

A Neural Network Model of Visual Tilt Aftereffects

James A. Bednar (jbednar@cs.utexas.edu)

Risto Miikkulainen (risto@cs.utexas.edu)

Department of Computer Sciences

University of Texas at Austin

Austin, TX 78712 USA

Abstract

RF-LISSOM, a self-organizing model of laterally connected orientation maps in the primary visual cortex, was used to study the psychological phenomenon known as the tilt aftereffect. The same self-organizing processes that are responsible for the long-term development of the map and its lateral connections are shown to result in tilt aftereffects over short time scales in the adult. The model allows observing large numbers of neurons and connections simultaneously, making it possible to relate higher-level phenomena to low-level events, which is difficult to do experimentally. The results give computational support for the idea that direct tilt aftereffects arise from adaptive lateral interactions between feature detectors, as has long been surmised. They also suggest that indirect effects could result from the conservation of synaptic resources during this process. The model thus provides a unified computational explanation of self-organization and both direct and indirect tilt aftereffects in the primary visual cortex.

1 Introduction

The tilt aftereffect (TAE, gibson:adaptation) is a simple but intriguing visual phenomenon. After staring at a pattern of tilted lines or gratings, subsequent lines appear to have a slight tilt in the opposite direction (Figure ??). The effect resembles an afterimage from staring at a bright light, but it reflects changes in orientation perception rather than in color or brightness.

Most modern explanations of the TAE are based on the *feature-detector* model of the visual cortex ?. Individual orientation detectors become more difficult to excite during repeated presentation of oriented stimuli, and the desensitization persists for some time afterwards. This observation forms the basis of the *fatigue* theory of the TAE: if active neurons become fatigued over time, the set of neurons activated for a test figure will shift away from the adaptation orientation. Assuming the perceived orientation is some sort of average over the orientation preferences of the activated neurons, the perceived orientation would thus show the direct TAE ?.

The fatigue theory has been discredited because it has become apparent that the adaptation is mediated by the lateral connections between neurons, rather than changes occurring within the neurons themselves ?. The now-popular *inhibition* theory postulates that tilt aftereffects result from changing inhibition between neurons ?, perhaps by increases in the strength of lateral connections between them.



Figure 1: **Tilt aftereffect patterns.** Fixate your gaze upon the circle inside the square at the center for at least thirty seconds, moving your eye slightly inside the circle to avoid developing strong afterimages. Now fixate upon the figure at the left. The vertical lines should appear slightly tilted to the right; this phenomenon is called the direct tilt aftereffect. If you fixate upon the horizontal lines at the right, they should appear barely tilted counterclockwise, demonstrating the indirect tilt aftereffect. (Adapted from campbell:vres71.)

Although the inhibition theory was first proposed in the 1970s, only recently has it become computationally feasible to test in a detailed model of cortical function. A Hebbian self-organizing process (the Receptive-Field Laterally Interconnected Syn-ergetically Self-Organizing Map, or RF-LISSOM; *miikkulainen:psylm97,sirosh:phd,sirosh:bc94,sirosh:npl96,sirosh:neuralc article) has been shown to develop feature detectors and specific lateral connections that could produce such after-effects. The RF-LISSOM model gives rise to anatomical and functional characteristics of the cortex such as topographic maps, ocular dominance, orientation, and size preference columns, and the patterned lateral connections between them. Although other models exist that explain

