A Hierarchical Model of Binocular Rivalry

Peter Dayan

Department of Brain and Cognitive Sciences
E25-210 Massachusetts Institute of Technology
Cambridge, MA 02139
dayan@psyche.mit.edu

June 5, 1997

Abstract

Binocular rivalry is the alternating percept that can result when the two eyes see different scenes. Recent psychophysical evidence supports the notion that some aspects of binocular rivalry bear functional similarities to other bistable percepts. We build a model based on the hypothesis (Logothetis & Schall, 1989; Leopold & Logothetis, 1996; Logothetis, Leopold & Sheinberg, 1996) that alternation can be generated by competition between top-down cortical explanations for the inputs, rather than by direct competition between the inputs. Recent neurophysiological evidence shows that some binocular neurons are modulated with the changing percept; others are not, even if they are selective between the stimuli presented to the eyes. We extend our model to a hierarchy to address these effects.

If one's eyes are presented with two different, but very low contrast stimuli, then the overall percept is of the sum or composition of the stimuli (Liu, Tyler & Schor, 1992). However, as the stimuli are given higher contrast, there comes a point when it appears as if the inputs from the eyes rival – first one dominates, then the other, with stochastic switching between the two. If the stimuli are large, then one single stimulus may not dominate across the entire field, but rather there will be a mosaic of patches, with different stimuli dominating in each patch (Wheatstone, 1838; Levelt, 1965). The dynamics of rivalry are sensitive to the contrast of the stimuli in the eyes if they are different, with such characteristic results as that increasing the contrast of one stimulus decreases the time during which that stimulus is suppressed much more than it increases the time that it is dominant (Levelt, 1965; Fox & Rasche, 1969; Blake, 1977; Mueller & Blake, 1989). There are also effects of the nature of the stimuli – for instance if two separate patterns are divided up between the two stimuli, then, in certain cases, the *patterns* will rival rather than the stimuli directly (Whittle, Bloor, Pocock, 1968; Kovacs, Papathomas, Yang & Feher, 1996), and there is some evidence that familiar patterns enjoy an advantage over unfamiliar ones during rivalry (see Yu & Blake, 1992).

It is natural to suppose that this rivalry is instantiated in the parts of the visual pathway that are still monocular, *ie* the lateral geniculate nucleus (LGN) and layer IV of V1. Indeed most models of rivalry implicitly or explicitly make this assumption (*eg* Matsuoka, 1984; Lehky, 1988; Blake, 1989; Mueller, 1990), using various forms of reciprocal inhibition between two pathways and thus capturing many of the intricacies of the dynamics of rivalry. It turns out that the activities of neurons in the LGN are not affected by rivalry (Lehky & Maunsell, 1996), leaving layer IV of V1 as the candidate for this class of models.

These models could be augmented with some top-down processing to capture the familiarity and pattern-based effect. However, they are directly challenged by psychophysical data from Logothetis, Leopold & Sheinberg (1996) and are hard to reconcile with the neurophysiological data from Schall & Logothetis (1989) and Leopold & Logothetis (1996). Logothetis, Leopold & Sheinberg (1996) switched rivalrous patterns quickly between the two eyes (see also Blake, Westendorf & Overton, 1980) whilst constantly flickering the stimuli. Subjects report that the perceptual switching time is much greater than the actual switching time, which is inconsistent with the hypothesis that there is a dominant *eye* rather than a dominant *pattern.* Of course, there could be eye-based competition as well (Wales & Fox, 1970; Fox & Check, 1972; Blake & Fox, 1974).

Leopold & Logothetis (1996) trained monkeys to report their percept during rivalrous and non-rivalrous stimuli whilst recording from neurons V1/2 and V4. They found that striate monocular neurons are unaffected by rivalry; that there are binocular neurons in all areas that are sensitive between the stimuli during binocular presentation and whose activities *are not* modulated with the monkey's percept; that there are binocular neurons in all areas that are sensitive between the stimuli during binocular presentation and whose activities *are* modulated with the monkey's percept; and that there are binocular neurons in all areas that are sensitive between the stimuli during binocular presentation whose activities are elevated during perceptual *suppression* of their preferred stimuli, and also binocular neurons that are not selective between the stimuli in binocular viewing, but whose activities are nevertheless modulated during rivalry.

Logothetis and his colleagues have long suggested an account of rivalry under which it is cortical *explanations* of sensory input that compete rather than the inputs themselves. Various recent models of cortical processing are based on the old notion of analysis by synthesis (MacKay, 1956; Grenander, 1976; Mumford, 1994; Carpenter & Grossberg, 1987; Pece, 1992; Hinton *et al,* 1995; Dayan *et al,* 1995; Olshausen & Field, 1996; Rao & Ballard, 1997). For these, the synthetic model, which is usually instantiated in top-down connections in cortex, exactly constructs a top-down explanation for input, and an analysis procedure finds which particular synthetic explanation is appropriate for a given input. In this paper we consider one form of analysis-by-synthesis model and show how it can exhibit rivalry between explanations in the case that the eyes receive different input. This model can provide an account for many of the behaviours described above.

Section 2 discusses a simplified case of rivalry to illustrate the basic principles of the model; section 3 describes a more complete model with three layers of units in a hierarchy; the implications of the model are discussed in section 4.

1 The Simple Model

Figure 1a shows a simple abstract model illustrating competition between cortical explanations. w_1 and w_2 model two binary-valued striate units and layer $\mathbf{z}$ models 32 binary-valued geniculate units, 16 each for left (L) and right (R) eyes. In the generative model, turning w_1 on activates two binocular horizontal bars in the input $\mathbf{z}$, and we therefore say that w_1 *explains* the input activity, if the input were really to consist of two binocular horizontal bars. Similarly, the activity of w_2 explains two binocular vertical bars in the input. More formally, the explanations arise as the analysis or recognition phase of an analysis-by-synthesis model of cortical function. The top-down, synthetic, model specifies successively probabilities $\mathcal{P}[\mathbf{w}]$ and $\mathcal{P}[\mathbf{z}|\mathbf{w}]$ according to:

$$\begin{aligned}\mathcal{P}[w_k{=}1] &= \sigma(b_w) \\ \mathcal{P}[z_i{=}1|\mathbf{w}] &= \sigma\left(b_z + \sum_{k=1}^{2} w_k J^{ki}_{\mathbf{wz}}\right)\end{aligned} \tag{1}$$

where

$$\sigma(x){=}\frac{1}{4000}\left\{1 + \frac{3998}{\sqrt{2\pi}}\int_{t=-\infty}^{x} e^{-t^2/2}dt\right\}$$

