LETTER Communicated by Gabriel Kreiman

A Gaussian Attractor Network for Memory and Recognition with Experience-Dependent Learning

Xiaolin Hu
xlhu@tsinghua.edu.cn
Bo Zhang
dcszb@tsinghua.edu.cn
State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, and Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Attractor networks are widely believed to underlie the memory systems of animals across different species. Existing models have succeeded in qualitatively modeling properties of attractor dynamics, but their computational abilities often suffer from poor representations for realistic complex patterns, spurious attractors, low storage capacity, and difficulty in identifying attractive fields of attractors. We propose a simple two-layer architecture, gaussian attractor network, which has no spurious attractors if patterns to be stored are uncorrelated and can store as many patterns as the number of neurons in the output layer. Meanwhile the attractive fields can be precisely quantified and manipulated. Equipped with experience-dependent unsupervised learning strategies, the network can exhibit both discrete and continuous attractor dynamics. A testable prediction based on numerical simulations is that there exist neurons in the brain that can discriminate two similar stimuli at first but cannot after extensive exposure to physically intermediate stimuli. Inspired by this network, we found that adding some local feedbacks to a well-known hierarchical visual recognition model, HMAX, can enable the model to reproduce some recent experimental results related to high-level visual perception.

1 Introduction

Attractor network theories suggest that memories are represented as steady states in recurrent neural networks (Hopfield, 1982; Amit, 1989; Fuster, 1995; Papp, Witter, & Treves, 2007), which is believed to underlie persistent neural activity observed in a wide variety of regions of the nerve system—for example, the prefrontal cortex (PFC) of monkeys (Fuster, 1995; Goldmanrakic, 1996), the anterior dorsal nucleus of the rodent thalamus (Sharp, Blair, & Cho, 2001), and the prepositus hypoglossi and the medial vestibular nucleus of mammals (Seung, Lee, Reis, & Tank, 2000). Emerging evidence

 © 2010 Massachusetts Institute of Technology

also demonstrated that dynamics can evolve from unstable states to steady states. For instance, single-cell recordings (Wills, Lever, Cacucci, Burgess, & O'Keefe, 2005) showed that when rats were exploring boxes of intermediate shapes varying between a square and a circle after they were familiarized with the square and circular boxes, the place cell activities in the CA1 region of the hippocampus were either similar to those when they were in the square box or similar to those when they were in the circular box. Though a different finding was reported in Leutgeb, Leutgeb, Treves et al. (2005), theoretical models indicated that it complemented rather than challenged attractor theories (Blumenfeld, Preminger, Sagi, & Tsodyks, 2006; Papp et al., 2007).

Similar categorization phenomena to the rats study (Wills et al., 2005) have been observed in human inferior temporal cortex (ITC) (Rotshtein, Henson, Treves, Driver, & Dolan, 2005), the highest purely visual area in the ventral visual stream. Subjects were exposed to sequences of facial images with incremental changes between pairs of famous faces. A behavioral experiment was performed to determine the psychological boundary of the two face categories on each sequence. Subsequent fMRI adaptation experiments showed that the right fusiform gyrus (FFG), including the independently defined right fusiform face area (FFA), could differentiate faces belonging to different categories but could not differentiate faces belonging to the same category. Rotshtein et al. (2005), Leutgeb, Leutgeb, Moser, and Moser (2005), and Lansner (2009) suggested that attractor dynamics was one of the most possible causes for these observations. However, if the morph sequences were created between unfamiliar source faces, the right FFG could differentiate two faces with a fixed distance between them on the sequence, despite their exact locations on the sequence (Jiang et al., 2006; Gilaie-Dotan & Malach, 2007).

The evidence also indicates that formation of attractors relies on past experiences. For instance, the seemingly incongruent results in Wills et al. (2005) and Leutgeb, Leutgeb, Treves et al. (2005) were speculated to be due to different training protocols (Blumenfeld et al., 2006), while the seemingly incongruent results in Rotshtein et al. (2005), Jiang et al. (2006), and Gilaie-Dotan and Malach (2007) were speculated to be due to different familiarity degrees to the stimuli. It was reported that accumulated experiences can make primates perform better in discrimination tasks (Logothetis, Pauls, & Poggio, 1995; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999), recognition (Rainer & Miller, 2000), and categorization tasks (Freedman, Riesenhuber, Poggio, & Miller, 2006). Supporting these facts, considerable evidence has indicated that training can increase the selectivity of neurons along the visual pathway of primates, including the V1 (Schoups, Vogels, Qian, & Orban, 2001), V4 (Raiguel, Vogels, Mysore, & Orban, 2006), ITC (Logothetis et al., 1995; Freedman et al., 2006), and PFC (Rainer & Miller, 2000) areas. Moreover, increased selectivity was observed even when monkeys

were passively viewing the stimuli (Freedman et al., 2006). A recent study demonstrated that extensive learning enhanced selectivity but degraded tolerance of ITC neurons (Zoccolan, Kouh, Poggio, & DiCarlo, 2007). From the viewpoint of attractor theory, these observations imply that practice can sharpen the attractive fields of patterns, a process that can be unconscious. Theoretical models have to take into account, or at least provide interfaces for incorporating, such experience-dependent learning strategies to model realistic nerve systems.

A number of attractor models have been proposed (Hopfield, 1982; Fuster, 1995; Cohen & Grossberg, 1983; Amit, 1989; Papp et al., 2007; Menghini, van Rijsbergen, & Treves, 2007; Zeng & Wang, 2007). However, their computational capabilities are greatly limited in several ways, including spurious attractors (attractors corresponding to unwanted memories), low storage capacity, inefficient pattern representations, and rather different structures from efficient information processing models. Therefore, it is difficult for them to process realistic complex patterns directly or in combination with other models such as in Mel (1997) and Riesenhuber and Poggio (1999). In addition, it is often hard to precisely manipulate the attractive fields of memories in these models, which could vary with experience. In some models (Hopfield, 1982; Cohen & Grossberg, 1983; Amit, 1989; Chua & Yang, 1988; Zeng & Wang, 2007), the attractive fields are determined by the representations of the stored memories, and it would be difficult to modify the attractive fields without affecting the representations of the memories; in other models (Papp et al., 2007; Menghini et al., 2007), a precise description of the entire dynamics is lacking, and it is hard to quantify the attractive fields analytically. These inherent deficiencies make these models difficult to incorporate in some learning strategies such as the sharpening of attractive fields.

Here we describe a simple two-layer network, termed gaussian attractor network (GAN), in which any attractor corresponds to an intentionally stored pattern, and the geometry of the attractive field of the pattern is independent of the pattern representation and can be precisely quantified. The memory capacity can be equal to the number of neurons in the output layer through unsupervised learning regardless of initial correlations among patterns to be memorized. The GAN offers a flexible framework to describe various physiological and psychological results that relate to memory and recognition by incorporating some experience-dependent learning strategies. More important, its architecture agrees well with an efficient feedforward visual recognition model, HMAX (Riesenhuber & Poggio, 1999; Serre, Oliva, & Poggio, 2007; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007), which emulates the primate ventral pathway from V1 to PFC. Based on the GAN, it is found that adding some local feedback can enable the HMAX to replicate some intriguing experimental findings mentioned earlier.

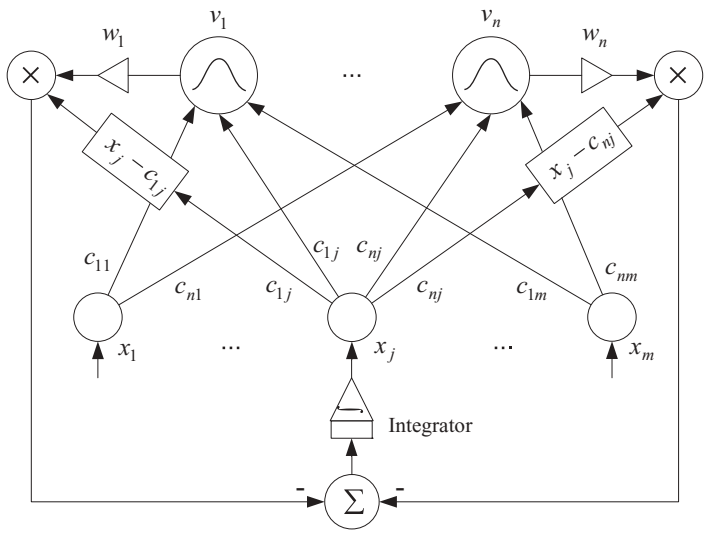

Figure 1: Schematic of the GAN. Only the feedback connections to the jth state are shown.

