Attending to Visual Motion: Localizing and Classifying Affine Motion Patterns [1]

John K. Tsotsos[1], Marc Pomplun[2], Julio C. Martinez-Trujillo[1], Kunhao Zhou[1]

[1]Centre for Vision Research, York University, Toronto, Canada M3J 1P3
[2]Department of Computer Science, University of Massachusetts at Boston, Boston, MA 02125, USA

Abstract. The Selective Tuning Model is a proposal for modelling visual attention in primates and humans. This paper describes ongoing research to include attention to motion stimuli within the model. The effort is unique because it seems that no past model presents a motion hierarchy plus attention to motion. We propose a biologically realistic model of the primate visual motion system attempting to explain how a hierarchical feed-forward network consisting of layers representing cortical areas V1, MT, MST, and 7a detects and classifies different kinds of motion patterns. The STM model is then integrated into this hierarchy demonstrating that successfully attending to motion patterns results in localization (segmentation) and labeling of those patterns.

1 Introduction

Attentive processing is a largely unexplored dimension of the computational motion field. No matter how sophisticated the methods become for extracting motion information from image sequences, it will not be possible to achieve the goal of human-like performance without integrating the optimization of processing that attention provides. Virtually all past surveys of computational models of motion processing completely ignore attention. However, the concept has crept into work in a variety of ways.

One can survey the current computer vision literature and realize that attentive processing is not much of a concern. Many recent reviews of various aspects of motion understanding have not made any mention of attentive processing of any kind [1, 2, 3, 4, 5]. The majority of authors do not consider attention simply because of assumptions that eliminate the issue. For example, algorithms start off by being given the regions of interest that correspond to each object that may be moving. No matter how high the quality of the ensuing process, the algorithms critically depend on reasonable regions of interest and are not designed to find that region of interest either independently or concurrently as they process the events in the scene. At best, semi-automatic methods find regions using statistical learning methods or hand-coded knowledge of scene contents, appearance or geometry. The problem is not that any one effort makes these assumptions; the problem lies in the fact that it is now almost universal to assume the unreasonable.

Attentive components have been included in systems not only through assumptions. At least three tools have appeared: the detection of salient tracking points/structures; search region predictions; and, Kalman filters and their extensions. Many examples have appeared [6, 7, 8, 9]. All are clearly strategies that help reduce search however, the overall result is an ad hoc collection of domain-specific methods.

A similar survey of computational neuroscience literature reveals many interesting motion models and better interest in motion attention. There, one finds a couple of proposals with similar goals, but none being demonstrated on real image sequences nor concurrently showing localization and labeling.

2 The Selective Tuning Model

Complexity analysis leads to the conclusion that attention must tune the visual processing architecture to permit task-directed processing [10]. In its original definition, [10], the Selective Tuning Model (STM), selection takes two forms: *spatial* selection is realized by inhibiting task-irrelevant locations in the neural network, and *feature* selection is realized by inhibiting the neurons that represent task-irrelevant features. When task constraints are available they are used to set priorities for selection; if not available, then there are default priorities (such as 'strongest response'). The two cornerstones of spatial and feature selection have since been experimentally supported [11, 12]. Only a brief summary is presented here since the model is detailed elsewhere [13].

The spatial role of attention in the image domain is to localize a subset of the input image and its path through the processing hierarchy such as to minimize any interfering or corrupting signals. The visual processing architecture is a pyramidal network composed of units receiving both feed-

[1] The version of this paper with colour figures may be downloaded at http://www.cs.yorku.ca/~tsotsos.

Proceedings of the First Canadian Conference on Computer and Robot Vision (CRV'04)
0-7695-2127-4/04 $20.00 © 2004 IEEE

forward and feedback connections. When a stimulus is first applied to the input layer of the pyramid, it activates in a feed-forward manner all of the units within the pyramid to which it is connected. The result is the activation of an inverted sub-pyramid of units whose degree of unit activation reflects the goodness-of-match between the unit and the stimulus it represents.

Attentional selection relies on a hierarchy of winner-take-all (WTA) processes. WTA is a parallel algorithm for finding the maximum value in a set of variables, which was first proposed in this context by Koch and Ullman [14]. WTA can be steered to favor particular stimulus locations or features but in the absence of such guidance it operates independently. The processing of a visual input involves three main stages. During the first stage, a stimulus is applied to the input layer and activity propagates along feed-forward connections towards the output layer. The response of each unit depends on its particular selectivities, and perhaps also on a top-down bias for task-relevant qualities. During the second stage, a hierarchy of WTA processes is applied in a top-down, coarse-to-fine manner. The first WTA process operates in the top layer and covers the entire visual field: it computes the unit or groups of contiguous units with the largest response in the output layer, that is, the *global winner*. In turn, the global winner activates a WTA amongst its input units in the layer immediately below. This localizes the largest response within the receptive field (RF) of the global winner. All of the connections of the visual pyramid that do not contribute to the winner are pruned (i.e., attenuated). This strategy of finding the winner within each receptive field and then pruning away irrelevant connections, is applied recursively through the pyramid, layer by layer. Thus, the global winner in the output layer is eventually traced back to its perceptual origin in the input layer. The connections that remain (i.e., are not pruned) may be considered the *pass zone* of the attentional beam, while the pruned connections an *inhibitory zone* around that beam. A final feed-forward pass then allows the selected stimulus to be processed by the network without signal interference from surrounding stimuli. This constitutes a single attentive processing cycle.