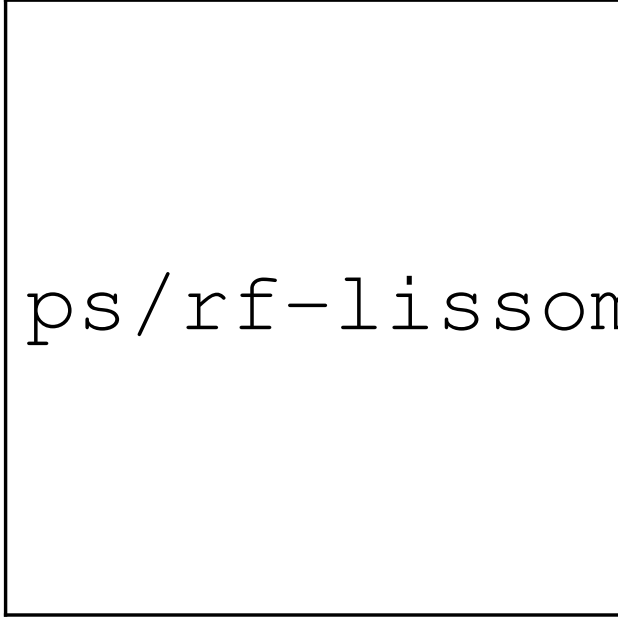


Figure 2: **Architecture of the RF-LISSOM network.** A tiny RF-LISSOM network and retina are shown, along with connections to a single neuron (shown as a large circle). The input is an oriented Gaussian activity pattern on the retinal ganglion cells. The afferent connections form a local anatomical receptive field on the simulated retina. Neighboring neurons have different but highly overlapping RFs. Each neuron computes an initial response as a dot product of its receptive field and its afferent weight vector. The responses then repeatedly propagate within the cortex through the lateral connections and evolve into an activity “bubble”. After the activity stabilizes, weights of the active neurons are adapted.

how the feature-detectors and afferent connections could develop by input-driven self-organization, RF-LISSOM is the only model that also shows how the lateral connections can self-organize as an integral part of the process. The laterally connected model has also been shown to account for many of the dynamic aspects of the visual cortex, such as reorganization following retinal and cortical lesions (miikkulainen:psylm97,sirosh:phd,sirosh:cns94;*sirosh:htmlbook96-article).

The current work is a first study of the *functional* behavior of the model, specifically the response to stimuli similar to those known to cause the TAE in humans. The RF-LISSOM model allows observing activation and connection patterns between large numbers of neurons simultaneously, making it possible to relate higher-level phenomena to low-level events, which is difficult to do experimentally. The results suggest that tilt aftereffects are not flaws in an otherwise well-designed system, but an unavoidable result of a self-organizing process that aims at producing an efficient, sparse encoding of the input through decorrelation (as proposed by barlow:aftereffects; see also dong:decorrelation,field:goal,foldiak:bc90,miikkulainen:psylm97;*sirosh:htmlbook96-article).

2 Architecture

The cortical architecture for the model has been simplified and reduced to the minimum necessary configuration to account for the observed phenomena. Because the focus is on the two-dimensional organization of the cortex, each “neuron” in the model cortex corresponds to a vertical column of cells through the six layers of the human cortex. The cortical network is modeled with a sheet of interconnected neurons and the retina with a sheet of retinal ganglion cells (figure ??). Neurons receive afferent connections from broad overlapping patches on the retina. The $N \times N$ network is projected on to the retina of $R \times R$ ganglion cells, and each neuron is connected to ganglion cells in a circular area of radius r around the projections. Thus, neurons at a particular cortical location receive afferents from the corresponding location on the retina. Since the LGN accurately reproduces the receptive fields of the retina, it has been bypassed for simplicity.

Each neuron also has reciprocal excitatory and inhibitory lateral connections with itself and other neurons. Lateral excitatory connections are short-range, connecting each neuron with itself and its close neighbors. Lateral inhibitory connections run for comparatively long distances, but also include connections to the neuron itself and to its neighbors.

The input to the model consists of 2-D ellipsoidal Gaussian patterns representing retinal ganglion cell activations. For training, the orientations of the Gaussians are chosen randomly from the uniform distribution in the range $[0, \pi)$. The elongated spots approximate natural visual stimuli after the edge detection and enhancement mechanisms in the retina. They can also be seen as a model of the intrinsic retinal activity waves that occur in late pre-natal development in mammals ?. The RF-LISSOM network models the self-organization of the visual cortex based on these natural sources of elongated features.

