Motion Integration and Disambiguation by Spiking V1-MT-MSTl Feedforward-Feedback Interaction

Maximilian P. R. Lühr  
Inst. of Neural Information Processing  
Ulm University  
Ulm, Germany  
maximilian.loehr@uni-ulm.de

Daniel Schmid  
Inst. of Neural Information Processing  
Ulm University  
Ulm, Germany  
daniel-1.schmid@uni-ulm.de

Heiko Neumann  
Inst. of Neural Information Processing  
Ulm University  
Ulm, Germany  
heiko.neumann@uni-ulm.de

Abstract—Motion detection registers items within restricted regions in the visual field. Early stages of cortical processing of motion advance this estimate by integrating spatio-temporal input responses in area V1 to build feature representations of direction and speed in area MT of primate cortex. The neural mechanisms underlying such processes are not yet fully understood. We propose a neural model of hierarchically organized areas V1, MT, and MSTl, with feedforward and feedback connections. Each area serves a distinct purpose and is formally represented by layers of model cortical columns composed of excitatory and inhibitory spiking neurons with conductance-based activation dynamics. Recurrent connections enhance activations by modulatory interaction and divisive normalization. MT population activities allow to estimate motion direction and speed which we show for various stimuli. The importance of the feedback connections for disambiguation is demonstrated in simulated lesion studies.

I. INTRODUCTION

The reliable detection of moving objects in the surround and the estimation of self-motion and body posture are prerequisites of robust visual space and object perception. The primate visual system achieves such performance over a sequence of neural transformations of the retinal input. Different stages along the dorsal pathway in visual cortex build representations of feature and shape motion in the scene. In order to accomplish this, the visual system must integrate information derived from a set of local detectors which are distributed over the visual field. Such cortical processing of motion starts in area V1 subsequently feeding into area MT and then forwarding resulting activity distributions to MST with its lateral and dorsal subdivisions [1]. Building coherent neural representations of moving shapes takes time such that localized 2D intrinsic shape features and 1D extended contour outlines are suitably weighted according to their relevance [2, 3]. Motion processing is driven by the feedforward sensory stream but is influenced by re-entrant feedback signals from higher-level areas which control the gain of neurons in V1 and MT [4, 5].

Here, we propose a detailed neural architecture of spiking neurons for motion detection and integration at the initial stages of the dorsal pathway in visual cortex. Spike-encoding is included for increased biological plausibility and for later model implementation in event-driven neuromorphic hardware platforms. Results of initial spatio-temporal filtering of driving visual input are laterally integrated and further enhanced by context-driven modulating feedback and down-modulating divisive outer-surround inhibition [6]. The model considers a hierarchy of stages corresponding to areas V1, MT, and MSTl. We suggest that different stages of hierarchical processing of input motion contribute different functions for the detection and integration of motion information. Our proposal builds upon a canonical neural circuit model (CNCM) which properties have been described and mathematically analyzed in [7]. This encoding scheme is here extended by transforming the circuit model into a spiking network architecture that incorporates the adaptive exponential firing (AdEx) mechanism to generate discrete events as outputs [8]. Along the feedforward pathway cells process spatio-temporal input signals over a hierarchy of increasing feature selectivity and spatial size. The hierarchical bottom-up stream of driving signals is combined with top-down modulatory feedback signals to build a counterstream network for top-down prediction. We demonstrate that V1-MT-MSTl feedforward and feedback interaction redistributes the neural responses to disambiguate initial motion responses in a context-dependent way and generates coherent shape motion representations to reliably encode direction and speed of moving objects. The work thus further contributes to reveal the roles different processing pathways play for feature detection, shape registration, the delivery of context and prediction by top-down feedback signals [9].