The processing exhibits serial search for scenes with multiple objects using a simple inhibition of return mechanism, that is, the pass zone pathways are inhibited for one processing cycle so that in the next feed-forward pass the second strongest responses form the global winner and the WTA hierarchy focuses in on the second strongest item in the display. The processing operates continuously in this manner.

The selective tuning model was developed with the dual goals of computational utility and biological predictive power. The predictions (appearing mostly in [10, 13]) and supporting evidence are briefly described:

- An early prediction was that attention is necessary at any level of processing where a many-to-one mapping between neural processes is found. Further, attention occurs in all the areas in a coordinated manner. The prediction was made at a time when good evidence for attentional modulation was known for area V4 only [15]. Since then, attentional modulation has been found in many other areas both earlier and later in the visual processing stream, and that it occurs in these areas simultaneously [16]. Vanduffel et al. [17] have shown that attentional modulation appears as early as the LGN. The prediction that attention modulates all cortical and even subcortical levels of processing has been borne out by recent work from several groups [17, 18, 19]:
- The notions of competition between stimuli and of attentional modulation of this competition were also early components of the model and these too have gained substantial support over the years [11, 16, 21].
- The model predicts an inhibitory surround that impairs perception around the focus of attention, a prediction that has strong support both psychophysically and neurophysiologically [17, 18, 22, 22, 24, 25].
- STM predicts that the latency of attentional modulations *decreases* from lower to higher visual areas. Although controversial, it seems that attentional effects do not appear until around 150 ms after the onset of a stimulus in IT cortex [26] while in V1 they appear after 230 ms [27].

Additional predictions of the selective tuning model concern the form of spatial and temporal modulations of visual cortical responses around the focus of attention, and the existence of a WTA circuit connecting cortical columns of similar selectivity. The selective tuning model offers a principled solution to the fundamental problems of visual complexity, a detailed perceptual account of both the guidance and the consequences of visual attention, and a neurally plausible implementation as an integral part of the visual cortical hierarchy. Thus, the model "works" at three distinct levels, computational, perceptual, and neural, offering a more concrete account and more specific predictions than previous models limited to one of these levels.

Previous demonstrations of STM were not without their weaknesses. The main one addressed here is that the levels of representation shown in [13] were not biologically plausible. The motion domain is chosen to demonstrate that STM has no difficulty with realistic representations. In addition, the effort is unique because it seems that no past

model presented a motion hierarchy plus attention to motion [28, 29, 30, 31, 32, 33, 34, 35, 36]. The remainder of this paper will focus on this issue.

3 The Feed-Forward Motion Pyramid

We propose a biologically realistic model of the primate motion processing hierarchy. The model aims to explain how a hierarchical feed-forward network consisting of neurons in the cortical areas V1, MT, MST, and 7a of primates detects and classifies different kinds of motion patterns. At best, the motion model is a first-order one with much elaboration left for future work. Indeed, some of the previous motion models offer better sophistication at one or another level of processing; however, none cover all these levels and incorporate selective attentional processes. The primary goal is to demonstrate that the STM functions not only as previously demonstrated on Gaussian pyramids but also on a more biologically realistic representation. The details of the various filters and other model aspects are in ([13], [48]).

Following [46], Longuet-Higgins and Prazdny [47] provide a classic definition of an affine model. Under perspective projection, the velocity vector at each point (x, y) of an image is given as the temporal derivative of spatial position, i.e., $(u, v) = (dx/dt, dy/dt)$. This 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 affine motions. Divergence is represented by $\mu = u_x + v_y$. Shear has two components ρ and σ and these are given by $\rho = u_x - v_y$ and $\sigma = u_y + v_x$. Finally, rotation is expressed as $\lambda = u_y - v_x$. Longuet-Higgins and Prazdny went a step further suggesting that perhaps biological vision systems included the same processing channels. This suggestion has received some neurobiological support recently ([41], [49]) but until now has not been considered by the computer vision community. The representation of motion presented here is based on this suggestion and biological evidence.

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 [37]). The functionality of layer V1 is realized by two types of artificial neurons, those performing spatiotemporal filtering and those integrating local filter unit activations. 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 RF of a neuron is oriented in such a way that local motion at its position in direction a and with speed s_v would induce constant intensity across the RF. V1 consists of neurons of three distinct speed selectivity types (following [37]): type 1 (low speed), type 2 (medium speed), type 3 (high speed). In the model, these neurons are implemented with three different preferred speed bands, which were centred at $s_1 = 0.5$ pixels/sec, $s_2 = 1$ pixel/sec, and $s_3 = 2$ pixels/sec. To limit the computational complexity of the model, only 12 different preferred directions were realized ($a = 0°, 30°, \ldots, 330°$), although it is known that a wider range of preferred directions exist in area V1 of macaques [43].

