rather overlap. An example is shown of two spatially overlapping motion patterns in Fig. 8 where two hexagonal disks are rotating one against the other. The two attentional fixations are shown.

### 4.2. Other motion types in the model

The model includes more motion types than just the ones described above. The methods of detecting onset and offset of events are included in the model and have been described previously [47,2]. This method has proven to be effective within any of the above representations. For example, assume an object is displaying a clockwise rotating motion, then stops and changes direction. The offset of the rotation signifies the end of the motion within the clockwise rotation representation. The onset within the counter-clockwise rotation representation would denote its beginning. Each of the representations above has a corresponding onset and offset computation and representation. It should be clear then that any motion processing model must include methods to detect the initiation and cessation of motions, yet most do not.

The previous examples were of motions generated by objects that did not change position in the image. Of course, objects in motion do change position (or camera motion induces such a motion) and motion occurs not only in pre-defined short image sequences but also continuously. To generate continuous output from continuous input, the straightforward solution might be to repeat the entire process of the algorithm in Section 3.1, that is, feed the network with time varying image sequences, perform the feed-forward computation and then feedback attentional selection. This idea that re-computing the pyramid and the beam for each image subset will definitely work, but is not efficient and has doubtful biological consistency due to the extra processing time required. A better strategy would be to modify the beam locally only as much as is needed to enable it to track the changes. The local beam structure could be defined in a dynamic manner, re-directed by new gating control signals at the top that propagate downwards [72].

In this approach, signals flow continuously through the pyramid. Input is continuous and flows upwards continuously. As WTA processes complete at the top, gating control signals begin their downward journey implementing the top-down WTA hierarchy. However, signals are potentially changing as this occurs and the selections made will not be the exact ones that led to the top winner, at least in precise location. Remember that the WTA as defined is guaranteed to find winners anywhere in the receptive field. The winner at the top is not restricted as a single location but rather can be a region. As long as motion speed and speed of signal propagation is matched, the system is blind to small location changes and operates correctly as expected. However the gating signals that would be generated at the top of the pyramid require time to propagate downwards and thus the control they may exert on lower levels of processing will reflect the past and not the present. This strategy features a time delay and a time period of ‘blindness’: discontinuities in motion that occur with a shorter duration than the propagation time are missed. This has been tested successfully on motion-defined object translation. The localization is rather simple because only the translation pyramidal stream is activated and the motion feature is uniform across the object.

## 5. Discussion

This paper makes several points: (1) it presents a new feed-forward motion processing hierarchy, (2) it presents examples of how the ST model can operate on this hierarchy to localize and label motion patterns, and (3) it shows how some aspects of recognition that require feature grouping (or binding) may be accomplished using a top-down attentional selection mechanism that does not depend on a single location-based saliency representation.

First, a new feed-forward motion analysis hierarchy is presented. The structure and computations are strongly inspired by biology, and the resulting network has a good degree of biological realism, although it is not biologically accurate in several ways. Due to the incorporation of functionally diverse neurons in the motion hierarchy, the output of the present model encompasses a wide variety of selectivities at different resolutions. This enables the computer simulation of the model to detect and classify various motion patterns in artificial and natural image sequences showing one or more moving objects as well as single objects undergoing complex, multiple motions. Most other models of biological motion perception focus on a single cortical area. For instance, the models by Zemel and Sejnowksi [10], Simoncelli and Heeger [4], and Beardsley and Vaina [5] are biologically relevant approaches that explain some specific functionality of MT or MST neurons, but do not include the embedding hierarchy in the motion pathway. On the other hand, there are hierarchical models for the detection of motion. Meese and Andersen [7] do not provide a computationally plausible version of the motion processing hierarchy. Giese and Poggio [6] describe a sophisticated, biologically motivated, and complex hierarchy for processing human movement patterns. However, they did not include any attentional influences. Further, they provide early input to their algorithm manually. Hand-tracked body joint positions were manually converted to stick figures where optic flow is easily computed. They cannot handle complex, overlapping, dense flow or discontinuous motions and certainly cannot process real image sequences directly. Lu and Sperling [73] present a motion hierarchy as well as attentive processes, but the model is not a computational one. However, it has strong biological plausibility in its function. They proposed that human visual motion perception is served by three separate motion systems: a first-order system that responds to moving luminance patterns, a second-order system that responds to moving modulations of feature types, and a third-order system that computes the motion of marked locations in a salience map. This third-order system of Lu and Sperling seems to be similar to the process of attending to motion in ST but without the computational details, it is difficult to draw too close a comparison.

Of course, this is only the beginning and we are actively pursuing several avenues of further work. The tuning characteristics of each of the neurons only coarsely model current knowledge of primate vision. The model includes little cooperative or competitive processing among units within a layer other than V1. Experimental work examining the relationship of this particular structure to human vision is also ongoing.