II. PREVIOUS WORK

Several works investigated the initial stages of motion detection along the cascade of areas V1 and MT of the dorsal pathway. These investigations identified canonical neural operations for the sequential filtering and competition to generate representations of moving patterns, such as spatio-temporal filtering at different scales, non-linear response integration, and opponent competition [10, 11]. Divisive normalization has been suggested to explain several nonlinear response properties of neurons at different neocortical levels [12, 13]. A cascaded model of linear-nonlinear operations was proposed...
by [14] in order to account for MT cell selectivities to component and pattern motion. The initial filtering of raw input built upon earlier investigations by [15–17] showing formal equivalence of filtering with competition and spatio-temporal correlation mechanisms [18]. Initial motion detection by V1 receptive fields (RF) can only measure flow that is normal to local contrast orientation. This so-called aperture problem is resolved over time in area MT with neuron population responses switching from signaling normal flow responses to pattern motion direction [19]. Motion can be signaled accurately already by V1 end-stopped neurons given localized features that are intrinsic to the target shape [20]. The contributions of initial filtering for feature motion detection in V1 and subsequent non-linear normalization operations for reliable MT motion integration was investigated in [21]. The authors demonstrated for oriented bars smaller in length than MT cell RFs that divisive inhibition plays a key role in the re-weighting of input strength for localized feature motions and aperture responses. More recently, hierarchical models have been suggested using hierarchical feedforward pooling schemes to process complex input motion for multiscale spatio-temporal filtering and pattern response integration in V1 and MT [22] and visual heading computation in the presence of occluding moving objects [23]. Several models suggested building speed selective representations in area MT operating either in the spatial or frequency domain [24–26]; the latter suggesting to combine sustained and transient output responses from V1 complex cells for generating a speed sensitive sensor. Several investigations emphasized that feedback signals are generated at higher stages along the feedforward pathway to reenter the motion computation at lower stages. Such computations can potentially help to resolve the uncertainties from initial feature detection. For example, in [27] the model proposed by [23] was extended by integrating inhibitory feedback connections from pooling cells in MT to V1 complex cells to improve their feature selectivity. An earlier model by [28] demonstrated that the integration of bidirectional signal flow between areas V1, MT, and MST helps to integrate and segregate local motion signals incorporating non-local context information, such as figure-ground properties and mutual occlusions. Similarly, motion feature selection and integration with and without guiding attention has been investigated by [29, 30] suggesting that re-entrant feedback signals are required to segregate semi-transparent motion. The latter models build upon a core model of bidirectionally coupled V1 and MT [31]. This work suggested disambiguation of normal flow in apertures via activity propagation initiated from localized features. The V1-MT counterstream interaction, with MT cell pooling over a larger spatial neighborhood, predicted motion disambiguation in both V1 and MT representations simultaneously. This was not confirmed experimentally [32]. The model presented here resolves this discrepancy predicting the growth cone of disambiguation in MT while V1 cells only improve their feature selectivity and reduce their responses to aperture motion. While the models mentioned above were not speed sensitive, [33] proposed a mechanism that incorporated different speed channels based on spatio-temporal filters.

The models discussed so far utilize rate-coded mechanisms to simulate population activities in response to motion inputs. A few spiking models have been suggested as well. For example, [34] investigated how a velocity-sensitive cell in MT might be learned using an STDP mechanism, [35] developed a spiking two-layer neural model to detect moving object boundaries in image sequences. The architecture in [36] suggested a feedforward spiking network model for sequence analysis in action recognition including areas V1 and MT, and [37] hypothesized that specific connectivity patterns in top-down pathways implement the generative model activation patterns in predictive spike coding for motion tracking. More recently, initial event-based (spiking) motion detector schemes have been realized and implemented on dedicated neuromorphic hardware platforms (e.g., [38]).

We here build upon such previous investigations and develop a spiking neural architecture of model V1, MT, and MSTl processing, incorporating driving feedforward processing together with top-down feedback signal pathways. Detailed investigations by selective lesioning of the model allow to value the individual contributions of these processing principles. The spiking network architecture is also well suited for later implementation on neuromorphic hardware platforms.

III. METHODS

A. Architecture

The proposed model architecture includes three cortical areas, namely V1 (primary visual area), MT (middle temporal area) and MSTl (lateral part of medial superior temporal area). Each is defined as a topographically organized layer of CNCMs corresponding to cortical columns and their lateral connections. The areas are hierarchically located on consecutive levels having different selectivities, such that the output of one level is fed as input to the subsequent layer (c.f. Fig. 1). Output activation at one level in the hierarchy generates re-entrant feedback signals to the stage below in that hierarchy. Each model area includes three computational operations, namely an input filtering using static spatio-temporal filters (see III-C), a re-entrant top-down modulation of activation, and a local spatio-temporal competition (see III-B) leading to response normalization over a pool of cells in a space-feature domain [7]. The input filtering and further mutual cell interaction plays a different computational role in each layer. Model V1 detects local movements in the input stream within a small RF. The input is a stream of events generated from temporal changes in the input luminance distribution, as registered in the retina. Local competition between cells at this stage reduce or suppress responses at locations of intrinsically 1D structure where local normal flow is signaled. This aids subsequent layers to detect the true motion direction which is mainly guided by locations of intrinsically 2D structure, e.g., corners and end-points. This first layer consists of $128 \times 128 \times 8$ pairs of excitatory-inhibitory (E-I) neurons. Each corresponds to a single spatial location in an $128 \times 128$ px image, which
is the input size of all further experiments, and one of eight discrete directions of motion, for increments of $\Delta = 45^\circ$.