2 Gaussian Attractor Network

2.1 Basic Model. The proposed model consists of two layers as shown in Figure 1. A set of inputs $x_1, \ldots, x_m$ is fed into every unit v_i in the output layer through synaptic weights c_{ij} where $i = 1, \ldots, n$, $j = 1, \ldots, m$. Each unit in the output layer has a gaussian activation function v_i centered at the afferent weights, which represent a pattern to be memorized. The function v_i is multiplied by a strength factor $w_i \geq 0$ and a gain factor $(x_j - c_{ij})$ whose sign determines the nature of the afferent x_j: excitatory or inhibitory. Sum the results across i together, and feed into an integrator, and we obtain the updated x_j. The dynamics of the network is described by

$$\frac{dx_j}{dt} = -\sum_{i=1}^{n}(x_j - c_{ij})w_i v_i(\mathbf{x}), \quad j = 1, \ldots, m \tag{2.1}$$

with the output equation

$$v_i(\mathbf{x}) = e^{-\|\mathbf{x}-\mathbf{c}_i\|^2/\sigma_i^2}, \quad i = 1, \ldots, n, \tag{2.2}$$

where $\|\cdot\|$ stands for the Euclidean norm, w_i, σ_i are scalars, and $\mathbf{x}$, $\mathbf{c}_i$ are m-dimensional vectors. In what follows, w_i and σ_i are, respectively, termed the strength and width of pattern $\mathbf{c}_i$, which are closely related to the notion of the attractive field of $\mathbf{c}_i$, defined as a set from which any initial state will converge to $\mathbf{c}_i$.

Note that the major components of the proposed model, the gaussian functions, have been considered basic elements in neural networks for decades (Poggio & Edelman, 1990; Park & Sandberg, 1991), and different mechanisms have been suggested for realizing them in the brain (Knoblich, Bouvrie, & Poggio, 2007; Kouh & Poggio, 2008). Therefore, the architecture in Figure 1 is also realizable by neuronal circuits. Similar infrastructures can be produced by using VLSI technology (Lin, Huang, & Chiueh, 1998; Peng, Hasler, & Anderson, 2007), which makes it possible to devise very fast application-specific circuits (Hopfield & Tank, 1986). In what follows, we focus on the analysis of the properties of the model.

Define an energy function

$$E(\mathbf{x}) = -\frac{1}{2}\sum_{i=1}^{n} w_i \sigma_i^2 v_i(\mathbf{x}), \tag{2.3}$$

which is bounded below. It is easy to check that $dx_j/dt = -\partial E/\partial x_j$. Taking the derivative of E with respect to time t, we have

$$\frac{dE}{dt} = -\sum_{j=1}^{m}\left(\frac{\partial E}{\partial x_j}\right)^2 \leq 0.$$

The energy function decreases until its gradient vanishes, and a local minimum, maximum, or saddle point is reached. Since isolated maxima and saddle points are not stable, in view of the widely spread noise in the brain, we concentrate on the analysis of local minima (both isolated and connected) and connected local maxima. Note that $E(\mathbf{x})$ is the weighted sum of a set of inverted-bell-shaped functions $f_i(\mathbf{x}) \triangleq -w_i\sigma_i^2 v_i(\mathbf{x})/2$. It is easy to see that if the centers of these functions are far from each other or their open widths are small enough so that the open areas do not overlap, then all centers are isolated local minima of $E(\mathbf{x})$, and the state of the network from any initial point within the open area of $f_i(\mathbf{x})$ will converge to $\mathbf{c}_i$. In this case, the attractive field of pattern $\mathbf{c}_i$ is indeed the open area of $f_i(\mathbf{x})$. If an input is located outside any attractive field, that is, in the connected local maxima (also local minima) region of the energy function, the network state is called marginally stable, and this input represents a new pattern to be stored through synaptic modifications rather than a spurious attractor.

Overlap-free among open areas of $f_i(\mathbf{x})$'s corresponds to the orthogonality of binary-coded patterns in some networks (Hopfield, 1982; Amit, 1989; Chua & Yang, 1988; Chartier & Proulx, 2005; Casali, Costantini, Perfetti, & Ricci, 2006; Zeng & Wang, 2007). In contrast to those networks, which can result in many unwanted attractors even in an orthogonal case, every isolated attractor in the GAN represents a pattern intended to be stored.

If the open areas of $f_i(\mathbf{x})$'s have overlaps, the patterns are said to be correlated, and several memories may merge into one. Blumenfeld et al. (2006)

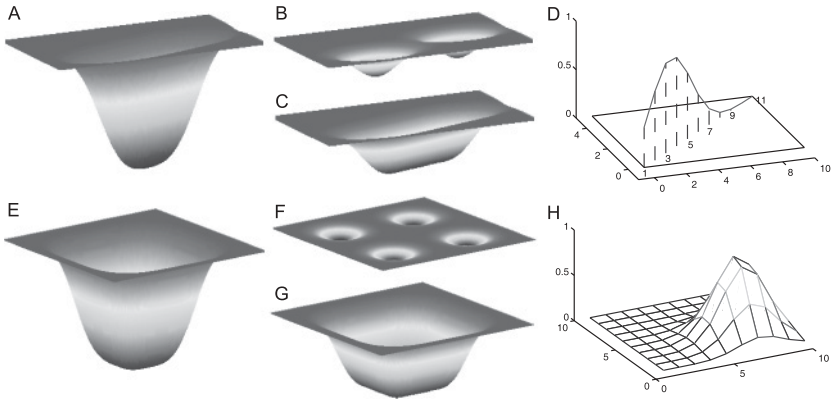

Figure 2: Storing correlated patterns represented by two-dimensional points. (A–D) Eleven patterns are evenly spread on a straight line between (0, 0) and (10, 5) with pattern widths $\sigma_i = 3$, and (E–H) 10×10 patterns are evenly spread on a square with pattern widths $\sigma_i = 2$. (A, E) Energy functions with all strengths equal to one. (B, F) Energy functions with strengths for extreme patterns equal to one and strengths for other patterns equal to zero. (C, G) Energy functions with strengths obtained through extensive training by using the learning rule, equation 2.4. (D, H) Bell-shaped steady response of the output neurons arranged on the line or the square for an arbitrary input after the formation of an approximate line attractor or plane attractor as shown in C and G. The response topographies, respectively, center around the two neurons whose centers correspond to the inputs.

demonstrated that if a sequence of gradually changing binary patterns is presented to the standard Hopfield network (Hopfield, 1982), the only attractor would be the pattern located at the middle of the sequence. This conclusion also holds for the GAN if we set identical widths and strengths for all patterns located equidistant on a straight line in the state space (see Figure 2A for a two-dimensional example and the appendix for analysis). The conclusion can be extended to the case of patterns evenly distributed in a hyperplane. If identical strengths are used, all patterns will be attracted to the center pattern as they can be regarded as densely distributed on multiple lines crossing the center pattern (see Figure 2E for a two-dimensional example).

Before moving to the next section, we point out an interesting fact. The energy function in equation 2.3 can be regarded as the output of a particular radial basis function (RBF) network with m input neurons, n hidden neurons, and one output neuron. In addition, the hidden neurons use gaussian functions as activation functions; the input weights are, respectively, the gaussian centers, and the output weights are, respectively, $-w_i\sigma_i^2/2$.

2.2 Experience-Dependent Learning. Formation of memories is a dynamic and experience-dependent process in the brain. As Blumenfeld et al.

(2006) suggested, the weights learning rule should reflect experimental observations: novel signals tend to induce higher neural activities and thus higher synaptic strength modifications than familiar signals. A modified history-dependent memory consolidation rule to that in Blumenfeld et al. (2006) is thus adopted here,

$$w_i(k+1) = w_i(k) + \alpha d + \delta, \tag{2.4}$$

where d is a parameter to quantify the level of novelty to the stimuli (the larger, the more novel), $\alpha > 0$ is learning efficiency, and $\delta > 0$ is a small increment to account for the instantaneous influence of the input. Throughout the letter, $\alpha = 0.2$ and $\delta = 0.005$ are used in simulations. Every time a stimulus is present, the corresponding weight is updated. Note that we do not specify a rule for decreasing the weights. To prevent them from increasing unboundedly, the weights are normalized by setting the maximum to one if it is greater than one.

A natural choice for d is the distance in the state space between the input and the pattern it converges to. Nevertheless, in a realistic system, d does not need to be so accurate. Here we let d be the (normalized) traveling distance of the state within a time window from the onset of the stimulus, which seems to be more biologically plausible. This is analogous to the first-step distance used by the discrete version of the Hopfield network in Blumenfeld et al. (2006). Throughout the letter, a time window of 20 time units was used in simulations. Simulation results with this time window being 50 or 100 time units were not radically different from those presented in the letter.

For storing uncorrelated patterns, d is equal to zero, and only δ takes effect. For storing correlated patterns, the above rule enables the GAN to exhibit continuous attractors dynamics, similar to the modified Hopfield network (Blumenfeld et al., 2006). Figures 2C and G visualize the energy functions for storing two sets of two-dimensional patterns. It is seen that a line attractor and a plane attractor are formed, respectively. Because of the gaussian activation function used in equation 2.2, the output of neurons elicited by any afferent is always bell shaped (see Figures 2D and H for examples).