The afferent weights are initially set to random values, and the lateral weights are preset to a smooth Gaussian profile. The connections are organized through an unsupervised learning process. At each training step, neurons start out with zero activity. The initial response η_{ij} of neuron (i, j) is calculated as a weighted sum of the retinal activations:

$$\eta_{ij} = \sigma \left(\sum_{a,b} \xi_{ab} \mu_{ij,ab} \right), \quad (1)$$

where ξ_{ab} is the activation of retinal ganglion (a, b) within the anatomical RF of the neuron, $\mu_{ij,ab}$ is the corresponding afferent weight, and σ is a piecewise linear approximation of the sigmoid activation function. The response evolves over a very short time scale through lateral interaction. At each time step, the neuron combines the above afferent activation $\sum \xi \mu$ with lateral excitation and inhibition:

$$\eta_{ij}(t) = \sigma \left(\sum \xi \mu + \gamma_e \sum_{k,l} E_{ij,kl} \eta_{kl}(t-1) - \right.$$

$$\gamma_i \sum_{k,l} I_{ij,kl} \eta_{kl}(t-1) \Big), \quad (2)$$

$$\eta_{ij}(t) = \sigma \left(\sum \xi \mu + \gamma_e \sum_{k,l} E_{ij,kl} \eta_{kl}(t-1) - \gamma_i \sum_{k,l} I_{ij,kl} \eta_{kl}(t-1) \right), \quad (3)$$

where $E_{ij,kl}$ is the excitatory lateral connection weight on the connection from neuron (k, l) to neuron (i, j) , $I_{ij,kl}$ is the inhibitory connection weight, and $\eta_{kl}(t-1)$ is the activity of neuron (k, l) during the previous time step. The scaling factors γ_e and γ_i determine the relative strengths of excitatory and inhibitory lateral interactions.

While the cortical response is settling, the retinal activity remains constant. The activity pattern starts out diffuse and spread over a substantial part of the map, but within a few iterations of equation ??, converges into a small number of stable focused patches of activity, or activity bubbles. After the activity has settled, the connection weights of each neuron are modified. Both afferent and lateral weights adapt according to the same mechanism: the Hebb rule, normalized so that the sum of the weights is constant:

$$w_{ij,mn}(t + \delta t) = \frac{w_{ij,mn}(t) + \alpha \eta_{ij} X_{mn}}{\sum_{mn} [w_{ij,mn}(t) + \alpha \eta_{ij} X_{mn}]}, \quad (4)$$

where η_{ij} stands for the activity of neuron (i, j) in the final activity bubble, $w_{ij,mn}$ is the afferent or lateral connection weight (μ , E or I), α is the learning rate for each type of connection (α_A for afferent weights, α_E for excitatory, and α_I for inhibitory) and X_{mn} is the presynaptic activity (ξ for afferent, η for lateral). The larger the product of the pre- and post-synaptic activity $\eta_{ij} X_{mn}$, the larger the weight change. Therefore, when the pre- and post-synaptic neurons fire together frequently, the connection becomes stronger. Both excitatory and inhibitory connections strengthen by correlated activity; normalization then redistributes the changes so that the sum of each weight type for each neuron remains constant.

At long distances, very few neurons have correlated activity and therefore most long-range connections eventually become weak. The weak connections can be eliminated periodically, resulting in patchy lateral connectivity similar to that observed in the visual cortex. The radius of the lateral excitatory interactions starts out large, but as self-organization progresses, it is decreased until it covers only the nearest neighbors. Such a decrease is necessary for global topographic order to develop and for the receptive fields to become well-tuned at the same time.

3 Experiments

The model consisted of an array of 192×192 neurons, and a retina of 24×24 ganglion cells. The circular anatomical receptive field of each neuron was centered in the portion of

the retina corresponding to the location of the neuron in the cortex. The RF consisted of random-strength connections to all ganglion cells less than 6 units away from the RF center. The cortex was self-organized for 30,000 iterations on oriented Gaussian inputs with major and minor axes of half-width $\sigma = 7.5$ and 1.5, respectively.¹ The training took 8 hours on 64 processors of a Cray T3D at the Pittsburgh Supercomputing Center. The model requires more than three gigabytes of physical memory to represent the more than 400 million connections in this small section of the cortex.