Model MT builds the first velocity, i.e. direction and speed selective, representation of the image features. At this level output activity of V1 is integrated over a larger spatial and temporal domain utilizing velocity sensitive filters. Several selective channels enable to discern different speeds of motion along an actual movement direction. Feedback signals are generated from MT velocity selective cells marginalizing over the speed dimension. This feedback reenters the more localized V1 direction selective activity distributions to enhance the gain of those components that might explain the current motion hypotheses. The net effect of such enhanced V1 responses is that the increase in local gain give these cells a competitive advantage in the subsequent competition for response normalization such that cells not receiving any re-entrant amplification will be down-modulated. The topographic structure of this second layer is formed by $128 \times 128 \times 6 \times 8$ E-I pairs where spatial locations and directions of motion correspond to those of layer V1. The additional explicit speed dimension is discretized into six channels to build explicit velocity representations. Speed selective cells have Gaussian tuning on log-speed input.

Inspired by [28] the model architecture includes area MSTl cells at the highest level which integrate MT output activity to find regions of homogeneous movement independent of the details and variations of speed. Again, output responses form re-entrant feedback signals to model MT neurons. At this stage the loopy feedforward and feedback interaction consolidates the responses of those MT cells that encode directions matching the movement hypothesis formed at the level of MSTl. In turn, this facilitates an activity spreading, or guided filling-in, for MT cell activation which evolves dynamically like a growth cone that is initiated at locations with unambiguous feature motion. The topographic structure of this last layer consists of $128 \times 128 \times 8$ linear neurons to discretely sample movements along eight directions. Any output visualization of motion is calculated from the population of model MT space-velocity representations.

**B. Model Neurons**

Model layers V1 and MT are populated by pairs of E-I cells resembling the interaction within a cortical column [7]. Their state and its respective temporal evolution is defined by membrane potentials, $r$, $p$, and their first order temporal changes. The spiking activity $u_r$ of an E subpopulation forms the output of the respective layer. The net input current to each type of model neuron is denoted by

$$I_r = (E_r^{ex} - r) \cdot g_{r}^{ex} \cdot (1 + \lambda \cdot g_{r}^{mod}) + (E_r^{in} - \gamma_r \cdot r) \cdot g_{r}^{in} \quad (1)$$

$$I_p = (E_p^{ex} - p) \cdot g_{p}^{ex} + (E_p^{in} - \gamma_p \cdot p) \cdot g_{p}^{in}, \quad (2)$$

which are driven by excitatory and inhibitory input conductances $g$. In the model, E-cells, $r$, receive excitatory input $g_{r}^{ex}$ via feedforward connections, modulatory input $g_{r}^{mod}$ by top-down feedback as well as by lateral E-cells$^2$, and inhibitory input $g_{r}^{in}$ from I-cells in the same layer. Similarly, I-cells, $p$, receive excitatory input $g_{p}^{ex}$ from E-cells of the regarding layer, as well as inhibitory input $g_{p}^{in}$ from I-cells of the same layer. The constants $E_r^{ex}$ and $E_r^{in}$ denote the excitatory and inhibitory reversal potentials, while $\lambda$ controls the strength of the gain modulation by feedback signals. We emphasize that feedforward and feedback signals have an asymmetric influence on the excitatory cells: while driving input signals are necessary to generate an activity, feedback signals can only amplify the response in the presence of coinciding input, whereas it cannot generate activity when driving input is absent. The integration of the different inputs specifies the temporal change of membrane potentials following first-order dynamics (using the notation $\partial_x x \equiv \dot{x}$).

$$\tau_r \dot{r} = f_r(r) - w_r + I_r \quad (3)$$

$$\tau_p \dot{p} = f_p(p) - w_p + I_p, \quad (4)$$

with a leak term $f_r(r)$ ($f_p(p)$ respectively) and an adaption term $w_r$ ($w_p$, respectively) for the spike-rate adaptation (see below). Steering parameters for the strength of divisive inhibition are given by $\gamma_r$ and $\gamma_p$, and $\tau_r$, $\tau_p$ denote the membrane time constants. For the definition of a spike-output

$^2$In the case of MT, lateral and feedback streams are additively combined.