To achieve the V1 computation, the model uses one hypercolumn of spatio-temporal filter 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 x 64 evenly distributed integrative neuron 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 x 256 pixels and integration units with RFs covering eight by eight neighbouring filter units are used, thereby creating substantial overlap of RFs. The 64 x 64 hypercolumns of integration units provide the input for the model's MT neurons. In Figure 1 the two layers of V1 computation are distinguished by the colour grey for the filter neurons and blue for the integrative neurons.

In area MT a high proportion of cells are tuned for a particular local speed and direction of movement, similar to direction and speed selective cells in V1 [39, 40]. A proportion of MT neurons are also selective for a particular angle between local movement direction and spatial speed gradient [41]. Movement is quantized into 12 directions and thus 12 neurons computing the spatial gradient of local velocity in each direction (and at each of three speeds). Both types of neurons are represented in the MT layer of the model, which is a 30x30 array of hypercolumns. Each MT cell receives input from a 4x4 field of V1 neurons with the same direction and speed selectivity.

Neurons in area MST are tuned to motion patterns: expand or approach, contract or recede, or rotation, with RFs covering most of the visual field [42, 43]. Two types of neurons are modeled: one type selective for translation (as in V1) and another type selective for spiral motion (clockwise and counterclockwise rotation, expansion, contraction and combinations). MST is simulated as a 5x5 array of hypercolumns. Each MST cell receives input from a large group (covering 60% of the visual field) of MT neurons that respond to a particular motion/gradient angle. Any coherent

motion/gradient angle indicates a particular type of spiral motion.

Finally, area 7a seems to involve at least four different types of computations [44]. Here, neurons are selective for translation and spiral motion as in MST, but they have even larger RFs. They are also selective for rotation (regardless of direction) and radial motion (regardless of direction). In the simulation, area 7a is represented by a 4x4 array of hypercolumns. Each 7a cell receives input from a 4x4 field of MST neurons that have the relevant tuning. Rotation cells and radial motion cells only receive input from MST neurons that respond to spiral motion involving any rotation or any radial motion, respectively.

Figure 1A 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. Grey rectangles represent translational motion of a particular speed and direction. Coloured rectangles represent speed gradient with respect to local motion (the grey arrow) in area MT, while in MST they represent generalized spiral motion, both coded using the colour wheel of part B of this figure. 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, direction and speed selective filters are represented by the sheet of grey rectangles while the integrative neurons in V1 are the blue sheet. 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 in Figure 1B represents the speed gradient coding with respect to local motion, in this case the larger grey arrow pointing upwards. There are a total of 690 filter planes in this model. 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.

4 An Example

The following figures present a simple illustration of the performance of the model. The input sequence has background formed by random noise in which a square, again filled with random noise, rotates counterclockwise; in other words the square is entirely motion-defined, there is no border or texture that differs from the background. The illustration will show the feed-forward output of each of the visual areas in turn, and then the final attended and labeled region.

Figure 2 gives the output of area V1 computations of the integrative units. Figure 3 showing the translation responses and Figure 4 showing the gradient responses. The gradient responses show only the largest value across all 12 directions at each pixel, color-coded using the scheme shown in Fig. 1. Part A of the figure shows the full set of responses while Part B shows the expanded view of the rightmost three directions, in order to show response details. Note how sparse the responses are and moreover how most, if final choices made this stage, would be incorrect. Figure 5 gives the output of two types of selectivity represented in MST, translation and spiral (using the color coding of Fig. 1). Figure 6 illustrates the output of area 7a computations.