It is important to put this new motion analysis framework into context of classic literature on motion starting with Koenderink and Van Doorn [74] and Longuet-Higgins and Prazdny [75]. On the assumptions that a moving object can be modeled (at least piecewise) by rigid planar patches, and that there is a fixation point on the surface in question, the motion may be estimated using an affine transformation. The affine model can be described on the basis of four quantities: image translation, image rotation, divergence, and shear. The first two terms specify respectively a rigid 2D translation and rotation of the fixated object. The third term describes an isotropic expansion (or contraction) that specifies a change in scale or a pure deformation.

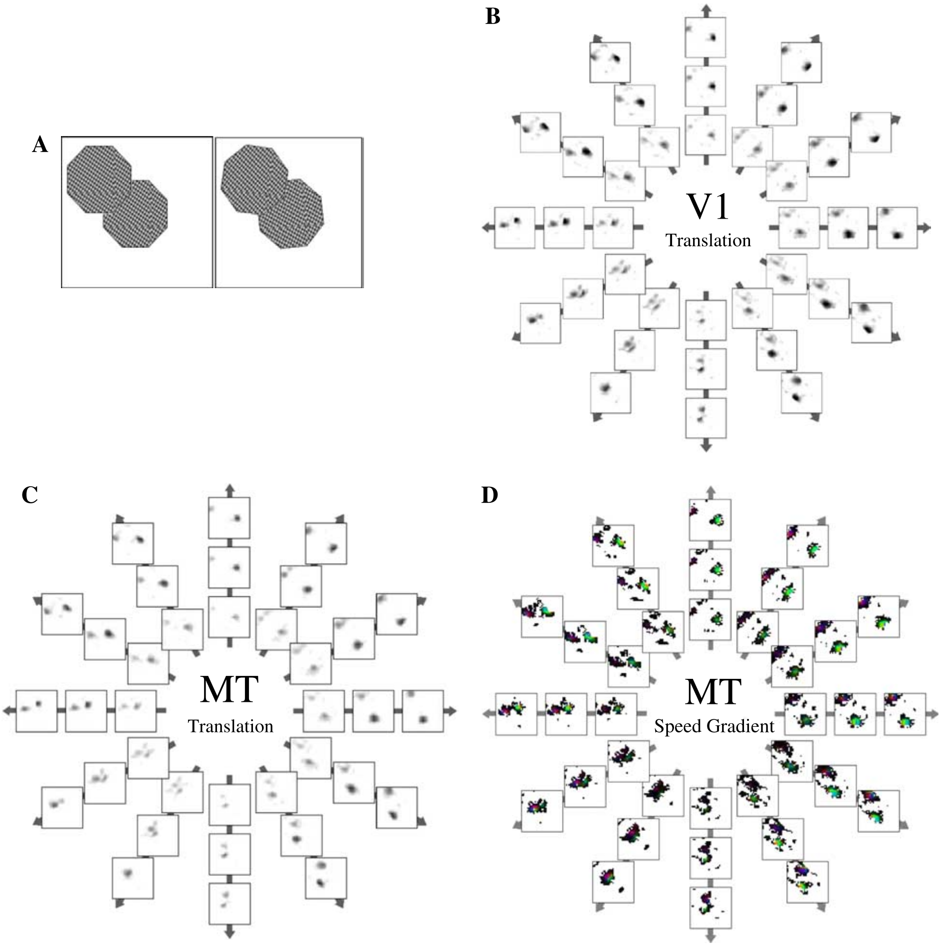

Fig. 8. Overlapping motion patterns—a feature binding example. (A) Two overlapping textured hexagons, the left one rotating clockwise while the right one rotates counter-clockwise. (B–J) The complete representation of all first pass feed-forward outputs is shown. (K) The two attentional fixations for the two hexagons. The lower one was attended first, localized quite well and labeled in green for counter-clockwise rotation, and the upper one second, localized well except for the overlap region and labeled in red for clockwise rotation. Note that even though the objects were overlapping, the motion labels were correct and the object localizations reasonable given the overlap. Note how the responses in area 7a are completely merged and no simple scheme could possibly disentangle this using only that output. However, the top-down search and feature binding strategy described here successfully separates the signals and localizes and labels the moving objects.

The shear term results in a distortion of the image pattern and corresponds to an expansion in a specified direction and a simultaneous contraction in the perpendicular one in such a way that the area of the pattern is preserved. Under perspective projection, the velocity vector at each point of an image is given by computing the derivative in time of the gray value changes at each pixel $(x, y)$ and yields $(u, v)$. This