![Fig. 1. Model architecture consisting of areas V1, MT and MSTl. Excitatory (E) cells are driven by the responses of their level-dependent feedforward filters which are driven by the output of the preceding layer. Modulatory input to E-cells is generated by top-down feedback and in the case of MT from lateral interaction as well. Inhibitory (I) cells are driven by integrated E-cell activity pooled over a larger neighborhood in the space-feature domain. Such pool activation of I-cells exerts inhibitory influence on E-cells at the corresponding location. Motion is estimated from area MT activity. Insets show examples of the spatio-temporal feedforward filters of V1 and MT, with light regions denoting positive and dark regions denoting negative coefficients.](image)
based dynamical neuron model, we incorporated mechanisms of the AdEx model [8] which considers leak and spike-rate adaptation terms, respectively, particularly:

$$f_r(r) = (E_r^{\text{leak}} - r) \cdot g_{r}^{\text{leak}} + g_{r}^{\text{leak}} \cdot \Delta_r \exp\left(\frac{r - V_r}{\Delta_r}\right)$$

$$f_p(p) = (E_p^{\text{leak}} - p) \cdot g_{p}^{\text{leak}} + g_{p}^{\text{leak}} \cdot \Delta_p \exp\left(\frac{p - V_p}{\Delta_p}\right),$$

for the leak and the additional dynamics for the adaptation

$$\tau_w \cdot \dot{w}_r = a_r \cdot (r - E_r^{\text{leak}}) - w_r$$

$$\tau_w \cdot \dot{w}_p = a_p \cdot (p - E_p^{\text{leak}}) - w_p,$$

with $\Delta_r$ and $\Delta_p$ as in [8], $g_{r}^{\text{leak}}$ denoting the leak strengths and $E_r^{\text{leak}}$ the reversal potentials in the spike-generation mechanisms. Spikes $u_r$ are generated whenever the potential of the E/I-cells exceeds the threshold $V_r^{\Theta}$ or $V_p^{\Theta}$, respectively. After spike-generation the membrane potential is set to a reset potential and the adaptation variable $w_r$ is incremented:

$$\text{if } r > V_r^{\Theta} : \quad u_r \rightarrow 1, \quad r \rightarrow E_r^{\text{leak}}, \quad w_r \rightarrow w_r + b_r$$

$$\text{else} : \quad u_r \rightarrow 0,$$

$$\text{if } p > V_p^{\Theta} : \quad u_p \rightarrow 1, \quad p \rightarrow E_p^{\text{leak}}, \quad w_p \rightarrow w_p + b_p$$

$$\text{else} : \quad u_p \rightarrow 0.$$

Spikes are transmitted to other neural sites by a filtering process via their overall connectivities and the associated weight coefficients (cf. Section II)

$$y^S(x, y, t, \theta) = (\Lambda^S(x, y, t, \theta) * u)(x, y, t, \theta),$$

with $\Lambda$ denoting a convolution operation over the respective spatial $(x,y)$, temporal $(t)$ and feature $(\theta)$ dimensions, $u$ being spikes emitted by a source (r or p cells), respectively. $\Lambda^S$ are filter matrices for different synapse types per layer and neuron population ($S \in \{\text{ex, in, mod}\}$). These filtered spikes $y^S$ are then integrated postsynaptically via

$$\tau^{\text{Syn}} y^S = -y^S + k^S \cdot y^S,$$

with the synaptic time constants $\tau^{\text{Syn}}$ which differ by connection, or synapse, and $k^S$ is a scaling parameter to define the impact of spikes. The time constants are parameterized according to [39]. The MST1 cell responses are defined by the simplified scheme utilizing a linear mapping

$$r_{\text{MST1}} = y_{\text{MST1,ex}}$$

The parameters for instantiating the different neuron models are defined in Table I.