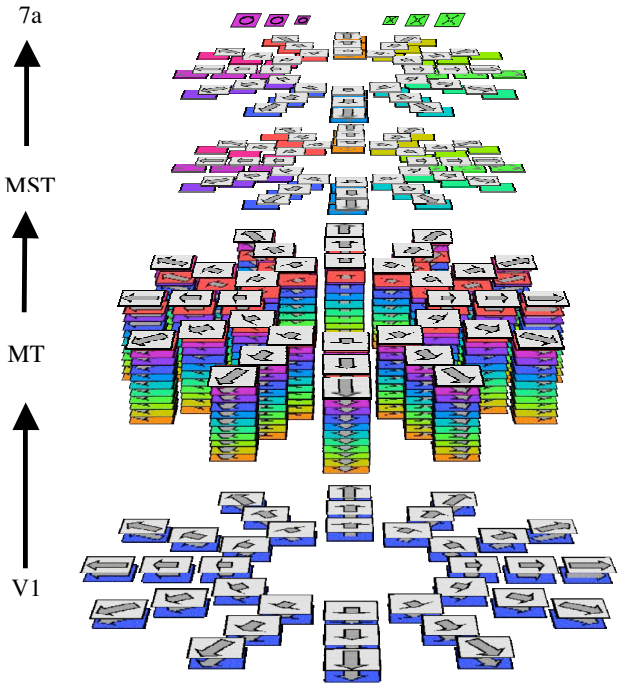


Figure 1A

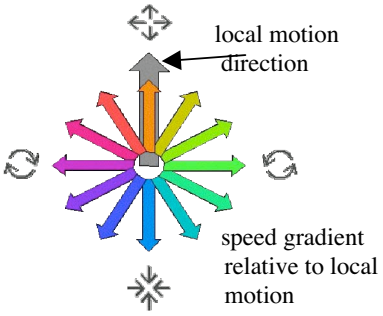


Figure 1B

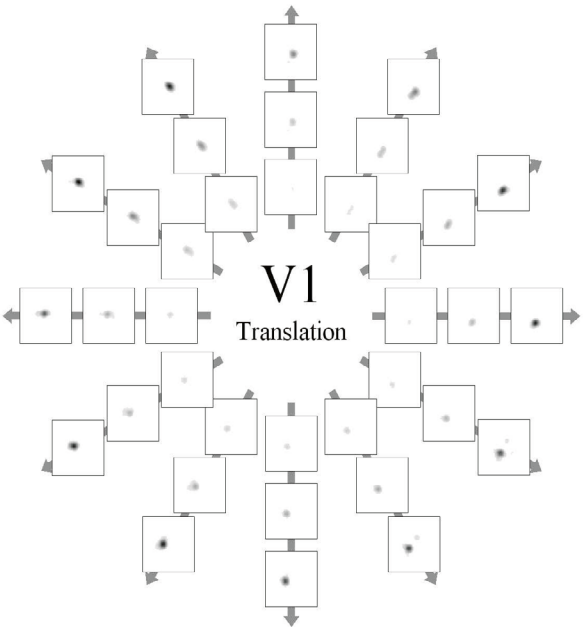


Figure 2

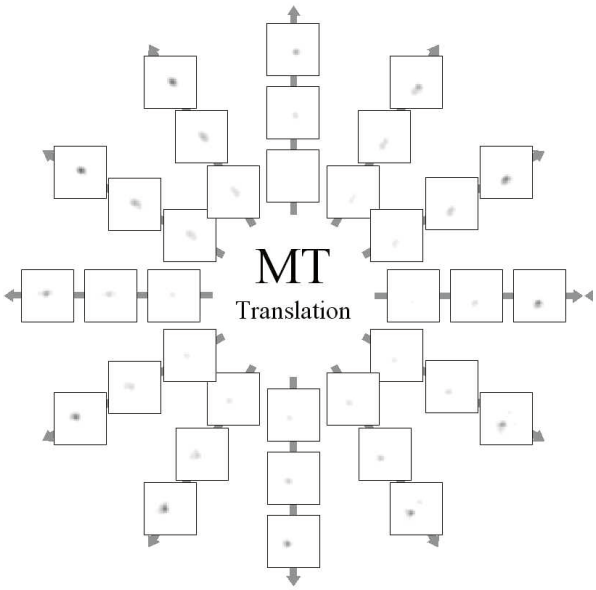


Figure 3

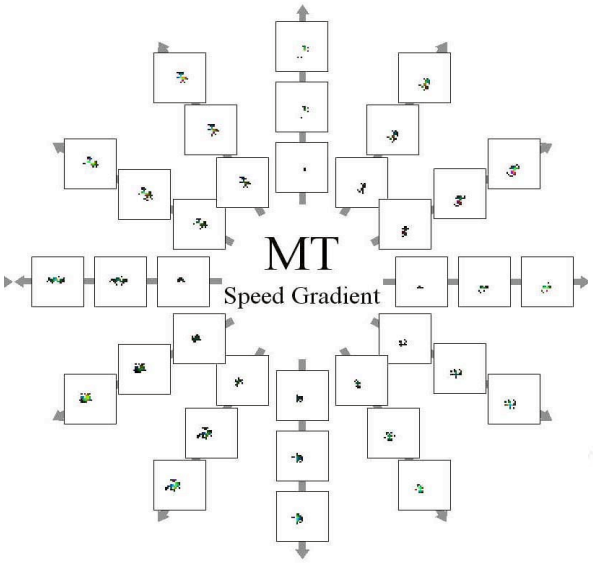


Figure 4A

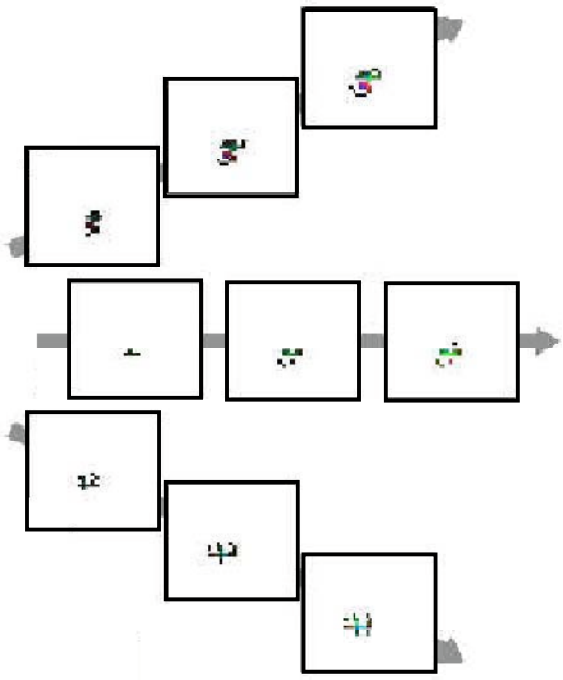


Figure 4B

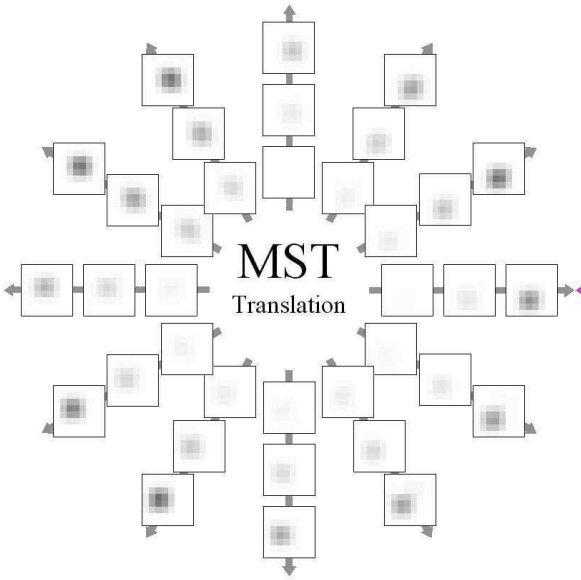


Figure 5A

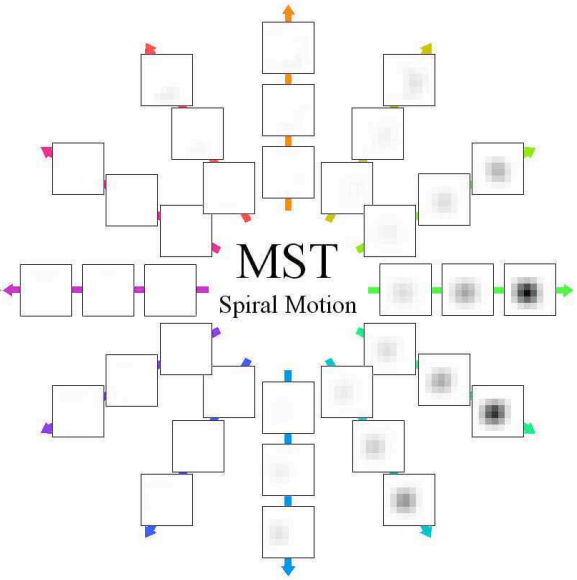


Figure 5B

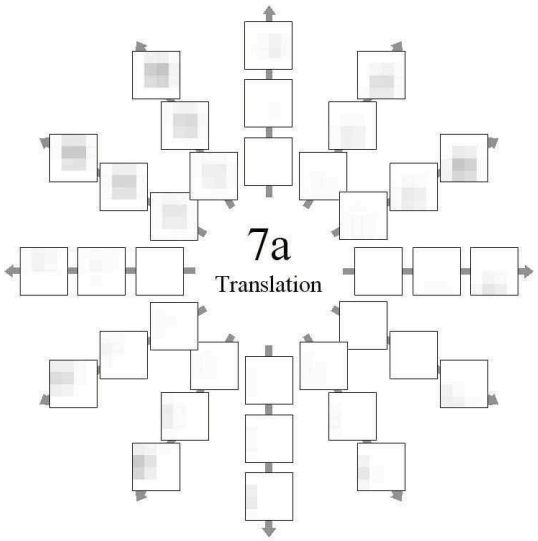


Figure 6A

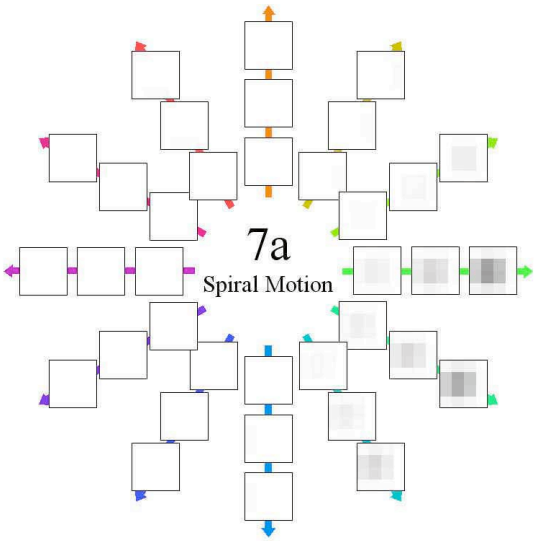
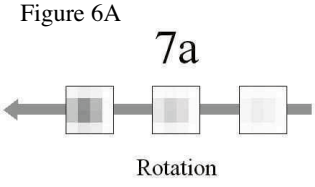


Figure 6B



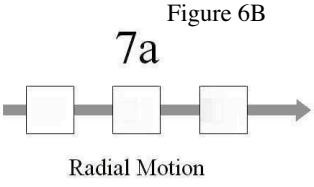


Figure 6C

5 Attending to 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 neurons in areas MT and MST are modulated by attention [45]. As a result of the model's feed-forward computations, the neural responses in the high-level areas (MST and 7a) roughly indicate the kind of motion patterns presented as an input but do not localize the spatial position of the patterns. The STM model was then applied to this feed-forward pyramid, adding in the required feedback connections, hierarchical WTA processes, and gating networks as originally defined in [10, 13]. The result is that the model attends to object motion, whether it exhibits a single or concurrent motion, and serially focuses on each motion in the sequence in order of response strength. Figure 7 shows the structure of the attention beam that localizes and labels the rotating square. The beam color is green, which signifies counterclockwise rotation (see the colour wheel in Figure 1). Note also the fact that its root is in a single representation of area 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 Figure 7 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.

