1 Saliency for velocity tuned features

We start by reviewing how the saliency of a generic set of features $\mathbf{Y}(l) = (Y_1(l), \ldots, Y_n(l))$ can be mapped to area V1 using the approach of [4].

1.1 Mapping saliency computation to area V1

When the features, $\mathbf{Y}(l)$, are of a bandpass nature, as is usual in biological vision, the feature responses follow a generalized Gaussian distribution (GGD) of scale *scale* α and *shape* β [7],

$$P_Y(y; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{|y|}{\alpha}\right)^{\beta}\right\}, \tag{1}$$

where $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\mathrm{d}t$, $t > 0$. In this case [4], $P_{C(l)|Y_k(l)}(1|y) = \sigma[g(y)]$, where $\sigma(y) = (1+e^{-y})^{-1}$ is a sigmoid, and $g(x)$ the log-likelihood ratio between the two class-conditional GGDs,

$$g(y) = \log \frac{P_{Y_k(l)|C(l)}(y|1)}{P_{Y_k(l)|C(l)}(y|0)} = \frac{|y|^{\beta_0}}{\alpha_0^{\beta_0}} - \frac{|y|^{\beta_1}}{\alpha_1^{\beta_1}} + T, \qquad T = \log \frac{\alpha_0\beta_1\pi_1\Gamma(1/\beta_0)}{\alpha_1\beta_0\pi_0\Gamma(1/\beta_1)}. \tag{2}$$

The scale parameters α_0 and α_1 are estimated by the maximum a posteriori probability (MAP) method, with conjugate (Gamma) priors, according to

$$\alpha_c^{\beta_c} = \frac{1}{\kappa_c}\left(\nu_c + \sum_{k\in\mathcal{W}_l^c} |y(k)|^{\beta_c}\right) \forall c \in \{0, 1\}. \tag{3}$$

The shape parameters β_c are quite consistent across image classes, and set to the value $\beta_c = 1$, which provides a good fit to natural images. Finally, replacing expectations by empirical averages, the saliency expression of Section 3 in the submission can be written as

$$S_k(l) = E_{Y(l)}\{\gamma[P_{C(l)|Y_k(l)}(1|y)]\} \tag{4}$$

$$\approx \frac{1}{|\mathcal{W}_l|}\sum_{j\in\mathcal{W}_l} \gamma\{\sigma(g[y(j)])\}, \qquad \mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1 \tag{5}$$

The computations of (2)-(5) can be mapped into a neural network that replicates the standard neurophysiological model of V1 neurons [3]. In particular, combining (2) and (3), it follows that $g[y(j)]$ computes a differential divisive normalization of the feature response $y(j)$ by the responses of the feature Y_k in the neighborhoods $\mathcal{W}_l^c$. Hence, $\sigma(g[y(j)])$ implements the computations of V1 simple cells under the standard model: a sequence of linear filtering, rectification, divisive normalization, and output saturation. (5) then pools the outputs of simple cells in $\mathcal{W}_l$ after passing them through the non-linearity $\gamma(x)$. These are the computations performed by V1 complex cells, under the standard model. The mapping is illustrated in Figure 3 in the submission.

So far, we have discussed the saliency model in terms of generic features $\mathbf{Y}(l)$, and have shown that the saliency computation can be mapped to area V1. We next show how the model can be extended to compute saliency of velocity tuned spatiotemporal features.

1.2 Constructing a model for an MT neuron

Velocity tuned neurons are found in area MT of the visual pathway. It is known that these MT neurons receive input from V1 complex cells [1]. Therefore, computational models for MT are usually constructed by combining the outputs of V1 complex cell afferents [9, 8]. As simple and complex cells in V1 are well modeled by the saliency network discussed above [4], a model for MT can be constructed by substituting this network model in place of the standard model for V1 in the approach of Simoncelli and Heeger [9].

In our model, as in [9], the linear filtering in the V1 simple cell stage is achieved using spatio-temporal Gabor features $Z_k(l)$ that are sensitive to motion. The output of the model V1 complex cell then computes bottom-up saliency for the corresponding spatio-temporal feature using (5), making

it selective to the component of stimulus velocity orthogonal to the spatial orientation of the feature, but not truly direction selective. By combining the responses of a set of such features, a unit responding to velocity in a specific direction can be constructed using the approach of Heeger [6].