In the following we describe the dynamic process of the GAN for storing a sequence of linearly changing patterns (see Figure 3). Suppose that the GAN has successfully memorized the two end patterns $\mathbf{c}_1$ and $\mathbf{c}_n$ (this can be attained by setting $w_1 = w_n = 1$ and other w_i's equal to zero). When pattern $\mathbf{c}_i$ is presented to the network, it can be predicted that during the first few trials, the state of the network with learning algorithm 2.4 will be attracted to $\mathbf{c}_1$ if $\mathbf{c}_i$ is located in the former half of the sequence, and it will be attracted to $\mathbf{c}_n$ if it is located in the latter half, which is confirmed by Figure 3A. Both gradual presentation (patterns in the sequence were presented from the first to the last) and random presentation (patterns were

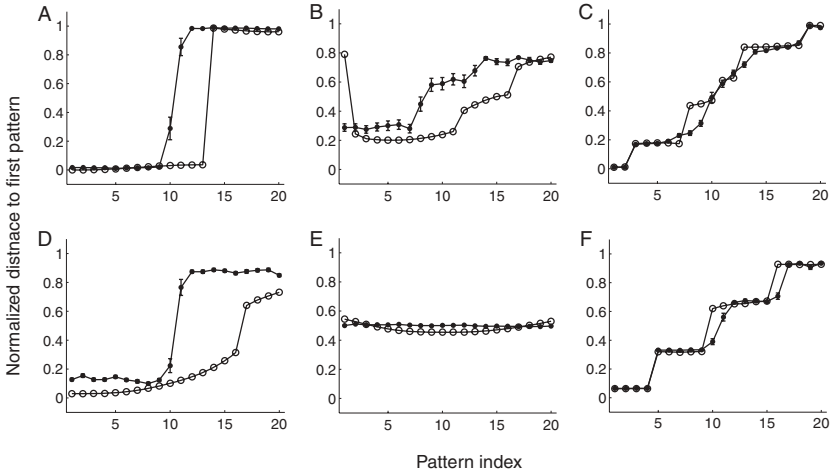

Figure 3: Learning a set of patterns represented by 20 equidistant points on a line in 100-dimensional real space with pattern widths $\sigma_i = 26$. The two extreme points are $(0, \ldots, 0)$ and $(10, \ldots, 10)$ in (A)–(C) and $(0, \ldots, 0)$ and $(4, \ldots, 4)$ in (D)–(F). (A) and (D) plot the normalized Euclidean distances of attractors to the first pattern after the first training epoch. (B) and (E) plot the normalized Euclidean distances of attractors to the first pattern after the 30th training epoch. (C) and (F) are the same as (B) and (E), respectively, except that the attractive field sharpening rule, equation 2.5, is used with $\sigma_{\min} = 6$. Filled circles: random protocol; empty circles: gradual protocol. Initially pattern strengths for the two extreme patterns were equal to one and for others were equal to zero. The maximum pattern strength was normalized to one before each training epoch if it was greater than one. For the random protocol, results were averaged over 30 independent runs (error bars: s.e.m).

presented in scrambled orders) resulted in step-like distance curves. But the curve of the gradual presentation shifts to the right slightly relative to that of the random presentation. It was found that the behavior of the network relied on the distance between the first and last patterns in the sequence. Figure 3D shows that if the distance is shortened, the step-like curve produced by the first trial of the random protocol is squeezed along the vertical axis, whereas the gradual protocol produces a semilinear curve. Figures 3A and 3D, respectively, resemble the electrophysiological data obtained on rats in Wills et al. (2005) and Leutgeb, Leutgeb, Treves et al. (2005); (see also section 4).

As the training is repeated, if the extreme patterns are distant, then each pattern tends to be drawn to itself regardless of training protocols (see Figure 3B); otherwise, all the memories merge to one (see Figure 3E). The complete merging prediction contradicts the common belief that practice

improves discrimination and recognition ability. As mentioned in section 1, evidence suggests that attractive fields of recurrent neuronal networks distributed in the nerve system may shrink with accumulated experiences. This effect can be easily accounted for by decreasing the pattern width σ_i's. Clearly the amount of the sharpening effect should also depend on experience. It is reasonable to assume that multiple exposures or a prolonged exposure time to a pattern would lead to further decrease of its width than a single short period exposure. In particular, if pattern $\mathbf{c}_i$ is presented repeatedly, we adopt the following simple rule in this letter (though other rules may be more appropriate),

$$\sigma_i(k+1) = \max(\beta\sigma_i(k), \sigma_{\min}), \tag{2.5}$$

where $0 < \beta < 1$ denotes the rate of decrease and $\sigma_{\min}$ denotes the biological limit ($\beta = 0.95$ is always used in this letter). If $\sigma_{\min}$ could be arbitrarily small, then eventually all patterns would be memorized since the correlations among them would fade as training progresses.

Figures 3C and 3F demonstrate the behavior of the network by adding the attractive field sharpening rule, equation 2.5. After extensive training, for the large distance case (see Figure 3C), this rule enhances the discrimination performance of the network slightly, while for the short distance case (see Figure 3F), it enhances performance dramatically. Clearly, with this rule, the prediction about discrimination ability after extensive training based on the GAN agrees with that in Blumenfeld et al. (2006).

However, if $\sigma_{\min}$ is not allowed to be so small, then all memories will merge to one as Figure 3E shows. We think that this is possible in biological systems. So it is speculated that there exist neurons in the nerve system that can differentiate the stimuli at first, but they will lose this ability after extensive exposure to many intermediate stimuli.

3 HMAX Model with Feedback Connections for Visual Recognition

HMAX model is a hierarchical feedforward network proposed by Riesenhuber and Poggio (1999) for mimicking the primate visual system. It consists of layers with linear and nonlinear units, that correspond to simple cells and complex cells in Hubel and Wiesel's paradigm. From lower to higher tiers, simple features are combined to build complex ones; meanwhile, perception invariance to translation and scale of objects is realized (see Figure 4). Each stimulus is represented by a shape-tuned unit (STU) in the top layer, modeled by a radial basis function (e.g., gaussian function). These STUs serve as inputs to task-relevant category-tuned units (CTUs, not shown), and the weights are determined by supervised learning. It was assumed that the STU-like neurons are mainly located in the ITC, while the CTU-like neurons are mainly located in the PFC (Riesenhuber & Poggio,

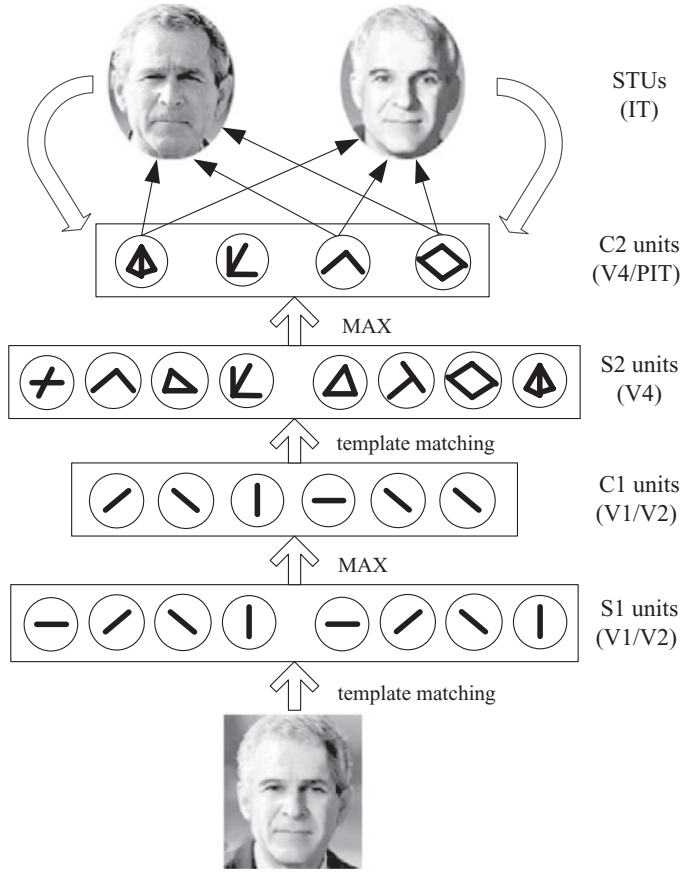

Figure 4: HMAX model with feedback connections. This particular HMAX implementation (Serre, Wolf et al., 2007) consists of five layers—S1, C1, S2, C2, and STU—whose major locations in the visual pathway are indicated on the right. The two S layers and the STU layer perform template matching (or weighted sum), realizing the increase of feature complexity, and the two C layers perform a MAX operation on their afferents, realizing the scale, orientation, and translation invariance. Gaussian functions are adopted here as activation functions of STUs. See Serre, Wolf et al. (2007) for details. Feedback connections from STU layer to C2 layer can be drawn as in Figure 1. (Photos courtesy of Pia Rotshtein.)