6 Saliency and Distributed, Local WTA

The Winner-Take-All scheme within STM 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. 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, θ. 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 [13]).

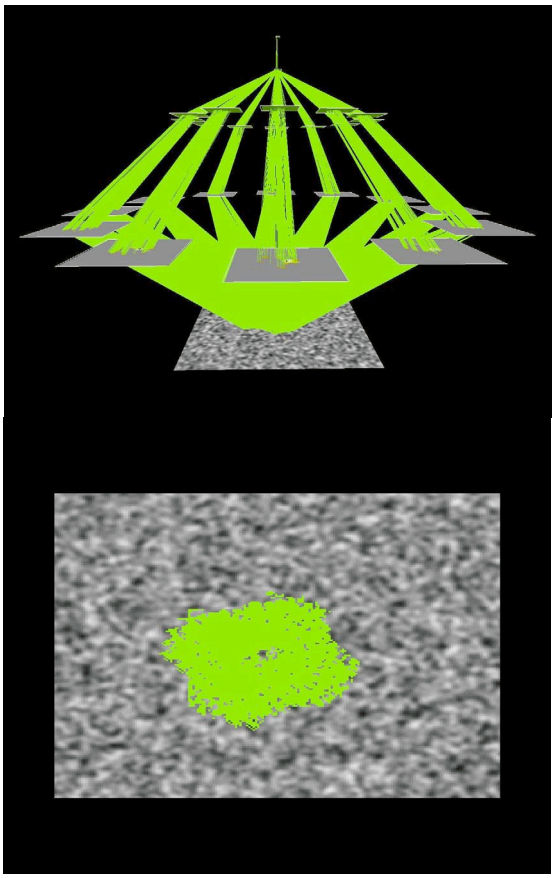


Figure 7

First, the WTA implementation uses an iterative algorithm with unit response values updated until convergence is achieved. Competition between units depends linearly on the difference between unit strengths in the following way. Some unit A will inhibit unit B in the competition if the response of A, denoted by $\rho(A)$ satisfies $|\rho(A)-\rho(B)|>\theta$. 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 $|\rho(A)-\rho(B)|$ for all units A. It has been shown [13] 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 θ and the maximum possible value, Z, across all unit responses. The reason for this is that because 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 $\log_2\left(\frac{A-\theta}{A-B}\right)$ 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 is, 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 θ. 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? 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). To allow full generality, define a receptive field as the set of z contiguous locations $R = \{r_i = (x_i, y_i)$, i=1...z$\}$. The neuron receives input from these locations from an arbitrary set of other neurons, not necessarily from the same representation or even from only the adjacent layers. Define the receptive field of a neuron as a set of arbitrarily shaped, contiguous, sub-fields $F = \{f_j = \{(x_{j,a}, y_{j,a})$, a=1..$b_j\}$, j=1...k$\}$, such that $\bigcup_{j=1,k} f_j = R$. Each sub-field is a retinotopic representation of a particular feature. This does not restrict features to be found only on adjacent layers of the hierarchy; rather, they may be from any other appropriate representation. The WTA competitions are defined on the sub-fields f_i. For spatially overlapping parts of these sub-fields, the features represented can be either mutually exclusive or can co-exist. The winning value is W, and this is determined by:

1. If k=1, that is, there is only a single sub-field f,

$$W = \max_{x,y} M^f_{x,y} \quad (1)$$

2. If F contains more than one sub-field, representing mutually exclusive features (sub-fields are fully overlapping in location), then

$$W = \max_{x,y}\left(\max_{i\in F} M^i_{x,y}\right) \quad (2)$$

3. If F contains more than one sub-field, all fully overlapping in location, 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

$$W = \max_{x,y}\left(\sum_{i\in F} M^i_{x,y}\right) \quad (3)$$

4. If F contains sub-fields 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:

$$W = \max_{x,y}\left[\sum_{b\in B} M^b_{x,y} + \max_{a\in A}\left(M^a_{x,y}\right)\right] \quad (4)$$

For sub-fields or portions thereof that are not spatially overlapping with any other sub-field, then the WTA process operates within that region following Rule 1.

As a result, there is no single saliency map in this model as there is in all other models. Indeed, there is no single WTA process necessarily, but several simultaneous WTA threads. 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 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. 1, 2, 3, or 4 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 next lower down competition to begin. The process continues until all layers have converged.

7 Feature Binding

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 Eq. 1 - 4. 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 and differently in each visual area and in practice 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;
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 process 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. This applies equally well for object recognition: faces are good examples of a grouping of features. In the demonstration above, 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. A motion pattern detector in layer MST simply sums responses of the corresponding MT units that feed it (see previous example). In layer MT, neurons sensitive to local motion direction within the object selective for gradients perpendicular to that direction respond as shown. Across all directions in the representation, one sees that the object has been '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. That 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. The WTA used is that shown in Eq. 3. 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 in [14]). No other model currently includes such a distributed definition of saliency. The binding is thus complete both in feature as well as location dimensions.