Let $S_{Z_j}(l)$ be the model output at the V1 stage for the j^{th} spatio-temporal feature, Z_j, using (5). Then the output of a unit tuned to velocity $\bar{v}_k$, corresponding to feature Y_k, is given by:

$$S_k(l) = \sum_j w_{jk}(\bar{v}_k) S_{Z_j}(l) \tag{6}$$

where the weights $w_{jk}(\bar{v}_k)$ are computed using the approach of [6] (see Appendix 1.3). This is equivalent to computing the saliency of *a complex spatio-temporal feature* Y_k, designed to respond to stimuli moving with a specific velocity $\bar{v}_k$, from a combination of simple spatio-temporal Gabor features:

$$Y_k(l) = \sum_j w_{jk}(\bar{v}_k) Z_j(l) \tag{7}$$

The expression for saliency of Y_k in (6) ignores the effect of dependencies between the simple spatio-temporal features, Z_j, which has been shown to be a reasonable approximation when the features are bandpass [11], as is the case for the Gabor features used in the construction of the model.

Finally, as in [9], the output of the unit is divisively normalized by the responses of other units:

$$S_k^{td}(l) = \frac{S_k(l)}{\sum_j S_j(l)} \tag{8}$$

Each model unit corresponding to a feature Y_k, tuned to velocity $\bar{v}_k$, can be thought of as being equivalent to a neuron in MT. The computed model output is analogous to the neuron's firing rate and responds maximally when the input stimulus moves with $\bar{v}_k$, the velocity to which the feature is tuned. This velocity is referred to as the *preferred velocity* of the model neuron. The interpretation, given by (8), is that velocity selective tuning is a reflection of the neuron's *function* as a detector of salient motion configurations in a particular velocity channel. The resulting network is illustrated in Figure 3 in the submission.

The model for MT proposed above is built from neurophysiologically plausible units, using the same architecture as [9]. So arguments for biological plausibility of the V1 stage [4] and of the architecture [9] extend to the proposed MT model. Further, by using a center-surround architecture in the V1 stage, the model accounts for the surround antagonism observed in MT neurons [10, 2], but not modeled by [9].

1.3 Computations of weights for the MT model

The weight w_{jk} used in (6) to compute the response of the k^{th} velocity tuned feature Y_k, from the j^{th} spatio-temporal Gabor feature Z_j can be evaluated using the Gabor energy approach of [6]. The feature response corresponding to Z_j is the output of the visual stimulus passed through a sine-phase three dimensional Gabor filter $\mathrm{g}_j(x, y, t)$ of the form:

$$\mathrm{g}_j(x, y, t) = \frac{1}{\sqrt{2}\pi^{\frac{3}{2}}\sigma_x\sigma_y\sigma_t} \sin(2\pi\omega_{x_j}x + 2\pi\omega_{y_j}y + 2\pi\omega_{t_j}t) e^{-\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}+\frac{t^2}{2\sigma_t^2}\right)} \tag{9}$$

where $\omega_{x_j}, \omega_{y_j}$ is the spatial frequency, ω_{t_j} the temporal frequency, and the 3D Gaussian envelope has standard deviations $\sigma_x, \sigma_y, \sigma_t$.

The Fourier transform of this Gabor filter is given by [5]:

$$\begin{aligned} \mathcal{F}_{\mathrm{g}_j}(\omega_x, \omega_y, \omega_t) = \frac{i}{2}\{ & e^{-2\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} - e^{-2\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2} + \\ & e^{-2\pi^2\sigma_y^2(\omega_y-\omega_{y_j})^2} - e^{-2\pi^2\sigma_y^2(\omega_y+\omega_{y_j})^2} + \\ & e^{-2\pi^2\sigma_t^2(\omega_t-\omega_{t_j})^2} - e^{-2\pi^2\sigma_t^2(\omega_t+\omega_{t_j})^2}\} \end{aligned} \tag{10}$$

The energy of feature responses to the filter of (9) can be computed if its power spectral density (PSD) is known. As the filter is separable in its three dimensions, we first illustrate the computation

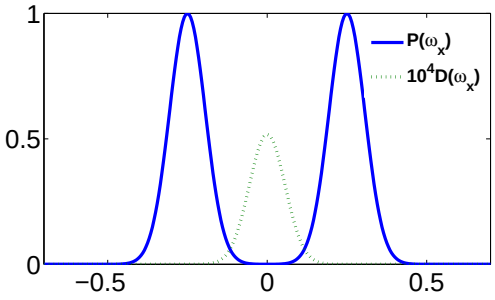

Figure 1: Approximation of the PSD of a sine phase Gabor filter in 1D. The thick blue curve shows the quantity in (14) for typical values of σ_x and ω_{x_j}, and the dotted curve shows 10^4 times the difference between the quantities in (14) and (13)

of PSD of a 1-D Gabor filter, $\mathrm{g}(x)$, from its Fourier response, $\mathcal{F}_{\mathrm{g}}(\omega_x)$ [5]:

$$\mathrm{g}(x) = \frac{1}{\sqrt{2}\pi^{\frac{3}{2}}\sigma_x}\sin(2\pi\omega_{x_j}x)e^{-\frac{x^2}{2\sigma_x^2}} \tag{11}$$

$$\mathcal{F}_{\mathrm{g}}(\omega_x) = \frac{i}{2}\{e^{-2\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} - e^{-2\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2}\} \tag{12}$$