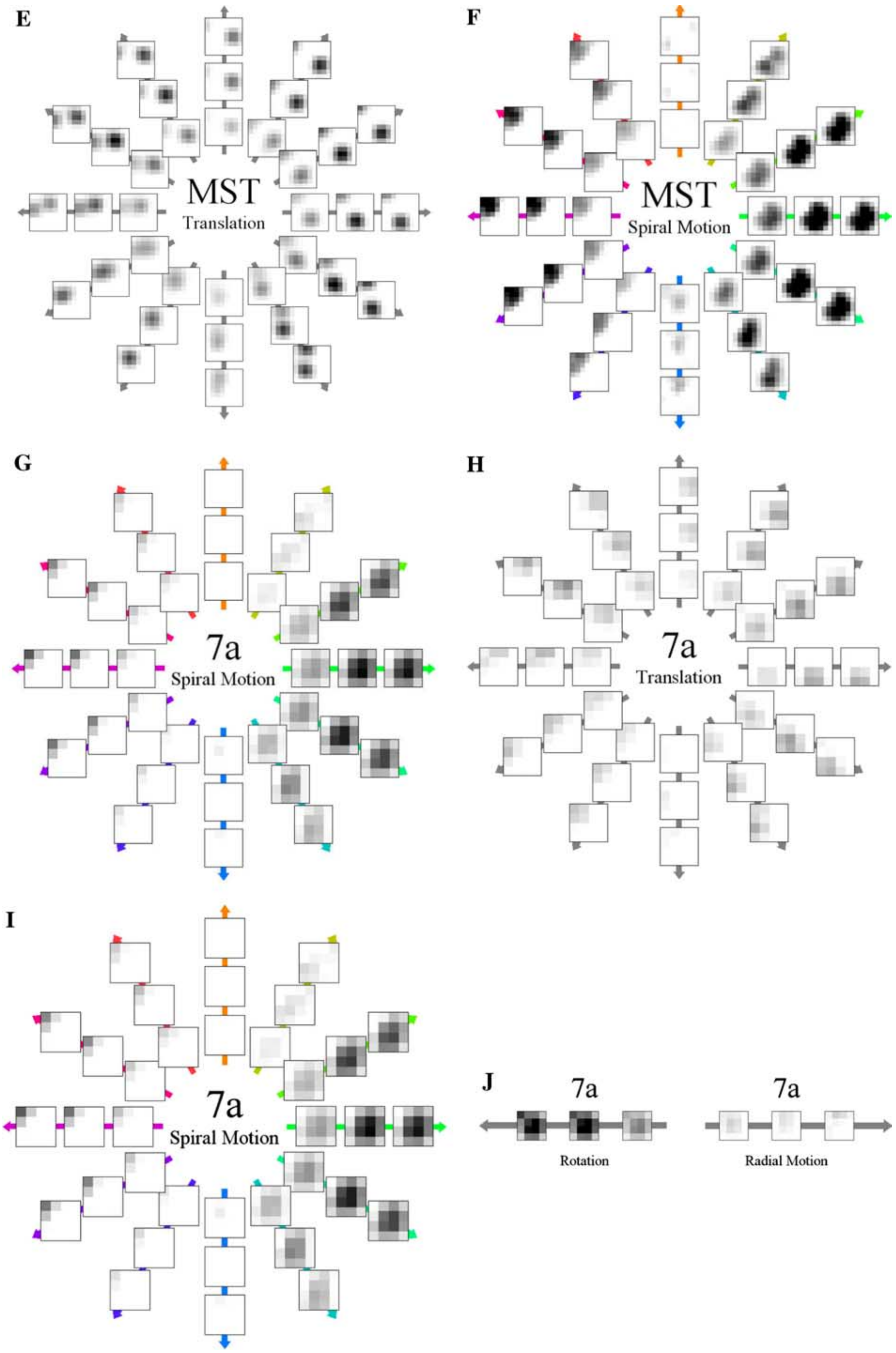

Fig 8. (*continued*)

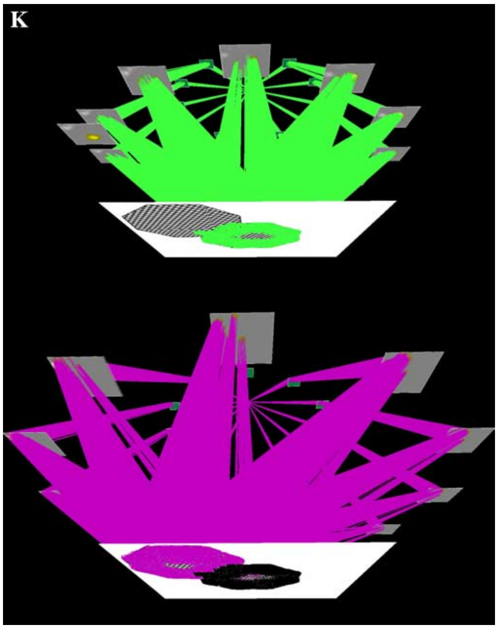

Fig 8. (*continued*)