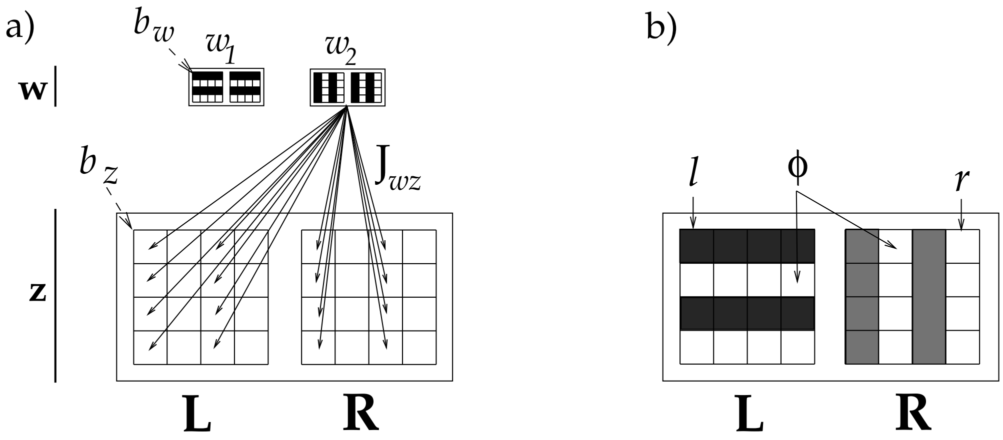

Figure 1: a) Simple generative model. Units w_1 and w_2 are shown in the form of their projective fields (*eg* w_2 generates two vertical bars binocularly in the 4×4 left (L) and right (R) geniculate units, $\mathbf{z}$), together with a subset of the weights. The other weights follow similarly. $b_w{=}-2$ and $b_z{=}-3$ are the generative biases, $J_{\mathbf{wz}}{=}5.8$ are the generative weights from $\mathbf{w}$ to $\mathbf{z}$. b) Rivalrous input pattern. Horizontal input is provided to the geniculate units for the left eye ($z_1 \ldots z_{16}$), with strength l (*eg* $\mathcal{P}[z_1{=}1]{=}\sigma(l)$); vertical input to those for the right eye ($z_{17} \ldots z_{32}$), with strength r. Silent units have input ϕ such that $\sigma(\phi){=}0.01$.

is a normal distribution function, squashed to avoid infinities, and all the w_k and z_i are independent given $\mathbf{w}$. The parameters of the generative model, the weights $J_{\mathbf{wz}}$ and the biases $b_{\mathbf{w}}$ and $b_{\mathbf{z}}$ are shown in the diagram. They were set by hand such that, in the generative model, w_1 and w_2 are active only rarely (*ie* activity in the $\mathbf{w}$ layer is sparse), but are almost sure to produce their favoured pattern in $\mathbf{z}$ if they do fire. In general these weights would be learned from experience of horizontal and vertical contours (Hinton *et al*, 1995; Dayan *et al*, 1995; Saul, Jaakkola & Jordan, 1996).

Since the units in the generative model are binary, we cannot model differing input contrasts directly by changing the level of activity of the z_i. Rather, we represent the input to z_i as d_i, where $\mathcal{P}[z_i{=}1] = \sigma(d_i)$ and all the z_i are independent. Write $\mathcal{P}[\mathbf{z}; \mathbf{d}]$ as the induced probability distribution over the input units.

Recognition is formally the statistical inverse to generation. For the network in figure 1a it should produce $\mathcal{P}[\mathbf{w}|\mathbf{z}]$ over the four choices for $\mathbf{w}$, given a particular input. In this case it would be easy to calculate these probabilities exactly. However, in general this is computationally intractable, since if there are n units, then there are 2^n probabilities. Further, we require a way of representing of these 2^n probabilities in terms of just the activities of the n units. Inspired by Saul, Jaakkola & Jordan (1996) and Jaakkola, Saul & Jordan (1996), we achieve both of these by using a mean field inversion method. This approximates $\mathcal{P}[\mathbf{w}|\mathbf{z}]$ by the parameterised factorial form

$$p\mathcal{Q}[\mathbf{w};\boldsymbol{\mu}] = \prod_i \sigma(\mu_i)^{w_i}\,(1-\sigma(\mu_i))^{1-w_i}\,. \tag{2}$$

This sets the mean activity of w_1 to be $\sigma(\mu_1)$. Note that μ_i are real values which, through equation 2, parameterise a distribution over the binary-valued w_i. We model the activities of cells as the real-valued $\sigma(\mu_i)$.

Mean field methods would use a descent method to optimise over the parameters $\boldsymbol{\mu}$ to minimise the mean Kullback-Leibler divergence between $\mathcal{Q}[\mathbf{w};\boldsymbol{\mu}]$ and $\mathcal{P}[\mathbf{w}|\mathbf{z}]$:

$$\mathcal{F}[\boldsymbol{\mu}] = \sum_{\mathbf{z}} \mathcal{P}[\mathbf{z};\mathbf{d}] \sum_{\mathbf{w}} \mathcal{Q}[\mathbf{w};\boldsymbol{\mu}] \log \frac{\mathcal{Q}[\mathbf{w};\boldsymbol{\mu}]}{\mathcal{P}[\mathbf{w}|\mathbf{z}]}.$$

The simplest model of gradient descent has:

$$\begin{aligned}
\mu_1(t+1) &= \mu_1(t) - \delta\nabla_{\mu_1}\mathcal{F}[\boldsymbol{\mu}(t)] \\
&= \mu_1(t) - \delta(\log\left[\frac{\sigma(\mu_1(t))}{\sigma(-\mu_1(t))}\frac{\sigma(-b_w)}{\sigma(b_w)}\right] + \\
&\quad \sigma'(\mu_1(t))\,(P_{10} - P_{00} + \sigma(\mu_2(t))[P_{11} - P_{10} - P_{01} + P_{00}]))
\end{aligned} \tag{3}$$

where

$$P_{ab} = \sum_i \sigma(d_i)\log\mathcal{P}[z_i{=}1|w_1{=}a, w_2{=}b] + \sigma(-d_i)\log\mathcal{P}[z_i{=}0|w_1{=}a, w_2{=}b]$$

and δ acts like an adaptation rate.

