Advanced Review

Predictive coding

Yanping Huang and Rajesh P. N. Rao*

Predictive coding is a unifying framework for understanding redundancy reduction and efficient coding in the nervous system. By transmitting only the unpredicted portions of an incoming sensory signal, predictive coding allows the nervous system to reduce redundancy and make full use of the limited dynamic range of neurons. Starting with the hypothesis of efficient coding as a *design principle* in the sensory system, predictive coding provides a functional explanation for a range of neural responses and many aspects of brain organization. The lateral and temporal antagonism in receptive fields in the retina and lateral geniculate nucleus occur naturally as a consequence of predictive coding of natural images. In the higher visual system, predictive coding provides an explanation for oriented receptive fields and contextual effects as well as the hierarchical reciprocally connected organization of the cortex. Predictive coding has also been found to be consistent with a variety of neurophysiological and psychophysical data obtained from different areas of the brain. *WIREs Cogn Sci* 2011 DOI: 10.1002/wcs.142

INTRODUCTION

Natural signals are highly redundant. This redundancy arises from the tendency toward spatial and temporal uniformity in these signals. For example, neighboring pixel intensities in natural images tend to be positively correlated because natural shapes extend over finite spatial regions; similarly, pixel intensities tend to be correlated over time because objects persist in time.[1–3] A direct representation of the raw image by the activity of an array of sensory receptors would thus be very inefficient. It has long been suggested based on the information theoretic considerations[4–6] that the role of early sensory processing is to reduce redundancy and recode the sensory input into an efficient form. One important model for achieving this goal is predictive coding.[7] Predictive coding postulates that neural networks learn the statistical regularities inherent in the natural world and reduce redundancy by removing the predictable components of the input, transmitting only what is not predictable (the residual errors in prediction).

Predictive coding provides a functional explanation for center–surround response properties and biphasic temporal antagonism of cells in the retina[7–10] and lateral geniculate nucleus (LGN).[11,12] In the predictive coding model, neural circuits in the retina/LGN actively predict the value of local intensity from a linear weighted sum of nearby values in space or preceding input values in time. Cells in these circuits convey not the raw image intensity, but the difference between the predicted value and the actual intensity. This decorrelates (or whitens)[7,9] the input signals by flattening the temporal and spatial power spectra, thereby reducing output redundancy. The resulting difference signal has a much smaller dynamic range [when the input signal-to-noise ratio (SNR) is high], and is therefore more economical for transmission through a visual pathway that has limited dynamic range.[6,9,13]

Neurons in the primary visual cortex (V1) respond to bars and edges at preferred orientations[14–16] whereas neurons in areas V2 and V4 respond to more complex shapes and contour features.[17,18] Neurons in medial superior temporal area (MST) respond to visual motion.[19,20] These response selectivities can be understood in terms of hierarchical predictive coding of natural inputs. For example, motivated by the fact that the visual system is hierarchically organized with reciprocal connections between cortical areas, Rao and Ballard[21] proposed a hierarchical neural network in which top-down feedback connections from higher order visual cortical areas carry predictions of lower-level neural activities, while the bottom-up connections convey the residual errors in prediction.[22,23] After training a model

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: rao@cs.washington.edu

Department of Computer Science and Engineering, University of Washington, Washington, DC, USA

DOI: 10.1002/wcs.142

network on image patches taken from natural scenes, they found that model neurons developed receptive field properties similar to those in V1, including oriented receptive fields, end-stopping, and other contextual effects. Jehee et al.[24] proposed a predictive coding model that captured the visual selectivity of MST neurons to optic flow when exposed to visual motion resulting from translation movements in space.

In this review article, we formally introduce predictive coding within an efficient coding framework and illustrate the concept using the examples of predictive coding in space and time. We then review how predictive coding has been used to understand the responses of neurons in various regions of the nervous system. We conclude with a brief summary and discussion of other experimental support for predictive coding in the brain.