The PSD of the filter is then

$$\begin{aligned}|\mathcal{F}_{\mathrm{g}}(\omega_x)|^2 &= \frac{1}{4}\{e^{-4\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} + e^{-4\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2} + 2e^{-2\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2-2\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2}\}\\ &= \frac{1}{4}\{e^{-4\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} + e^{-4\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2} + e^{-4\pi^2\sigma_x^2(\omega_x^2+\omega_{x_j}^2)}\} \qquad (13)\\ &\approx \frac{1}{4}\{e^{-4\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} + e^{-4\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2}\} \qquad (14)\end{aligned}$$

where the third term,

$$D(\omega_x) = e^{-4\pi^2\sigma_x^2(\omega_x^2+\omega_{x_j}^2)}, \tag{15}$$

can be ignored because it is upper-bounded by $e^{-4\pi^2\sigma_x^2\omega_{x_j}^2}$, a quantity that is much smaller than 1. This is illustrated in Figure 1.

Similarly, the PSD of the 3D Gabor filter can be given by,

$$\begin{aligned}|\mathcal{F}_{\mathrm{g}_j}(\omega_x,\omega_y,\omega_t)|^2 &\approx \frac{1}{4}\{e^{-4\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} + e^{-4\pi^2\sigma_x^2(\omega_x+\omega_{x_j})^2} + \qquad (16)\\ &\quad \frac{1}{4}\{e^{-4\pi^2\sigma_y^2(\omega_y-\omega_{y_j})^2} + e^{-4\pi^2\sigma_y^2(\omega_y+\omega_{y_j})^2}\} + \\ &\quad \frac{1}{4}\{e^{-4\pi^2\sigma_t^2(\omega_t-\omega_{t_j})^2} + e^{-4\pi^2\sigma_t^2(\omega_t+\omega_{t_j})^2}\}\\ &= P_j(\omega_x,\omega_y,\omega_t) \qquad (17)\end{aligned}$$

For a sinusoidal grating moving with a given velocity, $\bar{v}_k = v_{kx}\hat{e}_x + v_{ky}\hat{e}_y$, its energy in the frequency domain is contained in a plane defined by [12]:

$$v_{kx}\omega_x + v_{ky}\omega_y - \omega_t = 0 \tag{18}$$

To construct a unit tuned to velocity $\bar{v}_k$ we compute a weighted combination of the outputs of a set of 3D Gabor filters following the approach of [9]. The weight assigned to each Gabor filter in the set is in proportion to the energy contained in the intersection between the PSD of the filter and the plane corresponding to $\bar{v}_k$. This can be computed as:

$$\begin{aligned}w_{jk}(\omega_{x_j},\omega_{y_j},\omega_{t_j},\bar{v}_k) &\propto \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P_j(\omega_x,\omega_y,\omega_t)\; d\omega_x d\omega_y d\omega_t|_{v_{kx}\omega_x+v_{ky}\omega_y-\omega_t=0} \qquad (19)\\ &= \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(e^{-2\pi^2\sigma_x^2(\omega_x-\omega_{x_j})^2} + e^{-2\pi^2\sigma_y^2(\omega_y-\omega_{y_j})^2} + e^{-2\pi^2\sigma_t^2(v_{kx}\omega_x+v_{ky}\omega_y-\omega_{t_j})^2}\right) d\omega_x d\omega_y\end{aligned}$$

where the second step follows due to the symmetry between the two lobes of the PSD. The integral can be evaluated using the procedure outlined in [6].

In this work we use a total of 12 spatio-temporal filters, each with center frequency $(\omega_{x_j}, \omega_{y_j}, \omega_{t_j})$, $j = 0 \ldots 11$. We consider 12 neurons each tuned to motion with constant speed in one of 12 different directions spread uniformly in $(0^\circ, 360^\circ)$, corresponding to velocities $\bar{v}_k$, $k = 0 \ldots 11$.

w_{jk} is then the weight assigned to the j^{th} filter for computing motion in the k^{th} direction.