3.1 Orientation map organization

In the self-organization process, the neurons developed oriented receptive fields organized into orientation columns very similar to those observed in the primary visual cortex. The strongest lateral connections of highly-tuned cells link areas of similar orientation preference, and avoid neurons with the orthogonal orientation preference. Furthermore, the connection patterns of highly oriented neurons are typically elongated along the direction in the map that corresponds to the neuron's preferred stimulus orientation. This organization reflects the activity correlations caused by the elongated Gaussian input pattern: such a stimulus activates primarily those neurons that are tuned to the same orientation as the stimulus, and located along its length. Since the long-range lateral connections are inhibitory, the net result is *decorrelation*: redundant activation is removed, resulting in a sparse representation of the novel features of each input (barlow:aftereffects,field:goal; *sirosh:htmlbook96-article). As a side effect, illusions and aftereffects may sometimes occur, as will be shown below.

3.2 Aftereffect simulations

In psychophysical measurements of the TAE, a fixed stimulus is presented at a particular location on the retina. To simulate these conditions in the model, the position and angle of the inputs were fixed to a single value for a number of iterations, rather than having a uniform random distribution as in self-organization. To permit more detailed analysis of behavior at short time scales, the learning rates were reduced from those used during self-organization, to $\alpha_A = \alpha_E = \alpha_I = 0.00005$. All other parameters remained as in self-organization.

To compare with the psychophysical experiments, perceived orientations were compared before and after tilt adap-

¹ The initial lateral excitation radius was 19 and was gradually decreased to 1. The lateral inhibitory radius of each neuron was 47, and inhibitory connections whose strength was below 0.00025 were pruned away at 30,000 iterations. The lateral inhibitory connections were preset to a Gaussian profile with $\sigma = 100$, and the lateral excitatory connections to a Gaussian with $\sigma = 15$. The lateral excitation γ_e and inhibition strength γ_i were both 0.9. The learning rate α_A was gradually decreased from 0.007 to 0.0015, α_E from 0.002 to 0.001 and α_I was a constant 0.00025. The lower and upper thresholds of the sigmoid were increased from 0.1 to 0.24 and from 0.65 to 0.88, respectively. The number of iterations for which the lateral connections were allowed to settle at each training iteration was initially 9, and was increased to 13 over the course of training. The parameter settings were identical to those of sirosh:phd, and were not tuned or tweaked for the tilt aftereffect simulations. Small variations produce roughly equivalent results.