1999; Serre, Oliva et al., 2007; Freedman, Riesenhuber, Poggio, & Miller, 2003). The prediction of this network agrees with a number of findings on monkeys (Logothetis et al., 1995; Wang, Tanifuji, & Tanaka, 1998; Freedman et al., 2003; Cadieu et al., 2007) and humans (Serre, Oliva et al., 2007; Jiang et al., 2006) when they were engaged in visual recognition, discrimination, or categorization tasks.

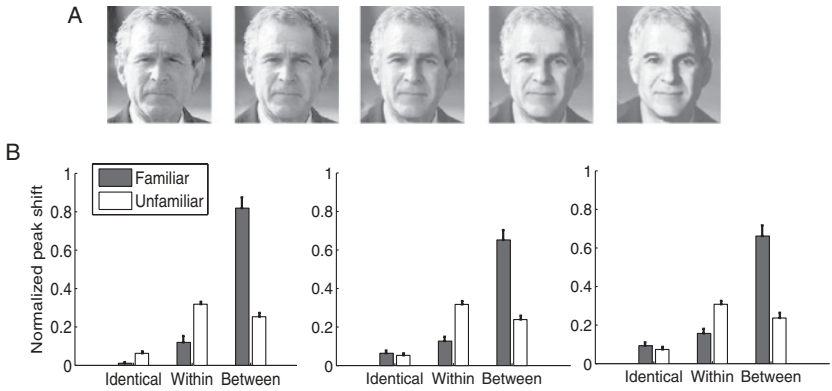

Figure 5: Simulation results of the modified HMAX for face processing in the right FFG of human brains. Forty-six sequences of images were generated by 92 facial images (sources), and each sequence consisted of 10 images with an approximately equal difference between two consecutive ones. (A) Five images (morph index: 1, 3, 5, 7, 9) in an example morph continuum. (B) Five pairs of morphs (identical: 4 versus 4, 7 versus 7; within: 1 versus 4, 7 versus 10; between: 4 versus 7) in each sequence were presented to the model in random order (which one in a pair came first was also random), and the peak shift over the morph index between the two resulting steady output responses was calculated to represent the difference in neuronal activities elicited by each pair of stimuli. Plotted is the mean + s.e.m. peak shift (normalized by dividing 9) across 46 independent runs. From left to right, the panels respectively show the results with d in 2.4 equal to the state shift within the first 20, 50, and 1000 time units. In the familiar case, initially w_1 and w_{10} were set to ones and the other w_i's were set to zeros. In the unfamiliar case, initially all w_i's were set to zeros. σ_1 and σ_{10} were set such that in the familiar case, the formal and latter five images were respectively attracted to the first and last images without learning rules. Other σ_i's were then linearly assigned values between σ_1 and σ_{10}. $\sigma_{\min}$ for each stimulus was equal to 80% of its initial value. (Photos courtesy of Pia Rotshtein.)

If a sequence of gradually morphed stimuli are presented, HMAX should differentiate every stimulus by the peak response of the STUs. However, this prediction was not validated in some regions of the ITC in an fMRI experiment (Rotshtein et al., 2005). Actually it was reported that when a pair of faces with 30% difference along a sequence of continuously morphed faces (see Figure 5A) between two famous faces (source faces) was presented to human subjects, the blood oxygenation level–dependent (BOLD) signal change in the right FFG, including the right FFA, was large if the pair crossed the perceived identity boundary obtained by behavioral tests, but small if the pair did not cross the boundary. In the latter case, the signal change in these areas was comparable with that for two identical images. However, other studies (Jiang et al., 2006; Gilaie-Dotan & Malach, 2007) with

similar experimental procedures showed that pairs 30% apart on a morph sequence could elicit a distinguishable BOLD signal change in the right FFG regardless of their exact locations on the sequence. This finding was ascribed to unfamiliar source faces used in the experiments. From attractor theories, the familiar sources imply the presence of preformed attractors, and the unfamiliar sources imply the absence of such attractors.

As purely feedforward architectures such as HMAX preclude attractor dynamics, recurrent or feedback connections must be taken into account in building models to interpret these findings. To endow the HMAX with this capability, a possible modification is to add feedback connections from the STU layer to the C2 layer as illustrated in Figure 1. Then every STU in Figure 4 corresponds to an output neuron v_i in Figure 1, and every C2 unit in Figure 4 corresponds to an input neuron x_j in Figure 1. The feedback signal from any STU to any C2 unit is the product of the output signal and the difference signal between the current state and the memorized state, modulated by a synaptic strength w_i. For simplicity, we consider only feedback from STUs to C2 units in Figure 4. In principle, such feedback can exist from the STU layer to any layer in the downstream direction if we add feedforward connections directly from the lower layers to the STU layer. In that case, the afferent to the STUs will contain more detailed but less invariant information.

This modification enables the model to replicate Rotshtein et al.'s (2005) findings. The experiment settings are as follows. The initial stimulus set consisted of 54 morph continua between 108 achromatic portraits of famous people, and each continuum consisted of 11 morphs representing gradual transitions from one source face to the other in steps of 10% change (see Figure 5A for an example). (The stimuli are by courtesy of Pia Rotshtein; more information can be found in Rotshtein et al., 2005.) All images were in gray scale and resized to 160×180 pixels. In our experiments, only the first 10 morphs were selected for convenience in deciding within- and between-category stimuli (therefore, the second source image in each continuum was the tenth image, not exactly a source for the morph algorithm). The category boundary for each continuum can be determined by inputting the morphs and checking the attractor locations with $w_1 = w_{10} = 1, w_2 = \cdots = w_9 = 0$ (i.e., familiar sources case). In Rotshtein et al. (2005), the boundary was determined by behavioral tests. But the actual location of the psychological boundary in a continuum is not essential in simulation studies. So, for convenience, it is desired that the category boundary is between the fifth and sixth images in every continuum. Then, similar to Rotshtein et al. (2005), morphs 1 versus 4 and morphs 7 versus 10 were chosen as two within-category pairs; morphs 4 versus 7 were chosen as a between category pair; morphs 4 versus 4 morphs and 7 versus 7 were chosen as two identical pairs. To satisfy the requirement, different σ_1 and σ_{10} should be used since the morph algorithm is not exactly linear. Here, they were randomly chosen between $0.2 - 0.4L$, where L denotes the Euclidean distance between the

C2 features of the two source patterns extracted by HMAX, such that in the familiar-sources case, the first five and the last five morphs were respectively attracted to the two source patterns without any learning rule. Other σ_i's were linearly assigned values between σ_1 and σ_{10} for each continuum. Eight morph continua that have too much nonlinearity (the difference of σ values for the two source patterns exceeds $0.2L$) were excluded in the experiments. Learning rules 2.4 and 2.5 were adopted with default parameter values.

An HMAX implementation described in Serre, Wolf et al. (2007) was used for simulations so the S2 units were learned instead of predefined as in Riesenhuber and Poggio (1999). To learn the S2 units, 400 patches of four sizes at random positions were extracted from each continuum at the C1 level, which led to 400 C2 features for representing an image. (More information is found in Serre, Wolf et al., 2007.)

After the presentation of each stimulus, the S and C layers extract the features and the outputs of the STUs from a gaussian response curve. Due to the feedback connections from the STU layer to the C2 layer, the values of C2 units may change with time. As a consequence, the peak location of STU response curve may also change with time. Eventually an equilibrium is reached, and both the C2 units and the STUs become constants. Then the peak location of the STU response curve indicates the identity of the stimulus from the perspective of the network. If two stimuli are presented, in the resting phase, the peak locations of the STU response curve can be identical or different, depending on the physical properties of the stimuli and inherent dynamics. The difference between the peak locations, or peak shift, can resemble the fMRI signal change evoked by a pair of stimuli. Clearly the larger the shift, the larger the difference between the stimuli from the perspective of the network. If the shift is equal to zero, the two stimuli are treated the same. This is consistent with the assumption of fMRI adaptation experiments.

Figure 5B (left panel) plots the statistics of the peak shifts by inputting different pairs of morphs (between, within, and identical conditions). In the familiar-sources case, a significantly larger peak shift of STUs in the between condition is observed than in the within condition, though the distances in the continuum between the two images in the two conditions are nearly the same. The difference of shifts between the within and identical conditions is small. In the unfamiliar-sources case, the peak shift in either the between or within condition is significantly greater than in the identical condition, but only a small difference is found between the between and within conditions. We have investigated the impact of the time window for d in learning rule 2.4. The middle and right panels of Figure 5 plot the results with the time window equal to 50 and 1000 time units. The results are similar to those in the left panel with the default time window (20 time units).