In this simple case, calculating these terms only requires operations local to each unit, although the operations are somewhat complicated. Jaakkola, Saul & Jordan (1996) provide a further approximation that simplifies these calculations in which unit z_i passes back to w_1 and w_2 information about how it is incorrectly predicted by w_1 and w_2:

$$\nabla_{\mathbf{w}}\left(d_i - b_z - \sum_k \sigma(w_k) J^{ki}_{\mathbf{wz}}\right)^2.$$

We found this model to work slightly less well.

The joint use of bottom-up and top-down models of processing is inspired by the hierarchical and reciprocal connectivity in cortex (Felleman & Van Essen, 1991) and is ubiquitous in this class of analysis-by-synthesis based models. Note that the mean-field model only affects the activities in the $\mathbf{w}$ layer and does not affect the inputs, even though there are top-down inputs to those units. Lehky & Maunsell (1996) resolved conclusively that the activities of neurons in the LGN of macaque monkeys are *not* modulated during rivalry (see Varela & Singer, 1987).

If a non-rivalrous input is presented, with just horizontal bars in both channels, then recognition assigns full responsibility to w_1. Rivalry results when different inputs are presented to the two eyes. For inputs such as figure 1b, $\{w_1{=}1; w_2{=}0\}$ and $\{w_1{=}0; w_2{=}1\}$ are equally good explanations (albeit worse than in the non-rivalrous case). Explanation $\{w_1{=}0; w_2{=}0\}$ is poor because it does not account for any input; $\{w_1{=}1; w_2{=}1\}$ is poor because activity across $\mathbf{w}$ should be sparse, according to the generative model, and $w_1{=}1$ *explains away* (Pearl, 1988) the need for $w_2{=}1$ for those elements of $\mathbf{z}$ that are common between horizontal and vertical bars. Note that w_1 and w_2 compete even though there are no explicit inhibitory interconnections between them in the *generative* model.

Note that the recognition model of a Helmholtz machine (Hinton *et al*, 1995) is *unsuitable* to model rivalry, since it acts in a purely bottom-up direction and therefore lacks the capacity to capture explaining away (Dayan & Hinton, 1996), on which this model of rivalry crucially depends. This is one reason why we used a mean field method instead (Saul, Jaakkola & Jordan, 1996).

If the dynamics were just determined by equation 3, then the activities would tend to one of the two equivalently good explanations (which are global minima of $\mathcal{F}$) and just stay there. We therefore implemented a simple oscillatory model with auxiliary variables $\mu'_k(t)$ implementing a form of fatigue process. The full dynamics for $\mu_1(t)$ and $\mu'_1(t)$ are:

$$\begin{aligned}
\mu_1(t+1) &= \mu_1(t) + \delta(-\nabla_{\mu_1}\mathcal{F}[\boldsymbol{\mu}(t)] + \alpha(\beta\mu_1(t)) - \mu'_1(t)) \\
\mu'_1(t+1) &= \mu'_1(t) + \delta(\mu_1(t) - \beta\mu'_1(t)),
\end{aligned}$$

where β is a decay term. A similar equation applies for $\mu_2(t)$ and $\mu'_2(t)$. In all the simulations, $\alpha{=}0.5, \beta{=}0.1$ and $\delta{=}0.01$. Factor $\frac{1}{\delta}$ now plays the role of a time constant for the network. μ_2 follows similar dynamics. As with most models of rivalry (see Lehky 1988 for a notable exception), we are modeling data on the mean dominance times and are ignoring the stochasticity of the data.

Based on this simple oscillatory process, the model effectively switches between horizontal ($\{w_1=1; w_2=0\}$) and vertical ($\{w_1=0; w_2=1\}$) explanations. Figure 2a shows the resulting activities of w_1 and w_2 for a case in which the strength of the input to the horizontal bars (l) is stronger than to the vertical (r). Alternations ensue, with a greater dominance period for w_1 than w_2. Figure 2b shows that, as empirically observed, when the input strengths for both patterns are increased together (modeling increasing contrast), the oscillations speed up (Levelt, 1965; Fox & Rasche, 1969), and when just r is varied, it has a significantly greater effect on the period for which the vertical explanation is *suppressed* (*ie* the horizontal explanation is dominant) than on the period for which it is *dominant* (Levelt, 1965; Fox & Rasche, 1969; Blake, 1977; Mueller & Blake, 1989; Leopold & Logothetis, 1996). This achieves the effect of mutual inhibition (Fox & Rasche, 1969; Matsuoka, 1984; Lehky, 1988; Mueller, 1990) between w_1 and w_2, dependent on input contrast (Mueller, 1990) by statistically justifiable means. Furthermore, for very weak inputs, both w_1 and w_2 are weakly activated, which is the model's account of the psychophysical observation that fusion rather than rivalry occurs for very low contrast stimuli. Also, if the eyes are provided with binocularly consistent inputs within a reasonable range of contrast, then the system does not oscillate.

We have therefore shown that it is possible to get rivalry between cortical explanations for input, using a mean field inversion method for a top-down generative model. In this case, the final model resembles existing models for rivalry in which there is competition amongst binocular oriented units rather than within in a monocular system (Grossberg, 1987). Indeed, Sengpiel, Blakemore & Harrad (1995) studied interocular suppression of activity in binocular cells when the two eyes were presented with gratings of orthogonal orientations. In the mean-field model, this suppression arises as a consequence of explaining away during the process of recognition, and has a precise relationship with the underlying top-down generative model.