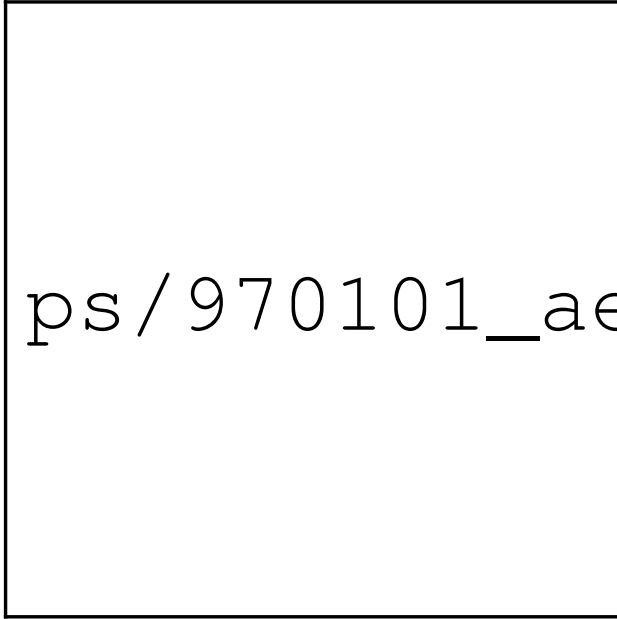


Figure 3: **Tilt aftereffect versus retinal angle.** The open circles represent the average tilt aftereffect for a single human subject (DEM) from mitchell:vres76 over ten trials. For each angle in each trial, the subject adapted for three minutes on a sinusoidal grating of a given angle, then was tested for the effect on a horizontal grating. Error bars indicate ± 1 standard error of measurement. The subject shown had the most complete data of the four in the study. All four showed very similar effects in the x-axis range $\pm 40^\circ$; the indirect TAE for the larger angles varied widely between $\pm 2.5^\circ$. The graph is roughly anti-symmetric around 0° , so the TAE is essentially the same in both directions relative to the adaptation line. The heavy line shows the average magnitude of the tilt aftereffect in the RF-LISSOM model over nine trials at different locations on the retina. Error bars indicate ± 1 standard error of measurement. The network adapted to a vertical adaptation line at a particular position for 90 iterations, then the TAE was measured for test lines oriented at each angle. The duration of adaptation was chosen so that the magnitude of the human data and the model match; this was the only parameter fit to the data. The result from the model closely resembles the curve for humans at all angles, showing both direct and indirect tilt aftereffects.

tation. Perceived orientation was measured as a vector sum over all active neurons, with the magnitude of each vector representing the activation level, and the vector direction representing the orientation preference of the neuron before adaptation. Perceived orientation was computed separately for each possible orientation of the test Gaussian, both before and after adaptation. For a given angular separation of the adaptation stimulus and the test stimulus, the computed magnitude of the tilt aftereffect is the difference between the initial perceived angle and the one perceived after adaptation. Figure ?? plots these differences after adaptation for 90 iterations of the RF-LISSOM algorithm. For comparison, figure ?? also shows the most detailed data available for the TAE in human foveal vision ?.

The results from the RF-LISSOM simulation are strikingly similar to the psychophysical results. For the range 5° to 40° ,

all subjects in the human study (including the one shown) exhibited angle repulsion effects nearly identical to those found in the RF-LISSOM model. The magnitude of this *direct* TAE increases very rapidly to a maximum angle repulsion at approximately 10° , falling off somewhat more gradually to zero as the angular separation increases.

The results for larger angular separations (from 45° to 85°) show a greater inter-subject variability in the psychophysical literature, but those found for the RF-LISSOM model are well within the range seen for human subjects. The *indirect* effects for the subject shown were typical for that study, although some subjects showed effects up to 2.5° .

In addition to the angular changes in the TAE, its magnitude in humans increases regularly with adaptation time ?. The equivalent of “time” in the RF-LISSOM model is an iteration, i.e. a single cycle of input presentation, activity propagation, settling, and weight modification. As the number of adaptation iterations is increased, the magnitude of the TAE in the model increases monotonically, while retaining the same basic shape of figure ?? ?. The curve that best matches the human data was shown in figure ??.

Due to the time required to obtain even a single point on the angular curve of the TAE for human subjects, complete experimental measurements of the angular function at different adaptation times are not available. However, when the time course of the direct TAE is measured at a single orientation, the increase is approximately logarithmic with time ?, eventually saturating at a level that depends upon the experimental protocol used ?. Figure ?? compares the shape of this TAE versus time curve for human subjects and for the RF-LISSOM model. The x axis for the RF-LISSOM and human data has different units, but the correspondence between the two curves might provide a rough way of quantifying the equivalent real time for an “iteration” of the model. The time course of the TAE in the RF-LISSOM model is similar to the human data. The TAE increases approximately logarithmically, but it does not completely saturate over the adaptation amounts tested so far. This difference suggests that the biological implementation has additional constraints on the amount of learning that can be achieved over the time scale over which the tilt aftereffect is seen.