represents translational motion. Spatial derivatives are then taken of each velocity component $u$ and $v$ in the $x$ and $y$ directions $(u_x, u_y, v_x, v_y)$. Combinations of these derivatives provide definitions of each of the remaining three affine motions, rotation, deformation, and divergence. Divergence is represented by $\mu = u_x + v_y$. Deformation or shear has two components $\rho$ and $\sigma$ and these are given by $\rho = u_x - v_y$ and $\sigma = u_y + v_x$. Finally, rotation is expressed as $\lambda = u_y - v_x$. The extraction of the affine estimates has essentially two components: the identification of an appropriate set of 2D spatial patches to represent each surface in a scene, and the tracking of the patches through the image sequence. The main point here is that the setting of a fixation point or the identification of the 2D patches to track is central to the definition and the majority if not all past uses of affine estimation make assumptions about where this fixation comes from. In this paper, we show a method for how it might be determined. More importantly, this is the first model to explicitly use local spatial derivates of velocity as an intermediate representation between local flow vectors and affine motion patterns. Evidence for such neurons was presented by Treue and Andersen [25]. Elsewhere, we show that this structure is biologically realistic by experimentally confirming that there are neural correlates in humans for the several layers of motion processing implied by these affine motion definitions [31].

Second, this paper shows that the earlier criticisms of the ST demonstrations, namely that the simple feature pyramids computed using Gaussian blurring were not biologically realistic nor useful, can now be forgotten. Although those criticisms were completely valid, they do not affect the original definition of ST, only those early demonstrations. As has been shown, ST operates perfectly well in this significantly more complex representation.

Another strength of our model is its mechanism of visual attention. To our knowledge, there are only three other computational motion models that address attention for motion. The earliest ones are due to Nowlan and Sejnowski [8] and Daniilidis [3]. In Nowlan and Sejnowski, processing is much in the same spirit as ours but very different in form. They compute motion energy with the goal of modeling MT neurons. This energy is part of a hierarchy of processes that include softmax for local velocity selection. They suggest that the selection permits processing to be focused on the most reliable estimates of velocity. There is neither top-down component nor a full processing hierarchy nor binding for complex patterns. Attentional modulation in motion neurons was described experimentally in [61] and appeared after their model was presented; thus, of course it is not developed and does not appear to be within the scope of their model. Based on the optical flow, Daniilidis computed 3D motion and structure. He fixated on an object to estimate ego motion in the presence of translation and rotation of the observer from the flow in the log-polar periphery. Computation of time to collision was a goal, not the definition of an attentive motion hierarchy. Although he used attentive fixations to advantage, the motion processing there was quite specific and based on log-polar representations and the connection to affine motion, implicated by the need to fixate, was not recognized. Finally, Grossberg et al. [9] present an integration and segmentation model for motion capture. Called the Formotion BCS model, their goal is to integrate motion information across the image and segment motion cues into a unified global percept. They employ models of translational processing in areas V1, V2, MT, and MST and do not consider motion patterns. Competition determines local winners among neural responses and the MST cells encoding the winning direction have an excitatory influence on MT cells tuned to the same direction. A variety of motion illusions are illustrated but no real image sequences are attempted. This model seems to be closest to ST here in goal and methodology. None of these models has the breadth of processing in the motion domain or in attentional selection as the current work.

An interesting comparison can be drawn between the performance of ST on this motion hierarchy and the population code strategy of neuronal representation. In the standard approach of population encoding, an assumption is made that there exists a unique and unambiguous value within the population that represents the stimulus (see [76] for review). A contrasting view put forward by Zemel and Dayan is that a population code can also contain additional information such as multiple values and uncertainty thus obviating the need for the restrictive assumption [77]. Is there a relationship of one or both of these approaches to the methods presented in this paper? The short answer is yes, there is a rather strong relationship at least at a qualitative level with the Zemel and Dayan approach. One of the examples presented earlier will be used to illustrate. The set of figures within Fig. 7 show the representations at each level of the hierarchy that arise due to two moving items within the visual field. Since there is no other clutter it will serve well to illustrate the representations for each item and how the overall population encodes these. First, the term 'population' must be clarified. In most population coding work, an assumption is made about where the neural responses come from and to what stimuli they respond. In Zemel and Dayan for example, responses from MT neurons in monkey are used; as is clear from the present paper, there is not only a single type of neuron in area MT. The Zemel and Dayan scheme does not separate the features, something that has been shown to have computational advantages [1]. Also, as the review by Pouget et al. shows, responses of behaving monkeys performing a task requiring attention are not considered. In other words, the population of neurons in the Zemel and Dayan work represent the combined result of all the feature maps of the first feed-forward pass as described in this paper, an unnecessarily large population. Population-coding strategies then, from this first pass encoding, attempt to remove noise and make inferences. ST employs attention for this task. The WTA process operates over a population in order to estimate the best response. This is an estimate because it is not necessarily the correct response. The next stage of processing, applying the top-down inhibitory surround around the selected location removes noise and permits a more precise representation of only the attended stimulus. A match of this result to a task representation then verifies or rejects the estimate. If there is more than one item in the visual field as in Fig. 7, the response populations due to each overlap as seen in the figure since the two peaks are readily visible. Importantly, Figs. 7F and G show the population code after the first attended stimulus has been removed by the attentive inhibition of return process. It has been sufficiently (but not completely) cleaned up so that the second peak in the population is correctly found. In summary, the distributional population-coding scheme of Zemel and Dayan and the ST scheme address very similar problems and accomplish very similar goals. The methods differ and the underlying assumption differs. That ST includes attention and the time course of attentional modulation within the process makes it a closer biological match. However, it would still be very interesting to do a more detailed analysis of the underlying mathematics for each to see where equivalences and differences can be found.