2 The Hierarchical Model

The simple model is too small to be able to have populations of units that are and are not modulated with rivalry, as in the neurophysiological data. We therefore extended it to a hierarchy of units covering a larger spatial array, incorporating various characteristics of

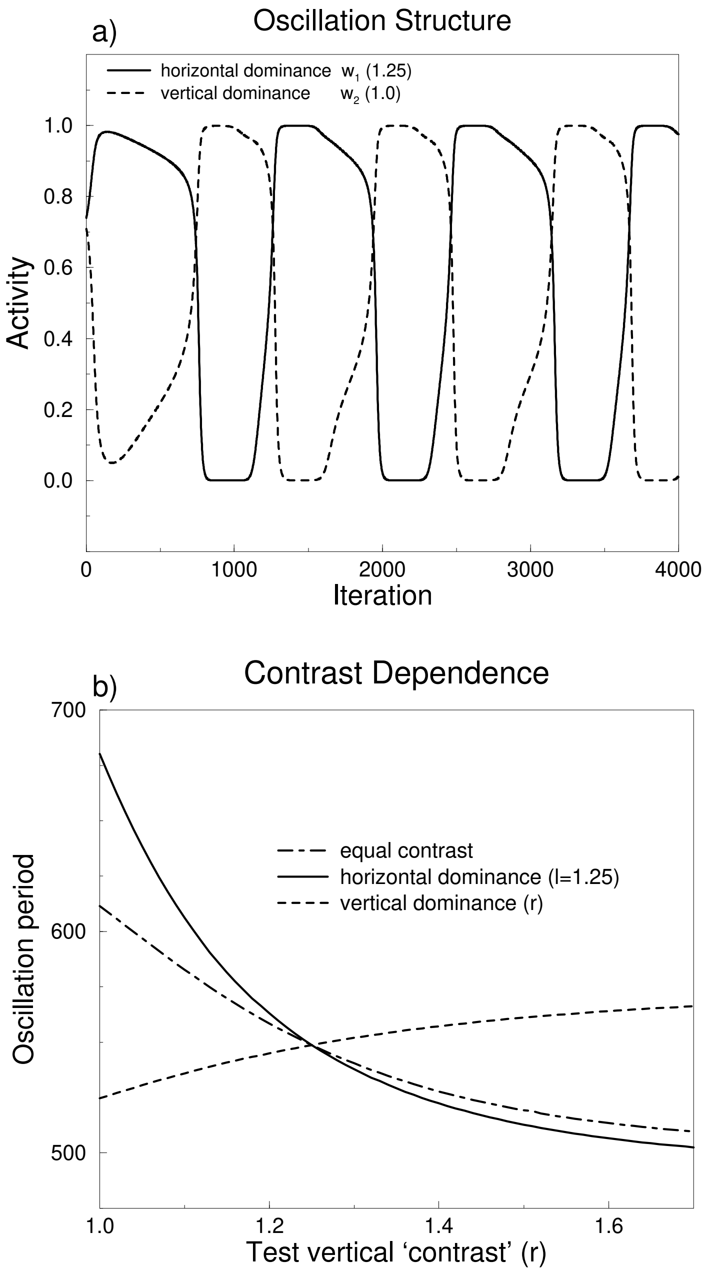

Figure 2: a) Development and maintenance of oscillations in the mean activities of $w_1 = \sigma(\mu_1)$ and $w_2 = \sigma(\mu_2)$ over time. Input strengths $l = 1.25$ and $r = 1.0$, so the horizontal bars dominate. b) Dependence on the input strength (modeling contrast) in l and r of the periods of suppression and dominance. Horizontal patterns are taken as dominant when the mean activity of w_1 is greater than the mean activity of w_2 — there is no switching reaction time. For the equal contrast case, l and r were varied together; for suppression and dominance plots, $l = 1.25$ was constant; r was varied.

cortical visual processing. Figure 3 shows the full generative model.

Units in layers $\mathbf{y}$ (modeling V1) and $\mathbf{x}$ and $\mathbf{w}$ (modeling early and late extra-striate areas) are all binocular and jointly explain successively more complex features in the input $\mathbf{z}$ according to a top-down generative model. Apart from the half bars in $\mathbf{y}$, the *generative* model is similar to that learned by the Helmholtz machine (Dayan *et al*, 1995) for which increasing complexity in higher layers rather than the increasing input scale is key.[1] In this case, for instance, w_2 specifies the occurrence of vertical bars anywhere in the 8×8 input grids; x_{16} specifies the rightmost vertical bar; and y_{31} and y_{32} the top and bottom half of this vertical bar. Again, these specifications are provided by a top-down generative model in which, as in equation 1, the activations of units are specified by probabilities such as:

$$\mathcal{P}[y_i = 1|\mathbf{x}] = \sigma\left(b_y + \sum_k x_k J_{\mathbf{xy}}^{ki}\right)$$

where the sum k is over all the units in the $\mathbf{x}$ layer.

In this more complicated model, activities of units in different layers could conflict. For instance, unit w_1 could be activated, suggesting that there are horizontal bars in the input; but units x_{15} and x_{16} could also be active, suggesting that there are two particular vertical bars. Such patterns of activity are unlikely, since they are inconsistent with the generative model, and we never observed them with the settings of the weights that we adopted. We therefore model the percept of the network as the activity in the $\mathbf{w}$ layer.

The same mean field method is used to perform recognition in this hierarchical model. The equivalent mean field distribution is:

$$\mathcal{Q}[\mathbf{w}, \mathbf{x}, \mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\psi}] = \mathcal{Q}[\mathbf{w}; \boldsymbol{\mu}]\mathcal{Q}][\mathbf{x}; \boldsymbol{\xi}]\mathcal{Q}[\mathbf{y}; \boldsymbol{\psi}]$$

which renders independent all the units in the model. The equivalent of $\mathcal{F}$ now depends on $\boldsymbol{\mu}, \boldsymbol{\xi}$ and $\boldsymbol{\psi}$:

$$\mathcal{F}[\boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\psi}] = \sum_{\mathbf{z}} \mathcal{P}[\mathbf{z}; \mathbf{d}] \sum_{\mathbf{w},\mathbf{x},\mathbf{y}} \mathcal{Q}[\mathbf{w}, \mathbf{x}, \mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\psi}] \log \frac{\mathcal{Q}[\mathbf{w}, \mathbf{x}, \mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\psi}]}{\mathcal{P}[\mathbf{w}, \mathbf{x}, \mathbf{y}|\mathbf{z}]}.$$

We adopted various heuristics to simplify the process of using this rather cumbersome mean field model. First, fatigue is only implemented for the units in the $\mathbf{y}$ layer, and

[1] Although the *recognition* model of the Helmholtz machine is not used, since it does not capture explaining away.