C. Parameterization of feedforward and feedback filters

In the proposed model, the weighted connections between neuron populations are represented by filter kernels which are specific to the input stage of the different levels. Such filters are applied to their respective input activity distribution using a convolution operation. Feedforward filter banks at different layers generate input to the E-cells of V1 and higher layers (see $g_{r}^{\text{leak}}$ in (1)). Feedback filters act along the reverse direction of the hierarchy with the role of the cells and their receptive and the projective field exchanged (c.f. $g_{r}^{\text{mod}}$). The I-cells of a layer are fed by the E-cells at this level (e.g., via $g_{r}^{\text{ex}}$ in (2)) which is the result of a convolution as well. Finally, there is lateral interaction within MT to support spreading of activity within the E-cell population. The qualitative shape of these kernels is described below and exact parameters can be found in Table II. All types of filter kernels are normalized such that their coefficients sum to one. Gabor filter coefficients sum to zero to achieve zero DC-level response. Feedback kernels are created such that their maximum coefficient amounts to one. Scaling factors $k^S$, c.f. (13), for filter responses are listed in Table III. Model parameters and kernel weights have been chosen according to physiologically plausible measures and selected to ensure stable dynamics.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Unit</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_{r}^{\text{leak}}$, $g_{p}^{\text{leak}}$</td>
<td>30</td>
<td>1</td>
<td>leak conductance</td>
</tr>
<tr>
<td>$E_r^{\Theta}$, $E_p^{\Theta}$</td>
<td>0</td>
<td>mV</td>
<td>excitatory reversal potential</td>
</tr>
<tr>
<td>$E_r^{\text{in}}$, $E_p^{\text{in}}$</td>
<td>-75</td>
<td>mV</td>
<td>inhibitory reversal potential</td>
</tr>
<tr>
<td>$E_r^{\text{leak}}$, $E_p^{\text{leak}}$</td>
<td>-70.6</td>
<td>mV</td>
<td>leak reversal potential</td>
</tr>
<tr>
<td>$V_r$, $V_p$</td>
<td>-50.4</td>
<td>mV</td>
<td>exponential term offset</td>
</tr>
<tr>
<td>$\Delta_r$, $\Delta_p$</td>
<td>2</td>
<td>mV</td>
<td>exponential term slope</td>
</tr>
<tr>
<td>$\tau_{\text{r}}$, $\tau_{p}$</td>
<td>1</td>
<td>1</td>
<td>divisive inhibition scaling</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.5</td>
<td>1</td>
<td>modulation scaling</td>
</tr>
<tr>
<td>$a_r$, $a_p$</td>
<td>4</td>
<td></td>
<td>sub-threshold adaptation strength</td>
</tr>
<tr>
<td>$b_r$, $b_p$</td>
<td>80.5</td>
<td>1</td>
<td>super-threshold adaptation strength</td>
</tr>
<tr>
<td>$V_r^{\Theta}$, $V_p^{\Theta}$</td>
<td>20</td>
<td>mV</td>
<td>spike threshold</td>
</tr>
<tr>
<td>$\tau_{\text{ex}}$, $\tau_{\text{in}}$</td>
<td>281</td>
<td>ms</td>
<td>membrane time constant E-cell</td>
</tr>
<tr>
<td>$\tau_{\text{r}}$, $\tau_{p}$</td>
<td>144</td>
<td>ms</td>
<td>adaptation time constant E-cell</td>
</tr>
<tr>
<td>$\tau_{\text{in}}$, $\tau_{\text{mod}}$</td>
<td>1</td>
<td>1</td>
<td>membrane time constant I-cell</td>
</tr>
<tr>
<td>$\tau_{\text{ex}}$, $\tau_{\text{in}}$</td>
<td>1.7</td>
<td>ms</td>
<td>adaptation time constant I-cell</td>
</tr>
<tr>
<td>$\tau_{\text{mod}}$</td>
<td>0.5</td>
<td>ms</td>
<td>inhibitory synaptic time constant</td>
</tr>
<tr>
<td>$\tau_{\text{in}}$</td>
<td>26</td>
<td>ms</td>
<td>modulatory synaptic time constant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer</th>
<th>Filter Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>feedforward: Gabor: $\sigma_{x, y} = 1.22$ px, $\sigma_{x, y} = 2.86$ px $\sigma_{x, y} = 33$ ms, $\sigma_{\text{half}} = 100$ px/s</td>
</tr>
<tr>
<td>MT</td>
<td>feedforward: DoG: $\sigma_{x, y}^{\text{in}} = 1.22$ px, $\sigma_{x, y}^{\text{in}} = 1.96$ px $\sigma_{x, y}^{\text{in}} = 24$, 40, 56, 72, 88, 104 px/s</td>
</tr>
<tr>
<td>feedback: Gaussian over space: $\sigma_{x, y} = 7.5$ px $\sigma_{x, y} = 10$ px</td>
<td></td>
</tr>
<tr>
<td>MSTI</td>
<td>feedforward: Gaussian over space: $\sigma_{x, y} = 13.44$ px $\sigma_{x, y} = 33$ ms, $\sigma_{\text{half}} = 100$ px/s</td>
</tr>
</tbody>
</table>