It was reported that subjects who were shown a series of morphed visual stimuli between two source stimuli were more likely to perceive the identity

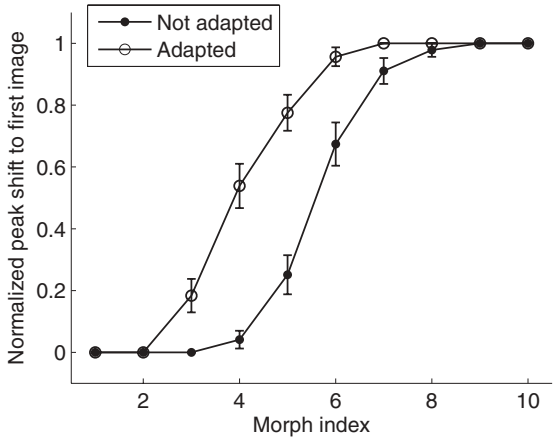

Figure 6: Simulation results of the modified HMAX for visual perceptual adaptation. All morphs in each of the 46 sequences were presented to the model in a scrambled order with and without a preceding presentation of the adapting stimulus (the first image in the sequence). Plotted is the mean $\pm$ s.e.m. normalized peak shift of the STU response curve at resting phase to the first image across 46 independent runs. For adapting stimuli, σ was set to $\sigma_{\min}$ (70% of its initial value) directly instead of using learning rule 2.5. Other parameter settings were the same as in the familiar case of Figure 5. Similar curves will be produced if we quantify the difference between attractors to the first image by Euclidean distance in the C2 space (as in Figure 3).

of one source in an ambiguous pattern after they had been exposed to the other source (called adapting stimuli) for a long time. The sources could be facial images (Webster, Kaping, Mizokami, and Duhamel, 2004; Fox & Barton, 2006), body images (Winkler & Rhodes, 2005) or object views (Fang & He, 2005), which were familiarized to subjects before testing. From the viewpoint of attractor theory, these findings can be interpreted as caused by attractive field sharpening. We tested this assumption as follows. First, we presented the 46 sequences of morphs used in Figure 5 to the modified HMAX model by presetting the strengths of sources as ones and the others as zeros. As expected, the perception boundary was found to be between the morphs 5 and 6 (see the curve with filled circles in Figure 6). Then with the same initial parameter settings, we repeated the experiments. But this time, before each morph presentation, an adapting image (the first morph in every continuum) was presented to the model for a much longer period, and this fact was accounted for by decreasing 30% of the corresponding σ. As indicated previously, prolonged presentation of the adapting stimulus will cause a significant sharpening of its attractive field, corresponding to a significant decrease of σ in the GAN. The curve with empty circles in

Figure 6 shows a large "perception" boundary shift toward the adapting stimuli, which is consistent with the results of the behavioral tests.

4 Discussion

The dynamics in every deterministic attractor network can be regarded as originating from a search process for a local minimum of an abstract energy function, and the proposed gaussian attractor network (GAN) implements this idea intuitively. The energy function of the model is a weighted sum of a set of inverted gaussian functions whose centers represent patterns to be memorized and whose open widths determine the attractive basins. This novel function offers many advantages for the GAN in modeling various results from single-cell recordings to a high-level perception, including no spurious attractors, large storage capacity, and dissociation of the memory representation and its attractive field. In addition, the structure of the GAN is compatible with a well-known visual recognition model HMAX (Mel, 1997; Riesenhuber & Poggio, 1999; Serre, Oliva et al., 2007; Serre, Wolf et al., 2007), and the combination of them can quantitatively reproduce some experimental data.

4.1 Experience-Dependent Learning. The novelty-facilitated learning rule, equation 2.4, and the correlation between the preformed memories can reconcile some apparently conflicting data on rats. Wills et al. (2005) reported that when rats were exploring different environments, which constituted a morph sequence, the firing rate pattern in the CA1 region was either similar to that in the first or that in the last environment on the sequence, and the two extreme environments were familiarized to them before these tests. But a different finding was reported in Leutgeb, Leutgeb, Treves et al. (2005). It was found that during tests, the firing rate patterns in both the CA1 and CA3 regions change continuously from the first to the last environments. There are two major discrepancies between the experimental procedures. First, the environments were presented to rats in a scrambled order in Wills et al. (2005) but in a gradual order in Leutgeb, Leutgeb, Treves et al. (2005). Second, before tests, the rats in Wills et al. (2005) exhibited stronger firing pattern dissociation in the two extreme environments than in Leutgeb, Leutgeb, Treves et al. (2005), which suggests more distinct representations of the two environments in the CA region in Wills et al. (2005). In the language of the GAN, this means a larger distance between the two extreme patterns in the state space. Therefore, the filled circle curve in Figure 3A and the empty circle curve in Figure 3D, respectively, resemble the two findings. Note that in Blumenfeld et al. (2006), only the exposure order was identified as a cause of these findings, whereas the initial correlation between the two extreme pattern representations was not identified.

If the two extreme patterns have large correlations, our simulations showed that without attractive field sharpening, all memories will merge

into one after extensive training (see Figure 3E). The attractive field sharpening strategy could enable the nerve system to memorize all patterns between the two extreme ones (see Figure 3F). In a realistic system, however, because of the biological limit on the attractive field width, we speculate that even with attractive field sharpening, if two stimuli are too similar, there exist neurons in the nerve system that can differentiate the stimuli at first (see Figure 3D), but will lose this ability after extensive training with many physically intermediate stimuli (see Figure 3E). In other words, we are questioning the notion that training always improves the discrimination ability of neurons, as Blumenfeld et al. (2006) claimed. A psychological study seems to support our prediction (Preminger, Sagi, & Tsodyks, 2007). After they were shown morph sequences of faces gradually and repeatedly, subjects tended to lose the ability to discriminate the morphs that were created between similar source faces. This effect did not occur when the perceived similarity between the source faces was low. Nevertheless, this effect did not occur with a random training protocol no matter how similar the source faces were, which cannot be explained by the GAN. Physiological studies are required to test this prediction at the neuron level.

Continuous attractors have been taken to explain many experimental observations, for example, line attractor for eye position memory (Seung et al., 2000), plane attractor for place cells of rodents (Samsonovich & McNaughton, 1997), and ring attractor (actually a one-dimensional line attractor along angular axis in the range of 2π) for orientation tuning cells in the V1 area of primates (Ben-Yishai, Bar-Or, & Sompolinsky, 1995) and head direction tuning cells of rodents (Zhang, 1996; Sharp et al., 2001). A major limitation of these models is that the connections are prespecified, though there are exceptions (Stringer, Rolls, Trappenberg, & de Araujo, 2002). In contrast, the GAN with rules 2.4 and 2.5 can learn to exhibit various dynamics, including discrete attractor (see Figures 2B and 2F), line attractor (see Figure 2C), and plane attractor (see Figure 2G) dynamics in a self-organizing manner. It thus has great potential for modeling related phenomena. For instance, inside the plane attractors formed by the GAN, the output neurons can generate a bump activity everywhere (see Figure 2H), representing a chart or a cognitive map inside the hippocampal field of rats (Samsonovich & McNaughton, 1997; McNaughton, Battaglia, Jensen, Moser, & Moser, 2000). The small increment signal δ in equation 2.4 drives the bump toward the correct location on the chart, representing the force exerted by external environmental stimuli or events. However, to emulate a parallel navigational system that is also able to calculate where an animal is at the moment but independent of environmental clues, path integration needs to be performed (Samsonovich & McNaughton, 1997; McNaughton et al., 2000).

4.2 Modified HMAX Model. A tacit assumption behind the purely feedforward architecture of the original HMAX model (Riesenhuber &

Poggio, 1999) is that during a very short time interval after the stimulus onset, top-down signals are not likely to be present given the number of processing stages involved and typical neural latencies (Serre, 2007). This assumption is supported by electrophysiological (Hung, Kreiman, Poggio, & DiCarlo, 2005) and psychological (Serre, Oliva et al., 2007) studies on primates. But it does not preclude transmission of local feedback signals inside ITC or from ITC to V4 (see Figure 4) after that short period. Note that a very short (<200 ms) presentation of stimuli might be sufficient for readout of the identity information of dissimilar objects (Hung et al., 2005) or for accomplishing the animal versus nonanimal categorization task (Serre, Oliva et al., 2007), but it is insufficient for recognizing similar objects. Actually, in the experiments that we have modeled with the modified HMAX (Rotshtein et al., 2005; Webster et al., 2004; Fox & Barton, 2006; Winkler & Rhodes, 2005; Fang & He, 2005), the stimuli were usually presented for more than 200 ms. Moreover, the images used in Hung et al. (2005) and Serre, Oliva et al. (2007) have large variations compared with the morph images used in Rotshtein et al. (2005); Webster et al. (2004), Fox and Barton (2006), Winkler and Rhodes (2005), and Fang and He (2005). It is easy to see that if the afferent patterns are not correlated, then the modified HMAX can give similar results to the original model in view of the fact that the minima of the energy function of the GAN roughly correspond to the individual efferent neurons' maxima in this case.