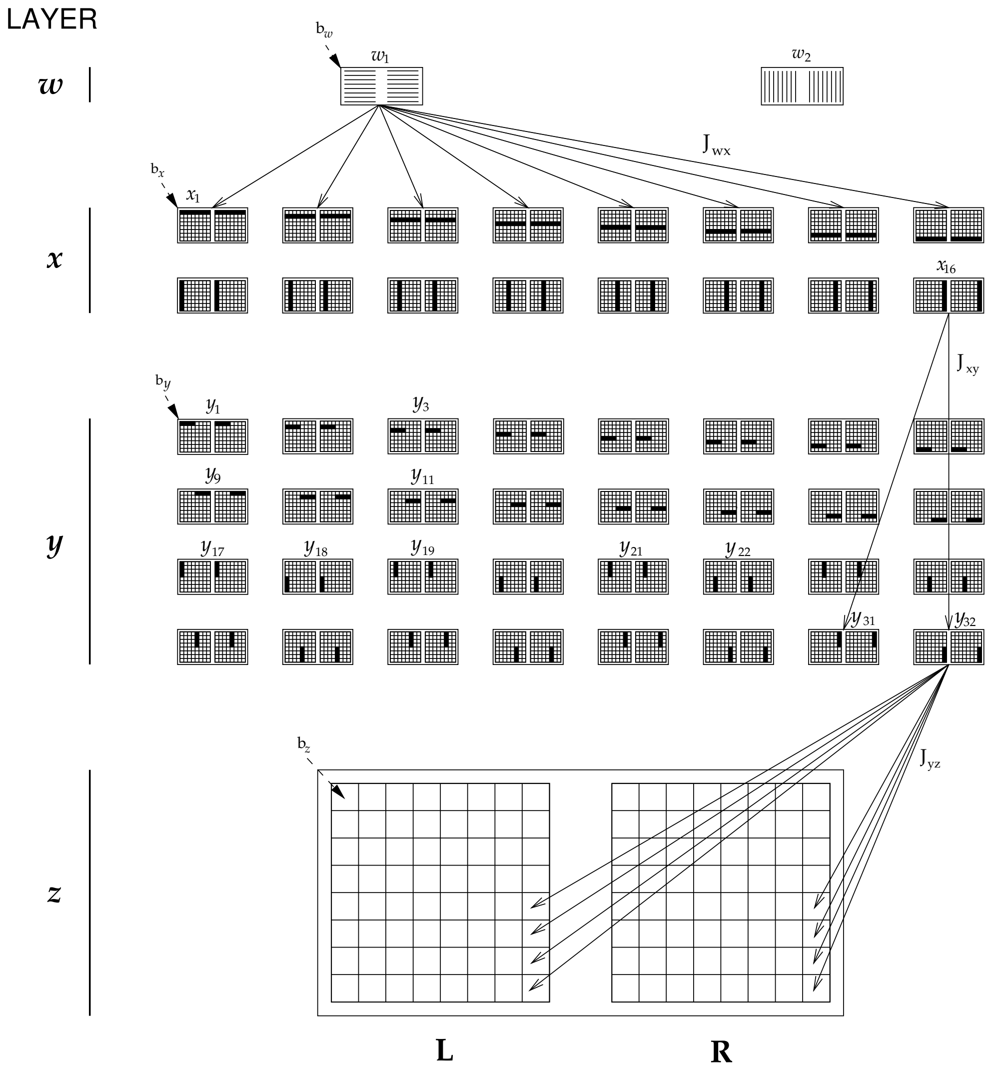

Figure 3: Hierarchical generative model for 8×8 bar patterns across the two eyes. Units are depicted by their net projective (generative) fields, and characteristic weights are shown. Even though the net projective field of x_1 is the top horizontal bar in both eyes, note that it generates this by increasing the probability that units y_1 and y_9 in the **y** layer will be active, not by having direct connections to the input **z**. Unit w_1 connects to $x_1, x_2, \ldots x_8$ through $J_{\mathbf{wx}}=0.8$; x_{16} connects to y_{31}, y_{32} through $J_{\mathbf{xy}}=1.0$ and y_{32} connects to the bottom right half vertical bar through $J_{\mathbf{yz}}=5.8$. Biases are $b_{\mathbf{w}}=-0.75$, $b_{\mathbf{x}}=-1.5$, $b_{\mathbf{y}}=-2.7$ and $b_{\mathbf{z}}=-3.3$. Each unit in the **z** layer is really a pair of units (as in Hinton *et al*, 1995), to increase the strength of the signal.

the ψ follow the equivalent of the dynamical equations above. Although adaptation processes can clearly occur at many levels in the system, their exact form is not clear. Bialek & DeWeese (1995) argue that the rate of a switching process should be adaptive to the expected rate of change of the associated signal on the basis of prior observations. This is clearly faster nearer to the input.

The second heuristic is that rather than perform gradient descent for the non-fatiguing units, the optimal values of $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$ are calculated on each iteration by solving numerically equations such as

$$\nabla_{\xi_i}\mathcal{F}[\boldsymbol{\mu},\boldsymbol{\xi},\boldsymbol{\psi}] = 0.$$

The dearth of connections in the network of figure 3 allows $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$ to be calculated locally at each unit in an efficient manner. Whether this is reasonable depends on the time constants of settling in the mean field model with respect to the dynamics of switching, and, more particularly on the way that this deterministic model is made appropriately stochastic.