Inspection of the patterns of activation shown in the figures also provides some perspective on the issue of the location of 'the saliency map.' The search for a neural correlate to the concept of a saliency map has led to much interesting experimental work each providing evidence for one or another particular location (superior colliculus [78–80]; LGN [81,82]; V1 [83]; V1 and V2 [84]; pulvinar [85–87]; FEF [88]; parietal [89]). In each of these, the correlate is found by locating maxima of response within a neural population that corresponds to the attended location. Consider the patterns of response shown in any of the figures in this paper involving visual areas MST or higher. In each case, this criterion of a maximum corresponding to an attended location can be seen. In each area, both feed-forward and feedback influences take effect. Perhaps this is why evidence has been found in so many areas for the neural correlate to the saliency map? Maybe saliency is a distributed computation, as shown in this paper, and like attention itself, evidence reflecting these computations can be found in many, if not all, neural populations.

Finally, our attention strategy also demonstrates a key aspect of the recognition process, that is, the separate computation of parts and their subsequent re-assembly guided by attention. The key to this solution is the abandonment of the single, location-based saliency representation that supports a single point-based WTA, a feature of most other attention models. Even for those that do not use a single map, attention is an emergent, stochastic feature. Here, saliency is a local and distributed, deterministic phenomenon and the WTA processes are hierarchical, region-based, and dynamically defined depending on task and neural selectivity. Although it would not be entirely inappropriate to claim this is a solution to the classic binding problem, it is too early to justify this claim. The solution here, however, does appear to have the right elements to solve the limited aspects of binding required for this domain.

This strategy for attention to motion can be considered as a precursor to more detailed analysis in order to extract precise velocity and direction of motion. For example, in a computer vision application, a determination of precise velocity depends on good object localization and elimination of outliers. An attentive process such as that presented here could provide a first estimate of object location and extent as well as point out which type of motion is present. New algorithms for velocity extraction could be developed that take advantage of the reduced search space.

Attention to motion has not received sufficient study in computer vision even though it is a critical component of a general solution to visual information processing [90]. Significant enough advances have been made in early attentive algorithms to warrant a closer look at how current models of attention, such as ST, can be usefully directed at complex computer vision problems.

## Acknowledgments

A number of people have contributed to this effort in important ways and we wish to express our thanks for their assistance: Sean Culhane, Stefan Treue, Guy Orban, Albert Rothenstein, Winky Wai, Yuzhong Liu, and Michael Tombu. Funding for this research was gratefully received from the Natural Sciences and Engineering Council of Canada, Communications and Information Technology Ontario, a Province of Ontario Centre of Excellence, and from the Institute for Robotics and Intelligent Systems, A Network of Centers of Excellence of the Government of Canada through grants to J.K.T. J.K.T. holds the Canada Research Chair in Computational Vision. J.C.M.-T. holds the Canada Research Chair in Neuroscience.

## References

[1] J.K. Tsotsos, Analyzing vision at the complexity level, Behav. Brain Sci. 13-3 (1990) 423–445.

[2] J.K. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning, Artif. Intell. 78 (1–2) (1995) 507–547.

[3] K. Daniilidis, Attentive visual motion processing: computations in the log-polar plane, Computing (Suppl.) 11 (1995) 1–20.

[4] E.P. Simoncelli, D.J. Heeger, A model of neuronal responses in visual area MT, Vision Res. 38 (5) (1998) 743–761.

[5] S.A. Beardsley, L.M. Vaina, Computational modeling of optic flow selectivity in MSTd neurons, Comput. Neural Syst. 9 (1998) 467–493.

[6] M.A. Giese, T. Poggio, Neural mechanisms for the recognition of complex movements and actions, Nat. Rev. Neurosci. 4 (2003) 179–192.

[7] T.S. Meese, S.J. Anderson, Spiral mechanisms are required to account for summation of complex motion components, Vision Res. 42 (2002) 1073–1080.

[8] S.J. Nowlan, T.J. Sejnowski, A selection model for motion processing in area MT of primates, J. Neurosci. 15 (2) (1995) 1195–1214.

[9] S. Grossberg, E. Mingolla, L. Viswanathan, Neural dynamics of motion integration and segmentation within and across apertures, Vision Res. 41 (2001) 2521–2553.

[10] R.S. Zemel, T.J. Sejnowski, A model for encoding multiple object motions and self-motion in area MST of primate visual cortex, J. Neurosci. 18 (1) (1998) 531–547.