In neurophysiological experiments, categorical cells are often found to be concomitant with cells tuning to individual stimuli, and the two types of cells are believed to respectively mediate the pattern completion and pattern separation channels for memory and recognition. When sets of morphed stimuli were shown to monkeys, both categorical cells and differentiating cells are found in the ITC and PFC (Freedman et al., 2003). However, the categorical cells in the ITC are less task relevant than in the PFC, suggesting that the categorical phenomenon in the ITC is more likely a result of attractors without supervised learning, whereas the phenomenon in the PFC is a result of supervised learning, which is consistent with the prediction of the modified HMAX. On the other hand, fine tuning to individual afferent patterns does not exclude the possibility that attractors with sufficiently small attractive fields are the underlying mechanism.

It is known that feedback in the vertebrate visual system exists almost everywhere. We have shown that the HMAX with feedback from the STU layer to the C2 layer is already capable of reproducing some experimental data in the IT cortex (see Figures 5 and 6). At present, it is unclear how to add the feedback connections between every two layers in HMAX. But in principle, neither the feedforward nor the feedback connections are necessary to be between two adjacent layers. For example, the C1 layer can have direct feedforward connections to the STU layer, while the STU layer can have direct feedback connections to the C1 layer, and the connections can be drawn similarly as between the C2 and STU layers (see Figure 4). It

depends on how detailed information is required by the STUs. In view of the agreement between the simulation results and the experimental data, it might be appropriate to think that the particular feedback connections in the modified HMAX shown in Figure 4 have accounted for the effect of all feedback connections along the ventral pathway. There also exist recurrent connections in the same tier in this pathway. The modified HMAX does not explicitly model this fact. But some lateral inhibitory connections are believed to be existent in the STU layer to do weights normalization (see, e.g., Kouh & Poggio, 2008). This is to prevent the weights from increasing too much with equation 2.4; otherwise all memories would completely corrupt to one with repeated presentations of stimuli.

Note that other attractor networks may also be combined with the HMAX on the top two layers. We speculate that such integrated architectures could produce similar results to those in Figures 5 and 6. The advantages of the GAN discussed in section 2 may not distinguish it from other models in reproducing these experimental data. From the viewpoint of the HMAX, it is the structural similarity with the HMAX's C2 and STU layers that distinguishes the GAN from other attractor networks. Note that the modified HMAX architecture, illustrated in Figure 4, has the same number of C2 neurons and STU neurons as in the original HMAX, while the STU neurons have the same bell-shaped tuning curves as in the original HMAX, which are believed to be a hallmark of IT neurons (Logothetis et al., 1995). This integrated model can be regarded as resulting from adding some feedback connections to the original HMAX as if no other models were incorporated. Existing attractor networks cannot be combined with the HMAX so well. For example, if a Hopfield network is used, the number of neurons in the C2 and STU layers must be different from those in the original HMAX.

The modified HMAX model can successfully reproduce the perceptual adaptation aftereffect (see Figure 6) based on the principle that prior experience to a stimulus sharpens the tuning curves of neurons (Schoups et al., 2001; Logothetis et al., 1995; Freedman et al., 2006; Rainer & Miller, 2000; Raiguel et al., 2006; Zoccolan et al., 2007). But this is not the only possible mechanism underlying this phenomenon. An alternative cause might be neural fatigue, which can suppress responses to repeated or adapted stimuli while enhancing responses to novel or nonadapted stimuli. A theoretical model implementing this idea by decaying firing rates for output neurons for a long time can also reproduce such aftereffects (Menghini et al., 2007). Further evidence is needed to validate the proposals.

Appendix

Here we show that a set of correlated patterns $\mathbf{c}_i$ evenly distributed on a line segment L in the state space with identical widths σ_i's and strengths w_i's can result in one attractor only at the middle of L for the GAN. What we only need to show is that if the number of patterns is large, the energy function

$E(\mathbf{x})$ defined in equation 2.3 has a unique minimum at the midpoint of L. Note that we do not consider the points far away from the line segment L, which constitute the connected local maxima region.

For convenience, let

$$f_i(\mathbf{x}) = -\frac{1}{2} w_i \sigma_i^2 v_i(\mathbf{x}). \tag{A.1}$$

Then the energy function defined in equation 2.3 becomes $E(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$. Before the analysis, we give a definition. A continuously differentiable function $g : R^m \to R$ is called a *pit function* if it (i) has a unique minimum at $\mathbf{c}$ and (ii) is symmetric about $\mathbf{c}$, i.e., $g(\mathbf{x}) = g(2\mathbf{c} - \mathbf{x})$. In addition, $\mathbf{c}$ is called the center of g. Clearly the function $f_i(\mathbf{x})$ defined in equation A.1 is a pit function. The aim is to show that $E(\mathbf{x})$ is also a pit function. The proof is given in three steps.

Step 1 (minima must be on L). It is easy to see that $E(\mathbf{x})$ has at least one local minimum. We need to ensure that any minimum must be located on L, which can be reasoned as follows. For any $\mathbf{x} \notin L$,

$$\nabla E(x) = \sum_{i=1}^{n} \nabla f_i = \sum_{i=1}^{n} \gamma_i(\mathbf{x})(\mathbf{x} - \mathbf{c}_i), \tag{A.2}$$

where ∇ denotes the gradient and $\gamma_i(\mathbf{x}) = w_i e^{-\|\mathbf{x}-\mathbf{c}_i\|^2/\sigma_i^2} > 0$. Let $\mathbf{u}$ denote the unit vector along L and define a linear subspace $V = \text{span}\{\mathbf{u}\}$. Decompose the vector $(\mathbf{x} - \mathbf{c}_i)$ into two components $(\mathbf{x} - \mathbf{c}_i)_{//} \in V$ and $(\mathbf{x} - \mathbf{c}_i)_{\perp} \in V^{\perp}$ where $V^{\perp}$ is the orthogonal complement of V in the m-dimensional real space. Then $(\mathbf{x} - \mathbf{c}_i)_{//} = \frac{\langle \mathbf{x}-\mathbf{c}_i, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} = \langle \mathbf{x} - \mathbf{c}_i, \mathbf{u} \rangle \mathbf{u}$ where $\langle \cdot \rangle$ denotes the inner product. It follows that

$$\begin{aligned}
(\mathbf{x} - \mathbf{c}_i)_{\perp} &= \mathbf{x} - \mathbf{c}_i - (\mathbf{x} - \mathbf{c}_i)_{//} \\
&= \mathbf{x} - \mathbf{c}_i - \langle \mathbf{x} - \mathbf{c}_i + \mathbf{c}_j - \mathbf{c}_j, \mathbf{u} \rangle \mathbf{u} \\
&= \mathbf{x} - \mathbf{c}_i - \langle \mathbf{x} - \mathbf{c}_j, \mathbf{u} \rangle \mathbf{u} - \langle \mathbf{c}_j - \mathbf{c}_i, \mathbf{u} \rangle \mathbf{u} \\
&= \mathbf{x} - \mathbf{c}_j - \langle \mathbf{x} - \mathbf{c}_j, \mathbf{u} \rangle \mathbf{u} = (\mathbf{x} - \mathbf{c}_j)_{\perp} \neq 0.
\end{aligned}$$

Let $\nabla_{//} E = \sum_{i=1}^{n} \gamma_i(\mathbf{x})(\mathbf{x} - \mathbf{c}_i)_{//}$ and $\nabla_{\perp} E = \sum_{i=1}^{n} \gamma_i(\mathbf{x})(\mathbf{x} - \mathbf{c}_i)_{\perp}$; then $\nabla E = \nabla_{//} E + \nabla_{\perp} E$ and the two components $\nabla_{//} E$ and $\nabla_{\perp} E$ are orthogonal. Obviously $\nabla_{\perp} E$ is in no way equal to zero. In other words, the gradient of $E(x)$ will never vanish outside L, and any minimum must be on L. See Figure 7A for the geometric interpretation of these arguments.

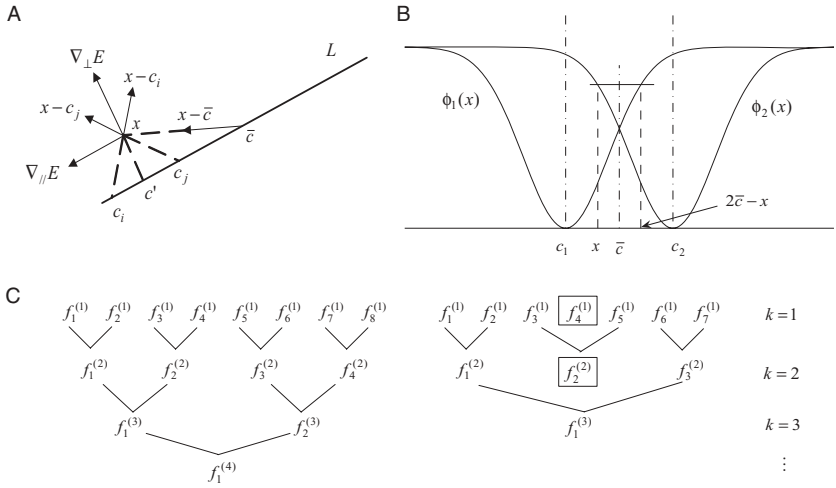

Figure 7: Illustration of complete merging of attractors for correlated patterns. (A) Decomposition of $\nabla E(\mathbf{x})$ for patterns on a line segment L in the state space when $\mathbf{x} \notin L$. (B) Summing two pit functions centered at $\mathbf{c}_1$ and $\mathbf{c}_2$ results in another pit function centered at $\bar{\mathbf{c}} = (\mathbf{c}_1 + \mathbf{c}_2)/2$. (C) Two examples for arranging the order of summations of n pit functions. (Left) At every step k, the number of pit functions $f_i^{(k)}$ to be paired consecutively is even, and the maximum interval $\theta^{(k)}$ required to result in the half number of new pit functions can be determined by any pair. (Right) At $k = 1$ and $k = 2$, the numbers of pit functions to be paired are both odd. So at $k = 1$, $f_4^{(1)}$ is excluded from the sum, and $\theta^{(1)}$ is determined by $f_3^{(1)}$ and $f_5^{(1)}$, not by $f_1^{(1)}$ and $f_2^{(1)}$, or $f_6^{(1)}$ and $f_7^{(1)}$. Likewise, at $k = 2$, $f_2^{(2)}$ is excluded from the sum, and $\theta^{(2)}$ is determined by $f_1^{(2)}$ and $f_3^{(2)}$.