8 Discussion

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 Simoncelli and Heeger [28] and Beardsley and Vaina [29] are biologically relevant approaches that explain some specific functionality of MT and MST neurons, respectively, but do not include the embedding hierarchy in the motion pathway. On the other hand, there are hierarchical models for the detection of motion (e.g., [30, 31]), but unlike the present model they do not provide a biologically plausible version of the motion processing hierarchy.

Another strength of our model is its mechanism of visual attention. To our knowledge, there are only 2 other motion models employing attention for motion. The earlier one is due to Nowlan and Sejnowski [32]. There, processing that is much in the same spirit as ours but very different in form takes place. They compute motion energy with the goal of modelling 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 focussed on the most reliable estimates of velocity. There is no top-down component nor full processing hierarchy. Attentional modulation does not appear to be within the scope of their model. The second one is from Grossberg, Mingolla, and Viswanathan [33], which is a motion 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 but 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. Neither model has the breadth of processing in the motion domain or in attentional selection as the current work.

Of course, this is only the beginning and we 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 no cooperative nor competitive processing among units within a layer. Experimental work examining the relationship of this particular structure to human vision is also on-going

Acknowledgements

The work is supported by grants from the Natural Sciences and Engineering Research Council of Canada, the Institute for Robotics and Intelligent Systems, a Government of Canada Network of Centres of Excellence and Communication and Information Technology Ontatio, a Province of Ontario Centre of Excellence. JKT holds the Canada Research Chair in Computational Vision.

References

1. Aggarwal, J.K., Cai, Q., Liao, W., Sabata, B. (1998). Nonrigid motion analysis: Articulated and elastic motion, *Computer Vision and Image Understanding 70(2)*, p142–156.
2. Shah, M., Jain, R. (1997). Visual recognition of activities, gestures, facial expressions and speech: an introduction and a perspective, in **Motion-Based Recognition**, ed. by M. Shah and R. Jain, Kluwer Academic Publishers.
3. Cedras, C., Shah, M. (1994). A survey of motion analysis from moving light displays, IEEE CVPR-94, Seattle, Washington, p214-221.
4. Cedras, C., Shah, M. (1995). Motion-based recognition: A survey, *Image and Vision Computing, 13(2)*, p129-155.
5. Hildreth, E. Royden, C. (1995). Motion Perception, **in The Handbook of Brain Theory and Neural Networks**, ed. by M. Arbib, MIT Press, p585 – 588.
6. Tsotsos, J.K. (1980). A framework for visual motion understanding, Ph.D. Thesis, Dept. of Computer Science, University of Toronto, May.
7. Dickmanns, E.D., Wünsche, H.J. (1999). Dynamic vision for perception and control of motion, **Handbook of Computer Vision and Applications Vol. 2**, ed by B. Jahne, H. Haubeccker, P. Geibler, Academic Press.
8. Dreschler, L., Nagel, H.H. (1982). On the selection of critical points and local curvature extrema of region boundaries for interframe matching, Proc. Int. Conf. Pattern Recognition, Munich, p542-544.
9. Wachter, S., Nagel, H.H. (1999). Tracking persons in monocular image sequences, *Computer Vision and Image Understanding 74(3)*, p174–192.
10. Tsotsos, J.K. (1990). Analyzing vision at the complexity level, *Behavioral and Brain Sciences 13-3*, p423 - 445.
11. Desimone, R., Duncan, J., (1995). Neural Mechanisms of Selective Attention, *Annual Review of Neuroscience 18*, p193 - 222.
12. Treue, S., Martinez-Trujillo, J.C., (1999). Feature-based attention influences motion processing gain in macaque visual cortex, *Nature*, 399, 575 – 579.
13. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N. & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, 78, 507-545.
14. Koch, C., Ullman, S., (1985). Shifts in selective visual attention: Towards the underlying neural circuitry, *Hum. Neurobiology 4*, p219 - 227.
15. Moran, J., Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex, *Science 229,* p782 - 784.