[11] C. Pack, S. Grossberg, E. Mingolla, A neural model of smooth pursuit control and motion perception by cortical area MST, J. Cognit. Neurosci. 13 (1) (2001) 102–120.

[12] J.A. Perrone, L.S. Stone, Emulating the visual receptive field properties of MST neurons with a template model of heading estimation, J. Neurosci. 18 (1998) 5958–5975.

[13] M. Pomplun, Y. Liu, J. Martinez-Trujillo, E. Simine, J.K. Tsotsos, A neurally-inspired model for detecting and localizing simple motion patterns in image sequences, in: Proc. 4th Workshop on Dynamic Perception, Bochum, Germany, November 14–15, 2002, pp. 47–52.

[14] J.K. Tsotsos, M. Pomplun, Y. Liu, J. Martinez-Trujillo, E. Simine, Attending to Motion: Localizing and Labeling Simple Motion Patterns in Image Sequences, Lecture Notes in Computer Science, vol. 2525, Springer-Verlag Berlin, 2002, pp. 439–452.

[15] J.K. Tsotsos, An inhibitory beam for attentional selection, in: L. Harris, M. Jenkin (Eds.), Spatial Vision in Humans and Robots, Cambridge University Press, Cambridge UK, 1993, pp. 313–331.

[16] J.K. Tsotsos, Towards a computational model of visual attention, in: T. Papathomas, C. Chubb, A. Gorea, E. Kowler (Eds.), Early Vision and Beyond, MIT Press/Bradford Books, 1995, pp. 207–218.

[17] J.K. Tsotsos, Complexity, vision and attention, in: L. Harris, M. Jenkin (Eds.), Vision and Attention, Springer-Verlag, New York, 2001, pp. 105–128.

[18] J.K. Tsotsos, S. Culhane, F. Cutzu, From theoretical foundations to a hierarchical circuit for selective attention, in: J. Braun, C. Koch, J. Davis (Eds.), Visual Attention and Cortical Circuits, MIT Press, Cambridge MA, 2001, pp. 285–306.

[19] D. Felleman, D. Van Essen, Distributed hierarchical processing in the primate visual cortex, Cereb. Cortex 1 (1991) 1–47.

[20] L. Lagae, S. Raiguel, G.A. Orban, Speed and direction selectivity of Macaque middle temporal neurons, J. Neurophysiol. 69 (1) (1993) 19–39.

[21] G.A. Orban, H. Kennedy, J. Bullier, Velocity sensitivity and direction sensitivity of neurons in areas V1 and V2 of the monkey: influence of eccentricity, J. Neurophysiol. 56 (2) (1986) 462–480.

[22] G.A. Orban, L. Lagae, S. Raiguel, D. Xiao, H. Maes, The speed tuning of medial superior temporal (MST) cell responses to optic-flow components, Perception 24 (3) (1995) 269–285.

[23] R.M. Siegel, H.L. Read, Analysis of optic flow in the monkey parietal area 7a, Cereb. Cortex 7 (1997) 327–346.

[24] S. Sunaert, P. Van Hecke, G. Marchal, G.A. Orban, Motion-responsive regions of the human brain, Exp. Brain Res. 127 (4) (1999) 355–370.

[25] S. Treue, R.A. Andersen, Neural responses to velocity gradients in macaque cortical area MT, Vis. Neurosci. 13 (4) (1996) 797–804.

[26] C.J. Duffy, R.H. Wurtz, MST neurons respond to speed patterns in optic flow, J. Neurosci. 17 (8) (1997) 2839–2851.

[27] D.J. Felleman, J.H. Kaas, Receptive field properties of neurons in middle temporal visual area (MT) of owl monkeys, J. Neurophysiol. 52 (1984) 488–513.

[28] M.S. Graziano, R.A. Andersen, R.J. Snowden, Tuning of MST neurons to spiral motions, J. Neurosci. 14 (1) (1994) 54–67.

[29] D.C. Van Essen, J.H. Maunsell, J.L. Bixby, The middle temporal visual area in the macaque: myeloarchitecture, connections, functional properties and topographic organization, J. Comp. Neurol. 199 (3) (1981) 293–326.

[30] D.J. Heeger, Optical flow using spatiotemporal filters, Int. J. Comput. Vision 1 (4) (1988) 279–302.

[31] J.C. Martinez-Trujillo, J.K. Tsotsos, E. Simine, M. Pomplun, R. Wildes, S. Treue, H.-J. Heinze, J.-M. Hopf, Selectivity for speed gradients in human area MT/V5, NeuroReport 16 (5) (2005) 435–438.

[32] J.K. Tsotsos, A 'complexity level' analysis of vision, in: Proc. 1st Internat. Conf. on Computer Vision London, England, 1987, pp. 346–355.

[33] J.K. Tsotsos, The complexity of perceptual search tasks, in: Proc. Internat. Joint Conf. on Artificial Intelligence, Detroit, 1989, 1571–1577.