In V1 driving feedforward activity is generated by a bank of spatio-temporal Gabor filters of four different orientations, each with direction selectivities orthogonal to the orientation of a Gabor-RF (yielding eight movement directions). We also
include oriented filters selective to zero motion, which are separable in space-time. To derive phase invariant responses for the retinal image both sine and cosine Gabor are used in quadrature. The vector of wave propagation for the filters is slanted towards the temporal axis (see the lower inset in Fig. 1). In order to yield causal impulse response properties filter weights are set to zero for negative times. The output of the spatio-temporally separable filters is used to subtrac-

tively inhibit the inseparable filters responding to normal flow orthogonal to their orientation preference. This rectified difference drives V1 E-cells, forming a simple mechanism of enhancing end-stop response. V1 inhibition pools excitatory activation over a Gaussian weighted spatial neighborhood for each direction tuning separately. Thus, V1 cells with different direction selectivities do not interact.

In MT feedback driving activity is generated by a filter bank consisting of 48 space-velocity selective filters similar to those shown in the upper inset of Fig. 1. These causal kernels are cylinders aligned along the temporal axis that have a Difference-of-Gaussians (DoG) cross-section (similar to [40]). Different speed tuning is generated by rotation towards or away from the t-axis and a rotation around the t-axis tunes the direction selectivity. The cylindrical weighting is extended to cover approximately 300 ms for low speeds. Its shape was chosen to sample just a few cells of V1 over its past temporal window while keeping a high sensitivity for direction or speed since the inhibitory Gaussian component punishes spatio-temporal misalignment. The MT inhibitory kernel for integrating over the pool of activations is Gaussian over space but uniform in both speed and direction. This allows strong activity of E-cells of a given feature combination to suppress nearby E-cells of a competing feature combination. To facilitate lateral integration of E-cell activities a kernel is specified that is Gaussian in space and over directions but uniform over speeds. Activities from cells with different speed tuning can facilitate each other as can activities of directly adjacent direction channels. To prevent neuronal self-excitation the central spatial coefficient of the kernel is zero.

Re-entrant feedback signals from levels higher in the processing hierarchy are generated by convolving the output activity distribution by kernels to specify the projective field of modulating signals. MT feedback signals are generated by convolving E-cell activations by a spatial Gaussian to define the modulatory cone for V1 cells. Speed selective responses in MT are marginalized over the different speed channels to facilitate V1 spatio-temporal cell responses of matching direction selectivity. Finally, direction selective model MSTl cells pool MT output using spatial Gaussian weights and uniform weighting for speeds. The speed integration makes MSTl selective to detect homogeneous directional motion fields. The resulting activation is fed back to enhance MT velocity representations for corresponding directions regardless of their speed characteristics.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Variable</th>
<th>Value</th>
<th>Unit</th>
<th>Post-multiplier type</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>(k_\text{exc}^{V1} )</td>
<td>20000</td>
<td>1</td>
<td>E-cell excitatory</td>
</tr>
<tr>
<td></td>
<td>(k_\text{inh}^{V1} )</td>
<td>1250</td>
<td>1</td>
<td>E-cell inhibitory</td>
</tr>
<tr>
<td></td>
<td>(k_\text{exc}^{V1} )</td>
<td>0.02</td>
<td>1</td>
<td>E-cell modulatory</td>
</tr>
<tr>
<td></td>
<td>(k_\text{exc}^{V1} )</td>
<td>10000</td>
<td>1</td>
<td>I-cell excitatory</td>
</tr>
<tr>
<td></td>
<td>(k_\text{inh}^{V1} )</td>
<td>1250</td>
<td>1</td>
<td>I-cell inhibitory</td>
</tr>
<tr>
<td>MT</td>
<td>(k_\text{exc}^{MT} )</td>
<td>20000</td>
<td>1</td>
<td>E-cell excitatory</td>
</tr>
<tr>
<td></td>
<td>(k_\text{inh}^{MT} )</td>
<td>5000</td>
<td>1</td>
<td>E-cell inhibitory</td>
</tr>
<tr>
<td></td>
<td>(k_\text{lat}^{MT} )</td>
<td>1</td>
<td>1</td>
<td>E-cell lateral modulatory</td>
</tr>
<tr>
<td></td>
<td>(k_\text{fb}^{MT} )</td>
<td>0.05</td>
<td>1</td>
<td>E-cell feedback modulatory</td>
</tr>
<tr>
<td></td>
<td>(k_\text{exc}^{MT} )</td>
<td>10000</td>
<td>1</td>
<td>I-cell excitatory</td>
</tr>
<tr>
<td></td>
<td>(k_\text{inh}^{MT} )</td>
<td>5000</td>
<td>1</td>
<td>I-cell inhibitory</td>
</tr>
</tbody>
</table>