16. Kastner, S., De Weerd, P., Desimone, R., Ungerleider, L. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI, *Science 282*, p108-111.
17. Vanduffel, W., Tootell, R., Orban, G. (2000). Attention-dependent suppression of metabolic activity in the early stages of the macaque visual system, *Cerebral Cortex 10,* p109-126.
18. Brefczynski J.A., DeYoe E.A. (1999). A physiological correlate of the 'spotlight' of visual attention. *Nat Neurosci*. Apr;2(4), p370-374 .
19. Gandhi S.P., Heeger D.J., Boynton G.M. (1999). Spatial attention affects brain activity in human primary visual cortex, *Proc Natl Acad Sci U S A*, Mar 16;96(6), p3314-9 .
20. Smith, A., Singh, K., Greenlee, M. (2000). Attentional suppression of activity in the human visual cortex, *NeuroReport*, Vol.11 ,No.27, p271–277.
21. Reynolds, J., Chelazzi, L., Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4, *The Journal of Neuroscience, 19(5)*, p1736–1753.
22. Caputo, G., Guerra, S. (1998). Attentional selection by distractor suppression, *Vision Research 38(5)*, p669 - 689.
23. Bahcall, D., Kowler, E. (1999). Attentional interference at small spatial separations, *Vision Research 39(1)*, p71 - 86.
24. Tsotsos, J.K., Culhane, S., Cutzu, F. (2001). From theoretical foundations to a hierarchical circuit for selective attention, **Visual Attention and Cortical Circuits**, ed. by J. Braun, C. Koch and J. Davis, p285 – 306, MIT Press.
25. Cutzu, F., Tsotsos, J.K. (2003). The selective tuning model of visual attention: Testing the predictions arisiing from the inhibitory surround mechanism, *Vision Research*, pp. 205 - 219.
26. Chelazzi, L., Duncan, J., Miller, E., Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search, *J. Neurophysiology 80*, p2918 - 2940.
27. Roelfsema, P., Lamme, V., Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey, *Nature 395*, p376 - 380.
28. Simoncelli, E.P. & Heeger, D.J. (1998). A model of neuronal responses in visual area MT. *Vision Research, 38 (5)*, 743-761.
29. Beardsley, S.A. & Vaina, L.M. (1998). Computational modeling of optic flow selectivity in MSTd neurons. *Network: Computation in Neural Systems, 9*, 467-493.
30. Giese, M.A. (2000). Neural field model for the recognition of biological motion. Paper presented at the Second International ICSC Symposium on Neural Computation (NC 2000), Berlin, Germany.
31. Meese, T.S. & Anderson, S.J. (2002). Spiral mechanisms are required to account for summation of complex motion components. *Vision Research, 42*, 1073-1080.
32. Nowlan, S.J., Sejnowski, T.J., (1995). A Selection Model for Motion Processing in Area MT of Primates, *The Journal of Neuroscience 15 (2)*, p 1195 – 1214.
33. Grossberg, S., Mingolla, E. & Viswanathan, L. (2001). Neural dynamics of motion integration and segmentation within and across apertures. *Vision Research, 41*, 2521-2553.
34. Zemel, R. S., Sejnowski, T.J., (1998). A Model for Encoding Multiple Object Motions and Self-Motion in area MST of Primate visual cortex, *The Journal of Neuroscience, 18(1)*, 531 – 547.
35. Pack, C., Grossberg, S. Mingolla, E., (2001). A nerual model of smooth pursuit control and motion perception by cortical area MST, *Journal of Cognitive Neuroscience, 13(1)*, 102 – 120.
36. Perrone, J.A. & Stone, L.S. (1998) Emulating the visual receptive field properties of MST neurons with a template model of heading estimation. *The Journal of Neuroscience*, 18, 5958-5975.
37. Orban, G.A., Kennedy, H. & Bullier, J. (1986). Velocity sensitivity and direction sensitivity of neurons in areas V1 and V2 of the monkey: Influence of eccentricity. *Journal of Neurophysiology, 56 (2)*, 462-480.
38. Heeger, D.J. (1988). Optical flow using spatiotemporal filters. I*nternational Journal of Computer Vision, 1 (4)*, 279-302.
39. Lagae, L., Raiguel, S. & Orban, G.A. (1993). Speed and direction selectivity of Macaque middle temporal neurons. *J Neurophysiology, 69* (1), 19-39.
40. Felleman, D.J. & Kaas, J.H. (1984). Receptive field properties of neurons in middle temporal visual area (MT) of owl monkeys. *Journal of Neurophysiology, 52*, 488-513.
41. Treue, S. & Andersen, R.A. (1996). Neural responses to velocity gradients in macaque cortical area MT. *Visual Neuroscience, 13*, 797-804.
42. Graziano, M.S., Andersen, R.A. & Snowden, R.J. (1994). Tuning of MST neurons to spiral motions. *Journal of Neuroscience, 14 (1)*, 54-67.
43. Duffy, C.J. & Wurtz, R.H. (1997). MST neurons respond to speed patterns in optic flow. *Journal of Neuroscience, 17(8)*, 2839-2851.
44. Siegel, R.M. & Read, H.L. (1997). Analysis of optic flow in the monkey parietal area 7a. *Cerebral Cortex, 7*, 327-346
45. Treue, S. & Maunsell, J.H.R. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature, 382*, 539-541.
46. Koenderink, J.J., van Doorn, A.J. (1976) *J. Opt. Soc. Am.* 66, 717.
47. Longuet-Higgins, H.C., Prazdny, K. (1980). *Proceeding of the Royal Society of London*, 208, 385 (1980).
48. Tsotsos, J.K., Liu, Y., Martinez-Trujillo, J., Pomplun, M., Simine, E., Zhou, K. (2004). Attending to Visual Motion, (submitted).
49. Martinez Trujillo JC., Tsotsos, J.K., Simine, E., Pomplun, M., Wildes, R., Treue, S., Heinze, J.H., Hopf, J.-M. (submitted). The human brain specialises in the processing of velocity gradients in optical flow.