[34] J.K. Tsotsos, On the relative complexity of passive vs active visual search, Int. J. Comput. Vision 7-2 (1992) 127–141.

[35] A. Zaharescu, A. Rothenstein, J.K. Tsotsos, Towards a biologically plausible active visual search model, in: Proc. ECCV WAPCV 2004, Lecture Notes in Computer Science, vol. 3368, Springer-Verlag Berlin, 2005, pp. 133–147.

[36] S. Thorpe, Ultra-rapid scene categorisation with a wave of spikes, in: H.H. Bulthoff et al. (Eds.), Biologically Motivated Computer Vision, Lecture Notes in Computer Science, 2525, Springer-Verlag, Berlin, 2002, pp. 1–15.

[37] M. Riesenhuber, T. Poggio, Are cortical models really bound by the 'binding problem'?, Neuron 24 (1999) 87–93.

[38] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum. Neurobiol. 4 (1985) 219–227.

[39] T.S. Lee, C. Yang, R.D. Romero, D. Mumford, Neural activity in early visual cortex reflects behavioral experience and higher order perceptual saliency, Nat. Neurosci. 5 (6) (2002) 589–597.

[40] L. Valiant, Parallelism in comparison problems, SIAM J. Comput. 4 (3) (1975) 348–355.

[41] A. Yuille, D. Geiger, Winner-take-all mechanisms, in: M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, MA, 1998, pp. 1056–1060.

[42] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[43] J. Wolfe, K. Cave, S. Franzel, Guided search: an alternative to the feature integration model for visual search, J. Exp. Psychol.: Hum. Percept. Perform. 15 (1989) 419–433.

[44] F.H. Hamker, A dynamic model of how feature cues can guide spatial attention, Vision Res. 44 (2004) 501–521.

[45] G. Deco, J. Zihl, Top-down selective visual attention: a neurodynamical approach, Visual Cogn. 8 (2001) 119–140.

[46] S. Culhane, J.K Tsotsos, An attentional prototype for early vision, in: G. Sandini (Ed.), Proc. 2nd European Conf. on Computer Vision, Santa Margherita Ligure, Italy, LNCS-Series, vol. 588, Springer-Verlag Berlin, 1992, pp. 551–560.

[47] W. Wai, J.K. Tsotsos, Directing attention to onset and offset of image events for eye-head movement control, in: Proc. IAPR Conf. on Pattern Recognition, Jerusalem, vol. A, 1994, pp. 274–279.

[48] G. Caputo, S. Guerra, Attentional selection by distractor suppression, Vision Res. 38 (5) (1998) 669–689.

[49] D.O. Bahcall, E. Kowler, Attentional interference at small spatial separations, Vision Res. 39 (1) (1999) 71–86.

[50] W. Vanduffel, R.B.H. Tootell, G.A. Orban, Attention-dependent suppression of metabolic activity in the early stages of the macaque visual system, Cereb. Cortex 10 (2) (2000) 109–126.

[51] F. Cutzu, J.K. Tsotsos, The selective tuning model of visual attention: testing the predictions arising from the inhibitory surround mechanism, Vision Res. 43 (2003) 205–219.

[52] J.D. Schall, T. Sato, K. Thompson, A. Vaughn, J. Chi-Hung, Effects of search efficiency on surround suppression during visual selection in frontal eye field, J. Neurophysiol. 91 (2004) 2765–2769.

[53] J. Mounts, Attentional capture by abrupt onsets and feature singletons produces inhibitory surrounds, Percept. Psychophys. 62 (2000) 1485–1493.

[54] M. Tombu, J.K. Tsotsos, Attention to orientation results in an inhibitory surround in orientation space, in: Proc. 14th Annual Meeting of Behaviour, Brain and Cognitive Science Society, June 12–14, 2004, St. John's Newfoundland, http://www.science.mcmaster.ca/~BBCS/2004/viewabstract.php?id=170.

[55] N. Muller, A. Kleinschmidt, The attentional 'spotlight's' penumbra: center-surround modulation in striate cortex, Neuroreport 15 (6) (2004) 977–980.

[56] D. O'Connor, M. Fukui, M. Pinsk, S. Kastner, Attention modulates responses in the human lateral geniculate nucleus, Nat. Neurosci. 5 (11) (2002) 1203–1209.

[57] A.D. Mehta, I. Ulbert, C.E. Schroeder, Intermodal selective attention in monkeys. I: distribution and timing of effects across visual areas, Cereb. Cortex 10 (4) (2000) 343–358.

[58] S. Kastner, P. De Weerd, R. Desimone, L. Ungerleider, Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI, Science 282 (1998) 108–111.

[59] K.H. Britten, Attention is everywhere, Nature 382 (6591) (1996) 497–498.