PREDICTIVE CODING: MODEL AND TWO ILLUSTRATIVE EXAMPLES

General Framework

The underlying assumption of predictive coding is that the visual system tries to learn an internal model of the external environment and uses this model to actively predict incoming signals.[21,25,26] This can be formalized using a generative model $P(\boldsymbol{I}|\boldsymbol{r})$, which is the probability of an image $\boldsymbol{I}$ given a set of hidden internal model parameters $\boldsymbol{r}$ (represented by firing rates in a network of neurons). For a given input image $\boldsymbol{I}$, the neural system is assumed to select the parameters $\boldsymbol{r}$ that maximize the posterior probability $P(\boldsymbol{r}|\boldsymbol{I}) = P(\boldsymbol{I}|\boldsymbol{r})P(\boldsymbol{r})/P(\boldsymbol{I})$ obtained using Bayes theorem, where $P(\boldsymbol{I})$ is a normalizing constant. Using an information theoretic point of view, let H be the overall description length (or information entropy), i.e., the sum of coding length $H_1 = -\log P(\boldsymbol{I}|\boldsymbol{r})$ of predicting $\boldsymbol{I}$ using $\boldsymbol{r}$, and the length $H_2 = -\log P(\boldsymbol{r})$ of the parameters $\boldsymbol{r}$ themselves. Minimizing the total description length $H = -\log P(\boldsymbol{I}|\boldsymbol{r}) - \log P(\boldsymbol{r})$ is thus equivalent to maximizing the posterior probability of the parameters under the predictive coding assumption. Therefore, the so-called minimum description length (MDL) framework[27,28] can be seen to be formally equivalent to Bayesian maximum a posteriori (MAP) estimation.

Other coding schemes such as sparse coding[16,29] and independent component analysis (ICA)[30] can also be understood under the above-mentioned Bayesian/MDL framework by imposing appropriate constraints on $P(\boldsymbol{I}|\boldsymbol{r})$ and $P(\boldsymbol{r})$. In sparse coding, for example the dimensionality of $\boldsymbol{r}$ is typically chosen to be larger than that of $\boldsymbol{I}$, i.e., the input is projected into a higher-dimensional space, and $P(\boldsymbol{r})$ is chosen to encourage sparseness in $\boldsymbol{r}$, i.e., most elements of $\boldsymbol{r}$ are zero (see Supporting Information for more details). In ICA, the goal is to make the elements of $\boldsymbol{r}$ as statistically independent as possible; in the case of Bell and Sejnowski's ICA algorithm,[30] this is achieved by assuming that $\boldsymbol{r}$ has the same dimensionality as $\boldsymbol{I}$ and minimizing the mutual information between the components of $\boldsymbol{r}$ (see Ref 31 for further details).

Predictive Coding in Space

Natural images are usually composed of many finite areas with relatively uniform intensity values. As a result, neighboring pixel intensities in most natural images tend to be spatially correlated over short distances. Figure 1(b) (blue curve) illustrates this using the spatial autocorrelation function measured from a natural scene (Figure 1(a)). The intensity at a particular pixel can thus be predicted based on the intensities surrounding it, allowing the input to be efficiently coded as the residual error between the actual intensity and the prediction based on the surrounding pixels. Suppose we would like to predict the pixel intensity x_0 based on the neighboring intensities $x_{-N}, \ldots, x_{-1}, x_1, \ldots, x_N$ at $2N$ different locations $-N, \ldots, -1, 1, \ldots, N$. Then, the statistically optimal linear prediction of x_0 is given by a weighted average of the $2N$ neighboring samples, i.e.,

$$\hat{x}_0 = \sum_i w_i x_i. \tag{1}$$

To compute the optimal weights w_i, suppose there are k such pixels x_0, we can stack these pixels as a $k \times 1$ vector $\mathbf{A}$. For each of these pixels, the neighboring intensities $x_{-N}, \ldots, x_{-1}, x_1, \ldots, x_N$ are used as the rows of a $k \times 2N$ matrix $\boldsymbol{B}$. We use the vector $\mathbf{W}$ to represent the weights w_i. Then, the weights $\mathbf{W}$ can be obtained by minimizing the total prediction error over all pixels:

$$E = \|\mathbf{A} - B\mathbf{W}\|^2/2, \tag{2}$$

where $\|\mathbf{Y}\|$ denotes the magnitude of vector $\mathbf{Y}$. It can be shown that minimizing the error E is equivalent to finding the MDL representation of the raw input image assuming that the intrinsic noise is Gaussian (Supporting Information). Taking the derivative of E with respect to $\mathbf{W}$, we have:

$$\frac{\partial E}{\partial \mathbf{W}} = -B^T(\mathbf{A} - B\mathbf{W}) = 0$$
$$B^T\mathbf{A} = B^TB\mathbf{W}. \tag{3}$$

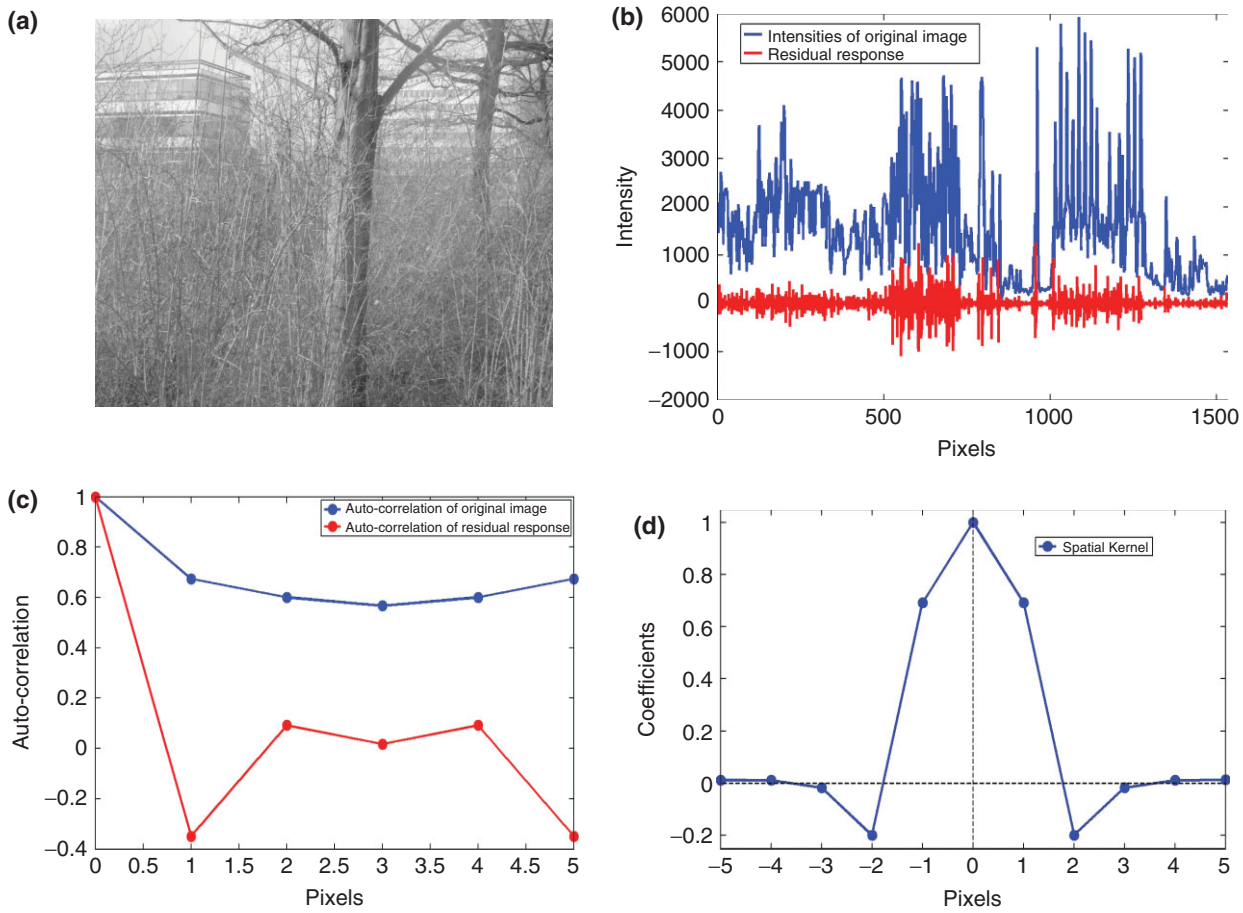

FIGURE 1 | Predictive coding in space. (a) An example natural image (Reprinted with permission from Ref 32. Copyright 1998 Royal Society Publishing), shown here in logarithmic scale for better visualization of pixel values. (b) Blue curve: pixel intensities measured along a horizontal line of the image in (a). Red curve: the residual error between the actual intensity and predicted intensity from neighboring pixels. (c) Blue curve: autocorrelation function of intensities shown in the blue curve in (b). Red curve: autocorrelation function of the residual error shown in the red curve in (b). (d) Optimal spatial weighting coefficients **W** calculated from Eq. 2 for this example.

Solving for the optimal linear weights, we obtain:

$$\mathbf{W} = (B^T B)^{-1} B^T \mathbf{A}. \tag{4}$$