Figure 4 shows the activities of units in response to binocular horizontal (i) and vertical (ii) bars. In these cases there are no oscillations. Figure 4 also shows the two equally likely explanations for rivalrous input (iii and iv). For rivalry, there is direct competition in the top left hand quadrant of $\mathbf{z}$, as in figure 1, which is reflected in the competition between y_1, y_3 and y_{17}, y_{21}. However, the input regions (top right of L and bottom left of R) for which there is no competition, require the constant activity of explanations y_9, y_{11}, y_{18} and y_{22}. Under the generative model, the coactivation of y_1 and y_9 *without* x_1 is quite unlikely ($\mathcal{P}[x_1{=}0|y_1{=}1, y_3{=}1] = 0.1$), which is why x_1, x_3 and also w_1 become active with y_1 and y_3.

Figure 5a shows the resulting activities during rivalry of units at various levels of the hierarchy including the fatigue process. Broadly, the competing explanations in figure 4(iii;iv), *ie* horizontal and vertical percepts, alternate, and units without competing inputs, such as y_9, are much less modulated than the others, such as y_1. The activity of y_9 is slightly elevated when horizontal bars are dominant, based on top-down connections. The activities of the units higher up, such as x_1 and w_1, do not decrease to 0 during the suppression period for horizontal bars, leaving weak activity during suppression. Leopold & Logothetis (1996) observed that many of their modulating cells were not completely silent during their periods of less activity. Figure 5b shows that the hierarchical version of the model

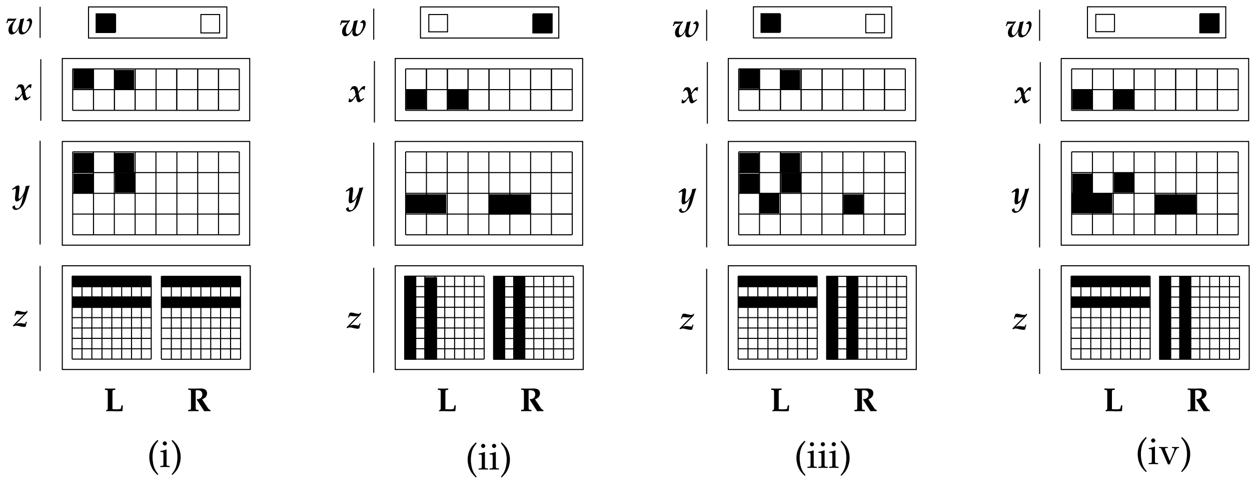

Figure 4: Recognition activity in the network for four different input patterns. The units are arranged in the same order as figure 3, and white and black squares imply activities for the units whose means are less than and greater than 0.5. (i) and (ii) represent normal binocular stimulation; (iii) and (iv) show the two alternative stable states during rivalrous stimulation, without the fatigue process.

also behaves in accordance with experimental results on the effects of varying the input contrast (Levelt, 1965; Fox & Rasche, 1969; Blake, 1977; Mueller & Blake, 1989; Leopold & Logothetis, 1996).

3 Discussion

Following Logothetis and his colleagues (Logothetis & Schall, 1989; Leopold & Logothetis, 1996; Logothetis *et al*, 1996; see also Grossberg, 1987) we have suggested an account of rivalry based on competing top-down hierarchical explanations. Neurons explain inputs in virtue of being capable of generating their activities through a top-down statistical generative model. Competition arises between higher-level explanations of overlapping active regions (*ie* those involving contrast changes) of the input rather than between inputs themselves.

The overall model mechanistically has much in common with models which place the competition in rivalry at the level of binocular oriented cells rather than between monocular cells (see Grossberg, 1987; Blake, 1989). Indeed, the model is based on an explanation-driven account for normal binocular processing, so this is to be expected. The advantage