[60] V. Mountcastle, R. Andersen, B. Motter, The influence of attentive fixation upon the excitability of the light-sensitive neurons off the posterior parietal cortex, J. Neurosci. 1 (1981) 1218–1225.

[61] S. Treue, J.H.R. Maunsell, Attentional modulation of visual motion processing in cortical areas MT and MST, Nature 382 (1996) 539–541.

[62] S. Treue, J.C. Martinez-Trujillo, Feature-based attention influences motion processing gain in macaque visual cortex, Nature 399 (6736) (1999) 575–579.

[63] A. Roskies (Ed.), Neuron, vol. 24, 1999.

[64] A. Roskies, The binding problem—introduction, Neuron 24 (1999) 7–9.

[65] F. Rosenblatt, Principles of Neurodynamics: Perceptions and the Theory of Brain Mechanisms, Spartan Books, Washington, DC, 1961.

[66] A. Treisman, H. Schmidt, Illusory conjunctions in the perception of objects, Cognit. Psychol. 14 (1982) 107–141.

[67] G. Ghose, J. Maunsell, Specialized representations review in visual cortex: a role for binding? Neuron 24 (1999) 79–85.

[68] C. von der Malsburg, The what and why of binding: review the modeler's perspective, Neuron 24 (1999) 95–104.

[69] H.B. Barlow, Single units and cognition: a neurone doctrine for perceptual psychology, Perception 1 (1972) 371–394.

[70] A. Treisman, G. Gelade, A feature-integration theory of attention, Cognit. Sci. 12 (1980) 99–136.

[71] D. Felleman, Y. Xiao, E. McClendon, Modular organization of occipito-temporal pathways: cortical connections between visual area 4 and visual area 2 and posterior inferotemporal ventral area in macaque monkeys, J. Neurosci. 17 (9) (1997) 3185–3200.

[72] K. Zhou, Modeling Motion with the Selective Tuning Model, MSc. Thesis, Dept. of Computer Science, York University, Toronto, Canada, 2004.

[73] Z.L. Lu, G. Sperling, The functional architecture of human visual motion perception, Vision Res. 35 (19) (1995) 2697–2722.

[74] J.J. Koenderink, A.J. van Doorn, Local structure of movement parallax of the plane, J. Opt. Soc. Am. 66 (1976) 717–723.

[75] H.C. Longuet-Higgins, K. Prazdny, The interpretation of a moving retinal image, Proc. Royal Soc. London B 208 (1173) (1980) 385–397.

[76] A. Pouget, P. Dayan, R. Zemel, Inference and computation with population codes, Annu. Rev. Neurosci. 26 (2003) 381–410.

[77] R. Zemel, P. Dayan, Distributional population codes and multiple motion models, NIPS-11: Adv. Neural Inform. Process. Syst. 11 (1999) 174–180.

[78] A.A. Kustov, D.L. Robinson, Shared neural control of attentional shifts and eye movements, Nature 384 (1996) 74–77.

[79] R.M. McPeek, E.L. Keller, Saccade target selection in the superior colliculus during a visual search task, J. Neurophysiol. 88 (2002) 2019–2034.

[80] G.D. Horwitz, W.T. Newsome, Separate signals for target selection and movement specification in the superior colliculus, Science 284 (1999) 1158–1161.

[81] C. Koch, A theoretical analysis of the electrical properties of an X-cell in the cat's LGN: does the spine-triad circuit subserve selective visual attention? Artif. Intell. Memo 787, MIT, Artificial Intelligence Laboratory, February, 1984..

[82] S.M. Sherman, C. Koch, The control of retinogeniculate transmission in the mammalian lateral geniculate nucleus, Exp. Brain Res. 63 (1986) 1–20.

[83] Z. Li, A saliency map in primary visual cortex, Trends Cognit. Sci. 6 (1) (2002) 9–16.

[84] D.K. Lee, L. Itti, C. Koch, J. Braun, Attention activates winner-take-all competition among visual filters, Nat. Neurosci. 2 (4) (1999) 375–381.

[85] S.E. Petersen, D.L. Robinson, J.D. Morris, Contributions of the pulvinar to visual spatial attention, Neuropsychologia 25 (1987) 97–105.

[86] M.I. Posner, S.E. Petersen, The attention system of the human brain, Annu. Rev Neurosci. 13 (1990) 25–42.

[87] D.L. Robinson, S.E. Petersen, The pulvinar and visual salience, Trends Neurosci. 15 (4) (1992) 127–132.

[88] K.G. Thompson, N.P. Bichot, J.D. Schall, Dissociation of visual discrimination from saccade programming in macaque frontal eye field, J. Neurophysiol. 77 (1997) 1046–1050.

[89] J. Gottlieb, M. Kusunoki, M.E. Goldberg, The representation of visual salience in monkey posterior parietal cortex, Nature 391 (1998) 481–484.

[90] J.K. Tsotsos, Motion understanding: task-directed attention and representations that link perception with action, Int. J. Comput. Vision 45 (3) (2001) 265–280.