Figure 1(d) shows the optimal linear weights **W** derived using this linear prediction model. By subtracting the linear prediction from the actual pixel intensity, the residual response $r = x_0 - \hat{x}_0$ (Figure 1(b), red curve) decorrelates the original image (compare Figure 1(c), red curve to the blue curve), thereby reducing redundancy.

Predictive Coding in Time

Predictive coding can also be applied to the time domain. Figure 2(a) shows a time-varying intensity profile measured from a fixed pixel in a natural movie. The corresponding autocorrelation function is shown in the blue curve of Figure 2(b). Given that temporally close intensities tend to be positively correlated, one can predict the current intensity as a weighted linear combination of preceding intensities using an analysis similar to the one used above for spatial predictive coding, except for one notable difference: spatial predictive coding is based on intensities from all surrounding neighbors, whereas temporal predictive coding is causal and based only on the past history of intensities. Figure 2(c) shows the optimal temporal weighting function derived from the temporal version of Eq. (1) with $j = -1, \ldots, -N$ over time. As expected, the dynamic range and the autocorrelation of the residual response are dramatically reduced after predictive coding (shown as red curves in Figure 2(a) and (b)). More sophisticated models of predictive coding in time rely on learning dynamic spatiotemporal models of the inputs and perform some form of optimal statistical filtering such as Kalman filtering of inputs.[33]