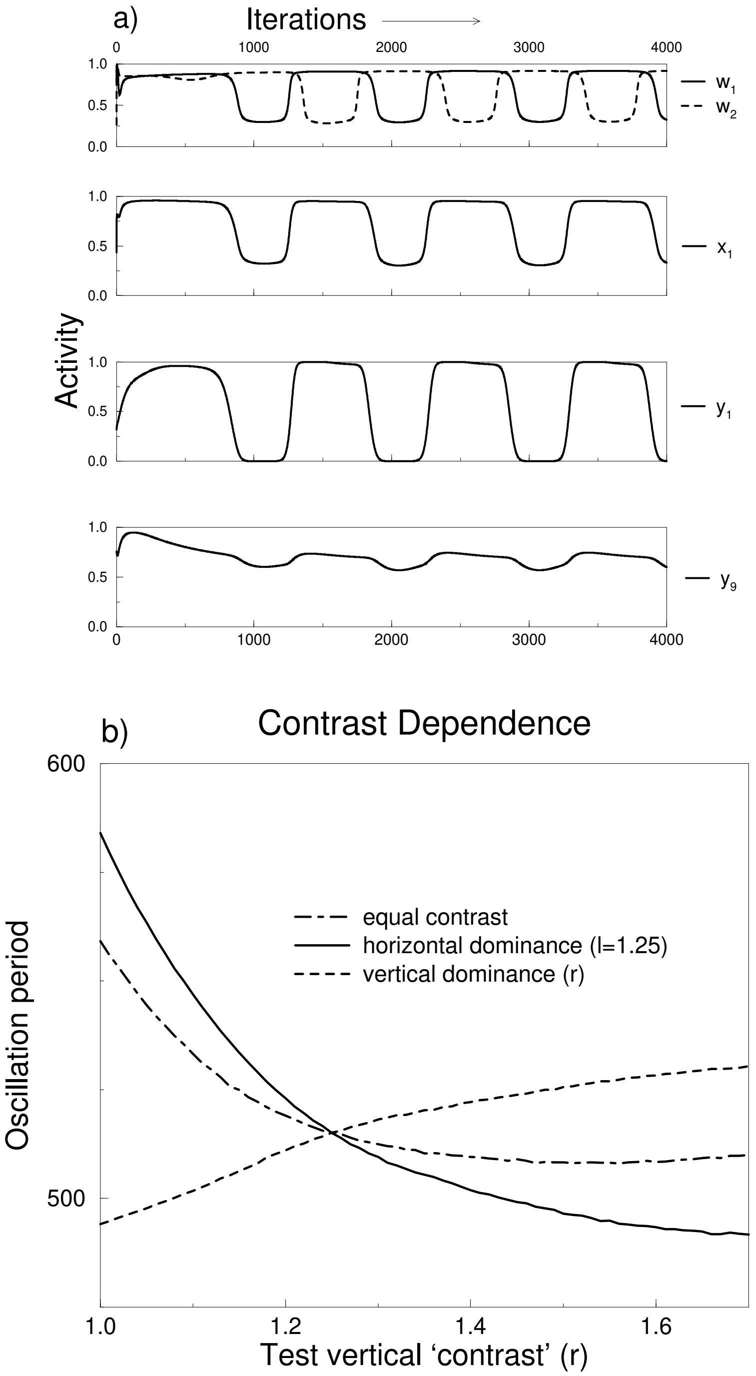

Figure 5: a) Mean activities of units at three levels of the hierarchy in response to rivalrous stimuli with input strengths l=r=1.75. b) Contrast dependence of the oscillation periods. The dash-dotted line shows the period when the contrasts in both eyes are varied together. The solid and dashed lines show the periods of dominance of the left and right eyes respectively when the contrast in the left eye is fixed (l=1.25) and the contrast in the right eye r is varied.

of couching rivalry in terms of explanations is that this provides a natural way of accounting for top-down influences, which are clear in such phenomena as the influence of perceptual organisation on rivalry (*eg* Whittle *et al*, 1968; Kovacs *et al*, 1996). In fact, one can hope to study top-down control through studying its effects on the behaviour of cells during rivalry. The model would also explain other sorts of alternation phenomena (such as those that arise with the Necker cubs) in terms of competition between top-down explanations. The top-down model governs which units should compete with each other.

The model correctly captures a number of the experimental characteristics of rivalry. If the input stimuli are weak, then there is no alternation, and instead both horizontal and vertical representing neurons are weakly activated (as in Liu, Tyler & Schor, 1992). If input stimuli are stronger, then alternation ensues. The period of the alternation increases as the contrast of both the stimuli decrease, and if the contrast of only one stimulus decreases, then the dominance period of the other stimulus increases substantially more than the suppression period of the given stimulus (as in Levelt, 1965; Fox & Rasche, 1969; Blake, 1977; Mueller & Blake, 1989). There are two classes of binocular units activated by the rivalrous stimulus. The activity of one class is substantially modulated during rivalry; the activity of the other is not (as in Leopold & Logothetis, 1996). Alternating the input between the two eyes has absolutely no effect on this behaviour of the model (as in Logothetis, Leopold & Sheinberg, 1996). The last effect arises since, apart from the input layer, on which there are no top-down influences, all the units are binocular, and there is no static or dynamic difference in the connections from the two eyes.

Although it captures these phenomena, the model is, of course, simplified and incomplete. In particular, it does not exhibit two of the phenomena that Leopold & Logothetis (1996) observed. The first is that there is no opportunity in the model for monocular cells to be unmodulated during rivalry, as they found. Given redundant inputs and an extra layer of monocular units between layers $\mathbf{z}$ and $\mathbf{y}$, this behavior would be expected. These units would explain away the redundancy in the input, and, like unit y_9 in figure 5, would have consistently to be activated during rivalry.

The second lacuna is that there are no units in the model that are selective between the stimuli when presented binocularly and are preferentially activated during *suppression* of their preferred stimuli during rivalry, or are not selective during binocular presentation but are selective during rivalry. In a model with more complicated stimulus contingencies, such units would emerge to account for the parts of the stimulus in the suppressed eye that are *not* accounted for by the explanation of the overlying parts of the dominant explanation, at least provided that this residual between the true monocular stimulus and the current explanation is sufficiently complex as to require explaining itself. This suggests the experimental test of presenting binocularly a putative form of the residual (*eg* dotted lines for competing horizontal and vertical gratings). We predict that these cells should be activated. One might expect some of these cells to participate in the explanation of the patterns when presented binocularly, whereas the activity of others would be explained away during binocular presentation, only to emerge during suppression.

Other extensions are also desirable. Foremost, it is necessary to model the stochasticity of switching between explanations (Fox & Herrmann, 1967; Levelt, 1965). The distributions of dominance times for both humans and monkeys have traditionally been characterised in terms of a Gamma distribution, and, more recently, in terms of a log normal distribution (Lehky, 1995), with independence between successive dominance periods. Our mean field recognition process is deterministic. The stochastic analogue would be some form of Markov chain Monte-Carlo method such as Gibbs sampling (see Neal, 1993). However, it is not obvious how to incorporate the equivalent of fatigue in a computationally reasonable way. In any case, the nature of neuronal randomness is subject to significant debate at present.

We have adopted a very simple mean field approach to recognition, giving up neurobiological plausibility for convenience. The determinism of the mean field model in any case rules it out as a complete explanation, but it does at least show clearly the nature of competition between explanations. The architecture of the model is also incomplete. The cortex is replete with what we would model as lateral connections between units within a single layer. We have constructed generative models in which there are no such direct connections, because they significantly complicate the mean field recognition method. These connections are certainly important for the recognition process (Dayan & Hinton, 1996), but modeling their effect would require representing them explicitly. This would also allow modeling of the apparent diffusive process by which patches of dominance spread and alter. In a complete model, it would also be necessary to account for competition between eyes in addition to competition between explanations (Wales & Fox, 1970; Fox & Check, 1972; Blake & Fox, 1974).