3.3 How does the TAE arise in the model?

The TAE seen in figures ?? and ?? must result from changes in the connection strengths between neurons, since no other component of the model changes as adaptation progresses. Simulations performed with only one type of weight (either afferent, lateral excitatory, or lateral inhibitory) adapting at a given time show that the inhibitory weights determine the shape of the curve for all angles ?. The small component of the TAE resulting from adaptation of either type of excitatory weights is almost precisely opposite the total effect. Although each inhibitory connection adapts with the same learning rate as the excitatory connections ($\alpha_I = \alpha_A = \alpha_E = 0.00005$), there are many more inhibitory connections than excitatory connections. The combined strength of all the small inhibitory changes outweighs the excitatory changes, and results in a curve with a sign opposite that of the components from the excitatory weights.

$$(0.6.37)[r][rt]$$

Figure 4: **Direct tilt aftereffect versus time.** The circles show the magnitude of the TAE as a function of adaptation time for human subjects MWG (unfilled circles) and SM (filled circles) from greenlee:vres87sat; they were the only subjects tested in the study. Each subject adapted to a single $+12^\circ$ line for the time period indicated on the horizontal axis (bottom). To estimate the magnitude of the aftereffect at each point, a vertical test line was presented at the same location and the subject was requested to set a comparison line at another location to match it. The plots represent averages of five runs; the data for 0 – 10 minutes were collected separately from the rest. For comparison, the heavy line shows average TAE in the LISSOM model for a $+12^\circ$ test line over 9 trials (with parameters as in figure ??). The horizontal axis (top) represents the number of iterations of adaptation, and the vertical axis represents the magnitude of the TAE at this time step. The RF-LISSOM results show a similar logarithmic increase in TAE magnitude with time, but do not show the saturation that is seen for the human subjects.

In what way do the changing inhibitory connections cause these effects? During adaptation, we see that the response to the 0° adaptation line becomes gradually more concentrated towards the central area of the Gaussian pattern presented. This is because the inhibition between active neurons increases, allowing only the most strongly activated neurons to remain active after settling (equation ??). However, the distribution of active orientation detectors is centered around the same angle, so the same angle is perceived.

The response to a test line with a slightly different orientation (e.g. 10°) is also more focused after adaptation, but the overall distribution of activated neurons has shifted. Fewer neurons that prefer orientations close to the adaptation line now respond, but an increased number of those that prefer distant angles do. This is because inhibition was strengthened primarily between neurons close to the adaptation angle, and not between those which prefer larger orientations, greater than the 10° test line. The net effect is a shift of the perceived orientation *away* from the adaptation angle, resulting in the direct TAE.

In contrast, the response to a very different test line (e.g. 60°) is broader and stronger after adaptation. Adaptation occurred only in activated neurons, so neurons with orientation preferences greater than 60° are unchanged. However, those with preferences somewhat less than 60° actually now respond more strongly. During adaptation, their inhibitory connections with other active neurons, i.e. those that represent orientations close to the 0° adaptation line, became stronger. Since the sum of inhibition is constant for each neuron (equation ??), the connections to neurons representing distant angles (e.g. 60°) became weaker. As a result, the 60° line now inhibits them less than before adaptation. Thus they are more active, and the perceived orientation has shifted towards 0° . This indirect effect is therefore true to its name, caused indirectly by the strengthening of inhibitory connections. The RF-LISSOM model thus shows computationally that both the direct and indirect effects could be caused by activity-dependent adaptation of inhibitory lateral interactions.

4 Discussion and Future Work

The results presented above suggest that the same self-organizing principles that result in sparse coding and reduce redundant activation may also be operating over short time intervals in the adult, with quantifiable psychological consequences such as the TAE. This finding demonstrates a potentially important computational link between development, structure, and function.

Even though the RF-LISSOM model was not originally developed as an explanation for the tilt aftereffect, it exhibits tilt

aftereffects that have nearly all of the features of those measured in humans. The effect of varying angular separation between the test and adaptation lines is similar to human data at all orientations, the time course is approximately logarithmic in each, and the TAE is localized to the retinal location which experienced the stimulus. With minor extensions, the model should account for other features of the TAE, such as higher variance at oblique orientations, frequency localization, movement direction specificity, and ocular transfer. For a discussion of the match between the model and data for humans from a variety of experiments, see bednar:aitr97.