Step 2 (two pit functions merge to one). Let $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ be two pit functions of the same shape centered at $\mathbf{c}_1$ and $\mathbf{c}_2$ on L, respectively, and $h(\mathbf{x}) = \phi_1(\mathbf{x}) + \phi_2(\mathbf{x})$. Clearly h has at least one local minimum. We prove the following result in what follows: if all local minima of h are on L and $\mathbf{c}_1$ and $\mathbf{c}_2$ are close enough, then h is also a pit function, centered at $\bar{\mathbf{c}} = (\mathbf{c}_1 + \mathbf{c}_2)/2$. Clearly, if ϕ_1 and ϕ_2 are inverted gaussian functions defined in equation A.1, according to step 1, the condition is satisfied.

The same shape of ϕ_1 and ϕ_2 implies that $\phi_1(\mathbf{x}) = \phi_2(\mathbf{c}_2 - \mathbf{c}_1 + \mathbf{x})$ and $\phi_2(\mathbf{x}) = \phi_1(\mathbf{c}_1 - \mathbf{c}_2 + \mathbf{x})$. It follows that

$$\begin{aligned} h(\mathbf{x}) &= \phi_2(\mathbf{c}_2 - \mathbf{c}_1 + \mathbf{x}) + \phi_1(\mathbf{c}_1 - \mathbf{c}_2 + \mathbf{x}) \\ &= \phi_2(2\mathbf{c}_2 - 2\bar{\mathbf{c}} + \mathbf{x}) + \phi_1(2\mathbf{c}_1 - 2\bar{\mathbf{c}} + \mathbf{x}) \\ &= \phi_2(2\bar{\mathbf{c}} - \mathbf{x}) + \phi_1(2\bar{\mathbf{c}} - \mathbf{x}) = h(2\bar{\mathbf{c}} - \mathbf{x}). \end{aligned}$$

Hence, h is symmetric about $\bar{\mathbf{c}}$. In addition, it is trivial to show that ϕ_1 and ϕ_2 are symmetric about $\bar{\mathbf{c}}$ (i.e., $\phi_1(\mathbf{x}) = \phi_2(2\bar{\mathbf{c}} - \mathbf{x})$).

Let $q(\phi_i) = \langle \nabla\phi_i, \mathbf{u} \rangle$ denote the projection of the gradient ϕ_i on L, where $i = 1, 2$. Without loss of generality, assume

$$q(\phi_1) \begin{cases} < 0, \forall \mathbf{x} \in (-\infty, \mathbf{c}_1), \\ = 0, \forall \mathbf{x} = \mathbf{c}_1, \\ > 0, \forall \mathbf{x} \in (\mathbf{c}_1, +\infty), \end{cases} \qquad q(\phi_2) \begin{cases} < 0, \forall \mathbf{x} \in (-\infty, \mathbf{c}_2), \\ = 0, \forall \mathbf{x} = \mathbf{c}_2, \\ > 0, \forall \mathbf{x} \in (\mathbf{c}_2, +\infty), \end{cases}$$

and $\mathbf{c}_1$ is on the left of $\mathbf{c}_2$ on L (i.e., $\mathbf{c}_1 \in (-\infty, \mathbf{c}_2)$; see Figure 7B), where ∞ denotes the infinity along the direction of L, and $(\cdot, \cdot)$ denotes a line segment between two points (parentheses indicate exclusion of the end point; similarly, braces indicate inclusion of the end point). Since by assumption, any local minimum of h is located on L, to show that h has a unique minimum at $\bar{\mathbf{c}}$, what is needed is to show

$$q(h) = q(\phi_1) + q(\phi_2) \begin{cases} < 0, \forall \mathbf{x} \in [\mathbf{c}_1, \bar{\mathbf{c}}), \\ = 0, \forall \mathbf{x} = \bar{\mathbf{c}}, \\ > 0, \forall \mathbf{x} \in (\bar{\mathbf{c}}, \mathbf{c}_2], \end{cases}$$

for sufficiently close $\mathbf{c}_1$ and $\mathbf{c}_2$. According to the symmetry of ϕ_1 and ϕ_2, it is needed only to prove the first line of the previous equation. The symmetry of ϕ_1 and ϕ_2 about $\bar{\mathbf{c}}$ implies $q(\phi_2(\mathbf{x})) = -q(\phi_1(2\bar{\mathbf{c}} - \mathbf{x}))$. Then

$$q(h) = q(\phi_1(\mathbf{x})) - q(\phi_1(2\bar{\mathbf{c}} - \mathbf{x})).$$

Since $q(\phi_1)$ is equal to zero when $\mathbf{x} = \mathbf{c}_1$ and greater than zero when $\mathbf{x} \in (\mathbf{c}_1, +\infty)$, there must exist $\tilde{\mathbf{c}} \in (\mathbf{c}_1, \infty)$ such that $q(\phi_1(\mathbf{x}))$ monotonically increases when $\mathbf{x}$ moves from $\mathbf{c}_1$ to $\tilde{\mathbf{c}}$ along L. Let $\mathbf{c}_2 = \tilde{\mathbf{c}}$. Consider $\mathbf{x} \in [\mathbf{c}_1, \bar{\mathbf{c}})$. If $c_{1,j} < c_{2,j}$ for some $j = 1, \ldots, m$, where $c_{s,j}$ denotes the jth component of $\mathbf{c}_s$, then $c_{1,j} \le x_j < \bar{c}_j < c_{2,j}$, which follows $x_j < 2\bar{c}_j - x_j = c_{1,j} + c_{2,j} - x_j \le c_{2,j}$. Similarly, if $c_{1,j} > c_{2,j}$ for some j, we can deduce $x_j > 2\bar{c}_j - x_j \ge c_{2,j}$. Taken together, we have $2\bar{\mathbf{c}} - \mathbf{x} \in (\mathbf{x}, \mathbf{c}_2]$ for $\mathbf{x} \in [\mathbf{c}_1, \bar{\mathbf{c}})$. It follows that $q(h) < 0$ for $\mathbf{x} \in [\mathbf{c}_1, \bar{\mathbf{c}})$. Therefore, h is a pit function.

Step 3 (n inverted gaussian functions merge to one). It is easy to see that if the sum of two pit functions of the same shape results in another pit function, then adding a third pit function centered at the middle of their centers, no matter whether it is the same shape, will also result in a pit function.

Denote each pit function f_i defined in equation A.1 by $f_i^{(1)}$, corresponding center by $\mathbf{c}_i^{(1)}$, the maximum interval between $\mathbf{c}_i^{(1)}$ and $\mathbf{c}_{i+1}^{(1)}$ required to make $f_i^{(1)} + f_{i+1}^{(1)}$ to be a pit function by $\theta^{(1)}$, and the actual interval between

$\mathbf{c}_i^{(1)}$ and $\mathbf{c}_{i+1}^{(1)}$ by δ. Suppose that n is even. Let $\delta = \theta^{(1)}$. We can pair the functions consecutively, sum respectively, and obtain $n/2$ pit functions, denoted by $f_i^{(2)}$'s with centers $\mathbf{c}_i^{(2)}$'s. Denote the maximum interval between $\mathbf{c}_i^{(2)}$ and $\mathbf{c}_{i+1}^{(2)}$ required to make $f_i^{(2)} + f_{i+1}^{(2)}$ be a pit function by $\theta^{(2)}$ (note that any minimum of $f_i^{(2)} + f_{i+1}^{(2)}$ must be on L because it is actually the sum of several $f_i^{(1)}$'s (see step 1). This merging effect can be ensured by letting $\delta = \min\{\theta^{(1)}, \theta^{(2)}/2\}$, because from the analysis in step 2, if an interval can make two pit functions merge to another pit function, a smaller interval also can. Suppose that $n/2$ is still even. We can pair the functions consecutively, sum respectively, and obtain $n/4$ pit functions denoted by $f_i^{(3)}$'s with centers $\mathbf{c}_i^{(3)}$'s. Denote the maximum interval between $\mathbf{c}_i^{(3)}$ and $\mathbf{c}_{i+1}^{(3)}$ required to make $f_i^{(3)} + f_{i+1}^{(3)}$ to be a pit function by $\theta^{(3)}$. If $n/2^2, n/2^3, n/2^4, \ldots$ are all even, then repeating this process will result in a pit function located at the midpoint of L, which is the sum of all $f_i^{(1)}$'s, by letting $\delta = \min\{\theta^{(1)}, \theta^{(2)}/2, \theta^{(3)}/4, \ldots\}$. See Figure 7C (left).