2 Derivation of Equation (7) in the submission

From Section 3.1 in the main submission, we have

$$P_{C(l^*)|F_k}(1|1) = 2S_k(l^*), \tag{20}$$

Denote the state of F_k and l^* at time t by F_k^t and l_t^*, respectively, and the sequence of target locations till time t by $\mathbf{l}_t^* = (l_t^*, l_{t-\tau}^* \ldots l_0^*)$. Using Bayes rule, the posterior probability of feature F_k being the most salient feature can be written as,

$$P_{F_k^t|C(\mathbf{l}_t^*)}(1|\mathbf{1}) = P_{F_k^t|C(l_t^*),C(\mathbf{l}_{t-\tau}^*)}(1|1,\mathbf{1}) \propto P_{C(l_t^*)|F_k^t,C(\mathbf{l}_{t-\tau}^*)}(1|1,\mathbf{1}) P_{F_k^t|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1}) \tag{21}$$

We do not assume an explicit motion model or motion extrapolation. However it is reasonable to assume that the probability of finding the target is uniformly distributed in a small neighborhood around $l_{t-\tau}^*$, and if the velocity of the target is not too high, l_t^* is in this neighborhood. We can then write,

$$P_{C(l_t^*)|F_k^t,C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1}) \propto P_{C(l_t^*)|F_k^t}(1|1) \tag{22}$$

Using this in (21), we get,

$$\begin{aligned}
P_{F_k^t|C(\mathbf{l}_t^*)}(1|\mathbf{1}) &\propto P_{C(l_t^*)|F_k^t}(1|1) P_{F_k^t|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1}) && (23)\\
&\propto P_{C(l_t^*)|F_k^t}(1|1) \sum_i P_{F_k^t,F_i^{t-\tau}|C(\mathbf{l}_{t-\tau}^*)}(1,1|\mathbf{1}) && (24)\\
&\propto P_{C(l_t^*)|F_k^t}(1|1) \sum_i P_{F_k^t|F_i^{t-\tau},C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1}) P_{F_i^{t-\tau}|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1}) && (25)
\end{aligned}$$

The probabilities $P_{F_k^t|F_i^{t-\tau},C(\mathbf{l}_{t-\tau}^*)}(1|1,\mathbf{1})$ encode the likelihood of transition from state i to state k. Since dominant features of the target tend to stay dominant for some time in the neighborhood of the last known position of the target, we assume $P_{F_k^t|F_i^{t-\tau},C(\mathbf{l}_{t-\tau}^*)}(1|1,\mathbf{1}) = 1$ if $i = k$ and null otherwise (this is the likelihood of transition only using the information from the previous time step, it does not preclude new features from being selected if they become salient at t). Using this, and (20), in (25) we get the recursion,

$$P_{F_k^t|C(\mathbf{l}_t^*)}(1|\mathbf{1}) = \frac{S_k(l_t^*) P_{F_k^{t-\tau}|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1})}{\sum_j S_j(l_t^*) P_{F_j^{t-\tau}|C(\mathbf{l}_{t-\tau}^*)}(1|\mathbf{1})}. \tag{26}$$

References

[1] R. Born and D. Bradley. Structure and Function of Visual Area MT. *Annu. Rev. Neurosci*, 28:157–89, 2005.

[2] R. T. Born and R. B. H. Tootell. Segregation of global and local motion processing in primate middle temporal visual area. *Nature*, 357:497–499, 1992.

[3] M. Carandini, J. Demb, V. Mante, D. Tolhurst, Y. Dan, B. Olshausen, J. Gallant, and N. Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25:10577–10597, 2005.

[4] D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principle, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, 21:239–271, Jan 2009.

[5] D. Heeger. Models for motion perception. *Ph.D.thesis, Univ. of Pennsylvania, 1987.*, 1987.

[6] D. Heeger. Optical flow from spatiotemporal filters. *International Journal of Computer Vision*, 1(4):279–302, 1988.

[7] J. Huang and D. Mumford. Statistics of Natural Images and Models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 541–547, 1999.

[8] N. Rust, V. Mante, E. Simoncelli, and J. Movshon. How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, 9(11):1421–1431, 2006.

[9] E. Simoncelli and D. Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761, 1998.

[10] K. Tanaka, K. Hikosaka, H. Saito, M. Yukie, Y. Fukada, and E. Iwai. Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey. *Journal of Neuroscience*, 6(1):134, 1986.

[11] M. Vasconcelos and N. Vasconcelos. Natural image statistics and low-complexity feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):228–244, 2008.

[12] A. Watson and A. Ahumada, Jr. Model of human visual-motion sensing. *Journal of the Optical Society of America A*, 2(2):322–341, 1985.