IV. RESULTS

To evaluate the function of the proposed model we ran simulations on several different inputs. We used synthetic input configurations similar to those in animal electrophysiology or behavioral studies [19, 41] as well as real-world address-event-representation (AER) data of [42] and of our own which was recorded with a neuromorphic DVS camera [43]. Using a moving oriented bar we investigate how spatio-temporal changes of intrinsically 1D and 2D intensity features are initially detected, how these responses are differently weighted to link them into a coherent moving shape, and how the different stages along the model visual pathway interact in order to disambiguate local measures to represent the target movement. We show how the aperture problem for the bar is resolved dynamically and how a representation of moving shape in model MT is generated. Our simulations show that reliable shape motion representations for different directions and speeds are only achieved at the level of MT but not in V1.

To determine the relative importance of each model component we conducted simulated lesion studies eliminating selected connections between layers and show how the generation of moving shape representations is impaired. We also demonstrate that the model architecture is capable of processing real camera input from a database provided in [42] (see Fig. 4). Finally, the representational differences of the modeled areas are compared.

For all simulations the model equations have been discretized using the explicit Euler scheme with a step size of 2 ms. To read out robust quantities for the response estimates and plots of this section we apply an exponential filter \((\tau_{res} = 13 \text{ time steps})\) on the output spikes. The resulting activation \(g_{r}^{MT, res}\) resembles a postsynaptic potential similar to (13).

A. Direction & Speed Estimation

To estimate individual motion vectors we employ a simple readout scheme of MT activity. At each spatial location \((x, y)\) neural responses of input filtering for combined direction \((\phi)\) and speed selectivity \((s)\) have been generated to yield velocity votes \(\hat{v}_{e}(\phi, s) = s(\cos(\phi), \sin(\phi))\). The final velocity
estimate is calculated by the sum of all votes weighted by the responses of all model MT neurons at that location:

$$\vec{v} = \sum_{\phi, s} \vec{v}_r(\phi, s) \cdot g^\text{MT, res}_s(\phi, s) / \sum_{\phi, s} g^\text{MT, res}_s(\tilde{\phi}, \tilde{s}) \quad (15)$$

We employ the average angular error (AAE) as metric to evaluate deviation between velocity estimates and the ground truth motion. Since high neuronal responses are interpreted as high confidence we weight the calculated average by the motion energy summing all cell responses at that pixel location.

$$\Delta_{\text{AAE}} = \frac{1}{N} \sum_{i=1}^{N} \left| \vec{v}_i - \vec{v}_i^\text{true} \right|$$

We define the average angular error as the mean of the absolute differences between the estimated and ground truth vector responses of all model MT neurons at each pixel location over time. Each bar stimulus is 1 px wide and moves along 45° at 64 px/s. AAE for six different bar lengths. Larger stimuli take longer to minimize the error. Feb 11.png

Responses are stronger at end-points and from such localized intrinsic feature locations the correct motion hypothesis is propagated through existing activity along the complete edge. Responses signaling locally ambiguous motion (normal flow along elongated contrast) are reduced.

The movement speed \(s_{x,y}\) is the length of the estimated motion vector. Feb 12.png

To ascertain the importance of each of the recurrent lateral and feedback pathways in our model we conducted several lesion studies. In such simulations the connections of either...
the MT-V1 and MSTl-MT feedback, the lateral connections within MT or combinations of those have been set to zero.