If at any step k (started from 1), $n/2^{(k-1)}$ is odd, we can take away the middle function and repeat the process as described above for the left $n/2^{(k-1)} - 1$ functions (see Figure 7C, right). The difference is that $\theta^{(k)}$ required to make $f_i^{(k)} + f_{i+1}^{(k)}$ to be a pit function may vary across i, and the smallest one should be recorded for calculating the required δ. In addition, δ is no longer equal to $\theta^{(k)}/2^{(k-1)}$; nevertheless, it can be determined from θ^k. Repeat this process until only one function has resulted. By taking δ as the minimum among the values calculated in all steps, the sum of all $f_i^{(1)}$'s will result in a pit function centered at the midpoint of L. Note that n can be very large and δ can be very small. The proof is completed.

Acknowledgments

We are grateful to Pia Rotshtein at University of Birmingham for providing the visual stimuli used in Figures 5 and 6 and Zhaoping Li at University College London for useful comments on the manuscript for this letter. The work was supported by the National Natural Science Foundation of China, grants 60805023, 60621062, and 60605003; National Key Foundation R&D Project, grants 2003CB317007, 2004CB318108, and 2007CB311003; China Postdoctoral Science Foundation, grants 20080430032 and 200801072; and Basic Research Foundation of Tsinghua National Laboratory for Information Science and Technology.

References

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge: Cambridge University Press.

Ben-Yishai, R., Bar-Or, R. L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. USA, 92*, 3844–3848.

Blumenfeld, B., Preminger, S., Sagi, D., & Tsodyks, M. (2006). Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron, 52*, 383–394.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C. E., Riesenhuber, M., & Poggio, T. (2007). A model of V4 shape selectivity and invariance. *J. Neurophysiol., 98*, 1733–1750.

Casali, D., Costantini, G., Perfetti, R., & Ricci, E. (2006). Associative memory design using support vector machines. *IEEE Trans. Neural Netw., 17*, 1165–1174.

Chartier, S., & Proulx, R. (2005). NDRAM: Nonlinear dynamic recurrent associative memory for learning bipolar and nonbipolar correlated patterns. *IEEE Trans. Neural Netw., 16*, 1393–1400.

Chua, L. O., & Yang, L. (1988). Cellular neural networks: Theory. *IEEE Trans. Circuits Syst., 35*, 1257–1272.

Cohen, M. A., & Grossberg, S. (1983). Absolute stability and global pattern formation and parallel memory storage by competitive neural networks. *IEEE Trans. Syst. Man. Cybern., 13*, 815–826.

Fang, F., & He, S. (2005). Viewer-centered object representation in the human visual system revealed by viewpoint aftereffects. *Neuron, 45*, 793–800.

Fox, C. J., & Barton, J. J. S. (2006). What is adapted in face adaptation? The neural representations of expression in the human visual system. *Brain Res., 1127*, 80–89.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci., 23*, 5235–5246.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2006). Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex, *Cereb. Cortex, 16*, 1631–1644.

Fuster, J. M. (1995). *Memory in the cerebral cortex*. Cambridge, MA: MIT Press.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nat. Neurosci., 2*, 568–573.

Gilaie-Dotan, S., & Malach, R. (2007). Sub-exemplar shape tuning in human face-related areas. *Cereb. Cortex, 17*, 325–338.

Goldmanrakic, P. S. (1996). Regional and cellular fractionation of working memory. *Proc. Natl. Acad. Sci. USA, 93*, 13473–13480.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational ability. *Proc. Natl. Acad. Sci., USA, 79*, 2554–2558.

Hopfield, J. J., & Tank, D. W. (1986). Computing with neural circuits: A model. *Science, 233*, 625–633.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310*, 863–866.

Jiang, X., Rosen, E., Zeffiro, T., VanMeter, J., Blanz, V., & Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. *Neuron, 50*, 159–172.

Knoblich, U., Bouvrie, J., & Poggio, T. (2007). *Biophysical models of neural computation: Max and tuning circuits* (Tech. Rep. CBCL). Cambridge, MA: MIT.

Available online at http://cbcl.mit.edu/projects/cbcl/publications/ps/CBCL-Paper-KBP-2007.pdf.

Kouh, M., & Poggio, T. (2008). A canonical neural circuit for cortical nonlinear operations. *Neural Comput., 20*, 1427–1451.

Lansner, A. (2009). Associative memory models: From the cell-assembly theory to biophysically detailed cortex simulations. *Trends Neurosci., 32*(3), 178–186.

Leutgeb, J., Leutgeb, S., Treves, A., Meyer, R., Barnes, C., McNaughton, B., et al. (2005). Progressive transformation of hippocampal neuronal representations in "morphed" environments. *Neuron, 48*, 345–358.

Leutgeb, S., Leutgeb, J. K., Moser, M. B., & Moser, E. I. (2005). Place cells, spatial maps and the population code for memory. *Curr. Opin. Neurobiol., 15*, 738–746.

Lin, S. Y., Huang, R. J., & Chiueh, T. D. (1998). A tunable Gaussian/square function computation circuit for analog neural networks. *IEEE Trans. Circuits Syst. II, 45*, 441–446.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol., 5*, 552–563.

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., & Moser, M. B. (2000). Path integration and the neural basis of the "cognitive map." *Nat. Rev. Neurosci., 7*, 663–678.

Mel, B. W. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput. 9*, 777–804.

Menghini, F., van Rijsbergen, N., & Treves, A. (2007). Modelling adaptation after-effects in associative memory. *Neurocomputing*, *70*, 2000–2004.

Papp, G., Witter, M. P., & Treves, A. (2007). The CA3 network as a memory store for spatial representations. *Learn. Mem., 14*, 732–744.

Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks, *Neural Comput., 3*, 246–257.

Peng, S. Y., Hasler, P. E., & Anderson, D. V. (2007). An analog programmable multi-dimensional radial basis function based classifier. *IEEE Trans. Circuits Syst., I, 54*, 2148–2158.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature, 343*, 263–266.

Preminger, S., Sagi, D., & Tsodyks, M. (2007). The effects of perceptual history on memory of visual objects, *Vision Res., 47*, 965–973.

Raiguel, S., Vogels, R., Mysore, S. G., & Orban, G. A. (2006). Learning to see the difference specifically alters the most informative V4 neurons. *J. Neurosci., 26*, 6589–6602.

Rainer, G., & Miller, E. K. (2000). Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron, 27*, 179–189.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci., 2*, 1019–1025.

Rotshtein, P., Henson, R. N. A., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nat. Neurosci., 8*, 107–113.

Samsonovich, A., & McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *J. Neurosci., 17*, 5900–5920.

Schoups, A., Vogels, R., Qian, N., & Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature, 412*, 549–553.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. USA, 104*, 6424–6429.

Serre, T., Wolf, T., Bileschi, T., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell., 29*, 411–426.

Seung, H. S., Lee, D. D., Reis, B. Y., & Tank, D. W. (2000). Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron, 26*, 259–271.

Sharp, P. E., Blair, H. T., & Cho, J. W. (2001). The anatomical and computational basis of the rat head-direction cell signal. *Trends Neurosci., 24*, 289–294.

Stringer, S. M., Rolls, E. T., Trappenberg, T. P., & de Araujo, I. E. T. (2002). Self-organizing continuous attractor networks and path integration: Two-dimensional models of place cells. *Network: Comput. Neural Syst., 13*, 429–446.

Wang, G., Tanifuji, M., & Tanaka, K. (1998). Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neurosci. Res., 32*, 33–46.

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*, 557–561.

Wills, T. J., Lever, C., Cacucci, F., Burgess, N., & O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science, 308*, 873–876.

Winkler, C., & Rhodes, G. (2005). Perceptual adaptation affects attractiveness of female bodies. *Br. J. Psychol., 96*, 141–154.

Zeng, Z., & Wang, J. (2007). Analysis and design of associative memories based on recurrent neural networks with linear saturation activation functions and time-varying delays. *Neural Comput., 19*, 2149–2182.

Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *J. Neurosci., 16*, 2112–2126.

Zoccolan, D., Kouh, M., Poggio, T., & DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J. Neurosci., 27*, 12292–12307.

Received February 7, 2009; accepted September 14, 2009.