The only prominent features of the TAE that do not directly follow from the model are saturation of the effect for long adaptations, and recovery of accurate perception even in complete darkness ???. These two features suggest that the inhibitory weights modified during tilt adaptation could actually be a set of small, temporary weights adding to or multiplying more permanent connections. Such a mechanism was proposed by vnderdalsburg:synaptic as an explanation of visual object segmentation; this idea was implemented for the RF-LISSOM model by choe:utctr96 and miikkulainen:psylm97. The TAE may be merely a minor consequence of this multi-level architecture for representing correlations over a wide range of time scales.

A main contribution of the RF-LISSOM model of the TAE is its novel explanation of the indirect effect. Proponents of the lateral inhibitory theory of direct effects have generally ignored indirect effects, or postulated that they occur only at higher cortical levels ?, partly because it has not been clear how they could arise through inhibition in V1. RF-LISSOM demonstrate that a quite simple, local mechanism in V1 is sufficient to produce indirect effects. If the total synaptic resources at each neuron are limited, strengthening the lateral inhibitory connections between active neurons weakens their inactive inhibitory connections. There is widespread biological evidence of competition for a limited number of synaptic sites ??????. There is also extensive computational justification for synaptic resource conservation, beginning with one of the first computational models of Hebbian adaptation ?. Without such normalization, connection weights governed by a Hebbian rule will increase indefinitely, or else each would reach a maximum strength ?. Neither outcome would appear biologically or computationally plausible, so the assumption of some form of normalization is well-motivated ?.

Through mechanisms similar to those causing the TAE, the RF-LISSOM model should also be able to explain simultaneous tilt illusions between spatially separated stimuli. Such an explanation was originally proposed by carpenter:interactions. However, it will be necessary to train the sys-

tem with inputs that have longer-range correlations between similar orientations, such as sinusoidal gratings (representing objects with parallel lines). With such patterns, long-range connections develop between widely separated orientation detectors, in addition to the relatively local connections now present. Trained with such patterns, RF-LISSOM should be able to account for tilt illusions as well as tilt aftereffects. Although such experiments require even larger cortex and retina sizes, they should become practical in the near future.

In addition, many similar phenomena such as aftereffects of curvature, motion, spatial frequency, size, position, and color have been documented in humans ?. Since specific detectors for most of these features have been found in the cortex, RF-LISSOM should be able to account for them by the same process of decorrelation mediated by self-organizing lateral connections.

5 Conclusion

The experiments reported in this paper lend strong computational support to the theory that tilt aftereffects result from Hebbian adaptation of the lateral connections between neurons. Furthermore, the aftereffects occur as a result of the same decorrelating process that is responsible for the initial development of the orientation map. This process tends to deemphasize constant features of the input, resulting in short-term perceptual anomalies such as aftereffects. The same model should also apply to other aftereffects and to simultaneous tilt illusions.

Because RF-LISSOM is a computational model, it can demonstrate many phenomena in high detail that are difficult to measure experimentally, thus presenting a view of the cortex that is otherwise not available. This type of analysis can provide an essential complement to experimental work with humans and animals. RF-LISSOM provides a comprehensive and fundamental account of how both cortical structure and function emerge by Hebbian self-organization in the primary visual cortex. It also shows how both indirect and direct tilt aftereffects could arise from simple, biologically plausible mechanisms in the primary visual cortex. Thus a single simple computational model may lead to significant insights into a variety of cortical phenomena, and thereby contribute to our understanding of the cortex.

A Acknowledgments

Thanks to Joseph Sirosh for supplying the RF-LISSOM code. This research was supported in part by the National Science Foundation under grant #IRI-9309273. Computer time for the simulations was provided by the Pittsburgh Supercomputing Center under grant IRI940004P.