Fig. 5(a) shows that eliminating any feedback signals reentering V1 and MT, akin to considering a pure feedforward-network, drastically impairs the computation of shape motion. The error reaches only $\Delta_{\text{AAE}} = 27^\circ$ after about 100 time steps and barely improves after that. Feedback from MSTl into MT alone (no feedback from MT into V1) converges to $\Delta_{\text{AAE}} = 25.6^\circ$ after 150 time steps. Feedback from MT into V1 (but removing re-entrant signals from MSTl to MT) leaves a final $\Delta_{\text{AAE}} = 17.3^\circ$ which shows that feedback at the lower levels has a large impact. However, all feedback signaling contributes to generating robust moving shape representations: Using both feedback streams but no lateral interaction in MT leads to an $\Delta_{\text{AAE}} = 2.1^\circ$. Including lateral interaction as well, the full model yields $\Delta_{\text{AAE}} = 1.6^\circ$ AAE at the end of the simulation and is still improving. When any feedback stream is missing, lateral connections alone show no noticeable effect.

C. Differences in V1, MT and MSTl Representations

All three layers of the model architecture serve different purposes (Section III-A). Fig. 6 shows the neuronal responses of each layer and the AAE estimated near the end of simulations. The detailed investigation considers the diagonally moving vertical bar (51 px) used above.

V1 neurons have a small RF and detect the true motion direction only at points of intrinsically 2D structure (e.g., at endpoints of the bar) or the localized feature points in complex patterns. The AAE there is low. Feedback from MT facilitates correct responses, however, the normal flow components of V1 are significantly reduced but not replaced by activity corresponding to the true direction. V1 cells are selective to the individual components of a moving pattern but insensitive to whole patterns [44]. This is reflected by the gray patches in

V. DISCUSSION & CONCLUSION

The proposed architecture explains how motion can be estimated in model area MT, utilizing feedforward and feedback streams between areas V1, MT and MSTl along the dorsal
pathway. Each layer has a different feature space and plays a different role. Feedback and lateral interaction govern the spread of information. This way motion hypotheses can be integrated over distances that exceed the receptive fields of MT and disambiguated through competitive normalization.

The importance of these recurrent connections has been demonstrated in simulated lesion studies. For a diagonally moving bar stimulus the error can drop below $\Delta_{\text{AEE}} = 2^\circ$ which is comparable to the results in [21], but removing any of the feedback streams drastically magnifies the error.

Still, there are limitations to the model. MT needs strong responses of V1 neurons at the end-points of the stimulus where the actual motion can be discerned in V1. Here, we decrease the response of all V1 movement neurons along the bar by subtracting the response of static contour neurons which is weaker at the endpoints, because the RF is only partially filled there. This works well for end-points of a bar and for $90^\circ$ corners in a contour. At T-junctions or crossings of long bars the inhibition is getting too strong and disambiguation takes significantly longer. Mechanisms for detecting intrinsic 2D features would stabilize the solution. These could be either signaled by a separate neuron population of end-stop neurons [21, 29] or be generated by scale-sensitive filter interactions within area V1 or between V1 and V2 [45].

A second issue is the current choice of feedforward filters in V1. Spatio-temporal Gabor filters impose a speed tuning so the contour inhibition mechanism will not work over an arbitrary range of different stimulus speeds. This can be seen in Fig. 4(b,c) were there is nearly no response close to the center of the windmill. One could switch to correlation detectors that have constant response for movements of any positive speed. Another possibility is to create multiple filters with different tuning speed and average their responses.

Future work can be pursued along several directions. The spatial size of each layer was chosen to be the same number of pixels. However, higher areas along the dorsal pathway have less spatial resolution than V1, thus, approximating a hierarchical pyramid of spatio-temporal filtering. Such a pyramidal representation might improve efficiency in simulation time. This will also support more rigorous testing of the proposed model on a greater variety of input data sets. Finally, creating a fully spike based model and realizing it on neuromorphic hardware is attempted as further extension.

Fig. 6. Representational differences of the three model areas. Stimulus is a vertical 51 px bar and activity was sampled at time step 650. (a)-(c) Angular error of estimations from the neuronal activity in layers V1, MT and MST. Same error legend as in Fig. 2(i) is used (white means zero error). Stimulus is shown in red. Increasing activation spread reflects the larger RF sizes. (d)-(e) Activities of V1 neurons selective to direction $45^\circ$ (the real stimulus movement) and $0^\circ$ (the stimulus orientation and direction of normal flow). Dark and light blue marks neurons selective to movement and to stationary contrasts, respectively, at the current horizontal position of the bar stimulus. (f)-(g) Activities of MT neurons selective to directions $45^\circ$ and $0^\circ$. Darker shades denote neurons of higher tuning speeds. (h)-(i) Activities of MST neurons of directions $45^\circ$ and $0^\circ$. All ordinate axes show the maximum range of neuronal activity.

**References**