Another extension is some form of contrast gain control (Carandini & Heeger, 1994). The model is quite sensitive to input contrast, which is obviously important for the effects shown in figures 2 and 5. However, the range of contrasts over which it works should be larger. Achieving this will likely require a statistical model with real-valued rather than binary-valued activities. It would be particularly revealing to explore the effects of changing the contrast in some parts of images and examine the consequent effects on the spreading of dominance.

Acknowledgements

I am very grateful to Bart Anderson, Adam Elga, Geoff Goodhill, Geoff Hinton, David Leopold, Nikos Logothetis, Earl Miller, Read Montague, Bruno Olshausen, Pawan Sinha, Rich Zemel, and particularly Zhaoping Li, Tommi Jaakkola, and two anonymous referees for discussion and comments on earlier drafts. This work was supported by NIMH grant 1 R29 MH 55541-01.

References

Bialek, W & DeWeese, M (1995). Random switching and optimal processing in the perception of ambiguous signals. *Physical Review Letters*, **74**, 3077-3080.

Blake, R (1977). Threshold conditions for binocular rivalry. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 251-257.

Blake, R (1989). A neural theory of binocular rivalry. *Psychological Review*, **96**, 145-167.

Blake, R & Fox, R (1974). Binocular rivalry suppression: Insensitive to spatial frequency and orientation change. *Vision Research*, **14**, 687-692.

Blake, R, Westendorf, DH & Overton, R (1980). What is suppressed during binocular rivalry? *Perception*, **9**, 223-231.

Carandini, M & Heeger, DJ (1994). Summation and division by neurons in primate visual cortex. *Science*, **264**, 1333-1336.

Dayan, P & Hinton, GE (1996). Varieties of Helmholtz machine. *Neural Networks*, **9**, 1385-1403.

Dayan, P, Hinton, GE, Neal, RM & Zemel, RS (1995). The Helmholtz machine. *Neural Computation*, **7**, 889-904.

Felleman DJ & Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, **1**, 1-47.

Fox, R & Check, R (1972). Independence between binocular rivalry suppression duration and magnitude of suppression. *Journal of Experimental Psychology*, **93**, 283-289.

Fox, R & Herrmann, J (1967). Stochastic properties of binocular rivalry alternations. *Perception and Psychophysics*, **2**, 432-436.

Fox, R & Rasche, F (1969). Binocular rivalry and reciprocal inhibition. *Perception and Psychophysics*, **5**, 215-217.

Grossberg, S (1987). Cortical dynamics of three-dimensional form, color and brightness perception: 2. Binocular theory. *Perception & Psychphysics*, **41**, 117-158.

Hinton, GE, Dayan, P, Frey, BJ & Neal, RM (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, **268**, 1158-1160.

Jaakkola, T, Saul, LK & Jordan, MI (1996). Fast learning by bounding likelihoods in sigmoid type belief networks. *Advances in Neural Information Processing Systems 8*, MIT Press.

Kovacs, I, Papathomas, TV, Yang, M & Feher A (1996). When the brain changes its mind: interocular grouping during binocular rivalry. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 15508-15511.

Lehky, SR (1988). An astable multivibrator model of binocular rivalry. *Perception*, **17**, 215-228.

Lehky, SR (1995). Binocular rivalry is not chaotic. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **259**, 71-76.

Lehky, SR & Blake, R (1991). Organization of binocular pathways: Modeling and data related to rivalry. *Neural Computation,* **3**, 44-53.

Lehky, SR & Maunsell, JH (1996). No binocular rivalry in the LGN of alert macaque monkeys. *Vision Research,* **36**, 1225-1234.

Leopold, DA & Logothetis, NK (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature,* **379**, 549-554.

Levelt, WJM (1965). *On Binocular Rivalry.* Soesterberg, The Netherlands: Institute for Perception RVO-TNO.

Liu, L, Tyler, CW & Schor, C (1992). Failure of rivalry at low contrast: Evidence of a suprathreshold binocular summation. *Vision Research,* **32**, 1471-1479.

Logothetis, NK, Leopold, DA & Sheinberg, DL (1996). What is rivalling during binocular rivalry. *Nature,* **380**, 621-624.

Logothetis, NK & Schall, JD (1989). Neuronal correlates of subjective visual perception. *Science,*, **245**, 761-763.

Matsuoka, K (1984). The dynamic model of binocular rivalry. *Biological Cybernetics,* **49**, 201-208.

Mueller, TJ (1990). A physiological model of binocular rivalry. *Visual Neuroscience,* **4**, 63-73.

Mueller, TJ & Blake, R (1989). A fresh look at the temporal dynamics of binocular rivalry. *Biological Cybernetics,* **61**, 223-232.

Neal, RM (1993). *Probabilistic Inference using Markov Chain Monte Carlo Methods.* Technical Report CRG-TR-93-1. Department of Computer Science, University of Toronto.

Olshausen, BA & Field, DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature,* **381**, 607-609.

Pearl, J (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann.

Rao, PNR & Ballard, DH (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation,* **9**, 721-764.

Saul, LK, Jaakkola, T & Jordan, MI (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research,* **4**, 61-76.

Sengpiel, F, Blakemore, C & Harrad, R (1995). Interocular suppression in the primary visual cortex: a possible neural basis of binocular rivalry. *Vision Research,* **35**, 179-195.

Varela, FJ & Singer, W (1987). Neuronal dynamics in the visual corticothalamic pathway revealed through binocular rivalry. *Experimental Brain Research,* **66**, 10-20.

Wales, R & Fox, R (1970). Increment detection thresholds during binocular rivalry suppression. *Perception and Psychophysics,* **8**, 90-94.

Wheatstone, C (1838). Contributions to the theory of vision. I: On some remarkable and hitherto unobserved phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London,* **128**, 371-394.

Whittle, P, Bloor, DC & Pocock, (1968). Some experiments on figural effects in binocular rivalry. *Perception & Psychophysics,* **4**, 183-188.

Yu, K & Blake, R (1992). Do recognizable figures enjoy an advantage in binocular rivalry? *Journal of Experimental Psychology: Humna Perception and Performance,* **18**, 1158-1173