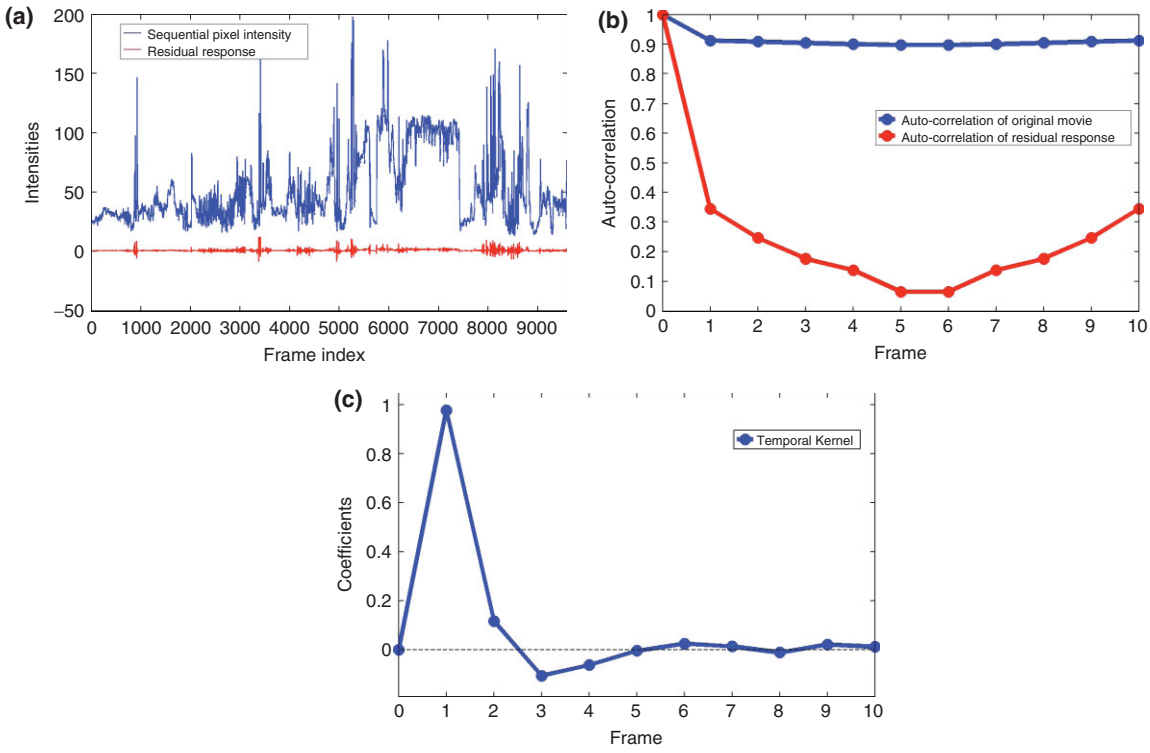

FIGURE 2 | Predictive coding in time. (a) Blue curve: time-varying intensities measured at a fixed pixel of a natural video (Reprinted with permission from Ref 32. Copyright 1998 Royal Society Publishing). The sampling frequency is 50 frames/second. Red curve: the residual error between the actual intensity and predicted intensity from past time steps. (b) Blue curve: autocorrelation function of intensities shown in the blue curve in (a). Red curve: autocorrelation function of the residual error shown in the red curve in (a). (c) Optimal temporal weighting coefficients for this example.

PREDICTIVE CODING IN THE NERVOUS SYSTEM

Predictive Coding in the Retina

In this section, we discuss how predictive coding provides an explanation for both the spatial and temporal receptive fields found in the retina.[8–10,7,13] The underlying assumption is that the retina tries to build an efficient (e.g., MDL) representation of the visual scene.

For the case of spatial receptive fields, consider an array of neurons in the retina, each of which receives an excitatory input from the center and an inhibitory input from the surround. The response at the center of the receptive field is estimated from a linear weighted average of surrounding intensity values.[1–3] By subtracting this prediction from the actual intensity via lateral inhibition, the range of the neural response can be minimized as demonstrated in the section on *Predictive Coding in Space*. The shape of the weighting function that minimizes the error between the estimated intensity value and its actual value was derived in the section on *Predictive Coding in Space* and closely resembles the classical center–surround receptive fields of retinal ganglion cells (Figure 3(a); compare with Figure 1(d)).

Srinivasan et al.[7] showed that the weighting function also depends on the SNR of the input signal. When SNR is low, the intensity value at the center can no longer be estimated reliably from its nearest neighbors. Instead, better estimation can be achieved by recruiting larger groups of surrounding points in order to cancel out the statistically independent noise. Thus, one expects the surround of the spatial receptive field to become weaker and more diffuse as SNR decreases (Figure 3(b)). Remarkably, this predicted phenomenon was observed by Srinivasan et al. in first-order interneurons in the compound eye of the fly (Figure 3(c)). Similarly, in the temporal domain, as SNR decreases, one expects more pixels from the past to be used in generating a prediction (Figure 3(d)). This was also observed in the fly eye (Figure 3(e)).

An efficient visual encoder should learn the statistical regularities of the input image and adapt its encoding strategy accordingly.[35,36] The center–surround antagonism can be viewed as adaptation to spatial image correlations. However, animals may also encounter environments where neighboring image pixels do not share similar intensities. Under these conditions, the intensity at the center can no longer be predicted using a simple weighted average of the

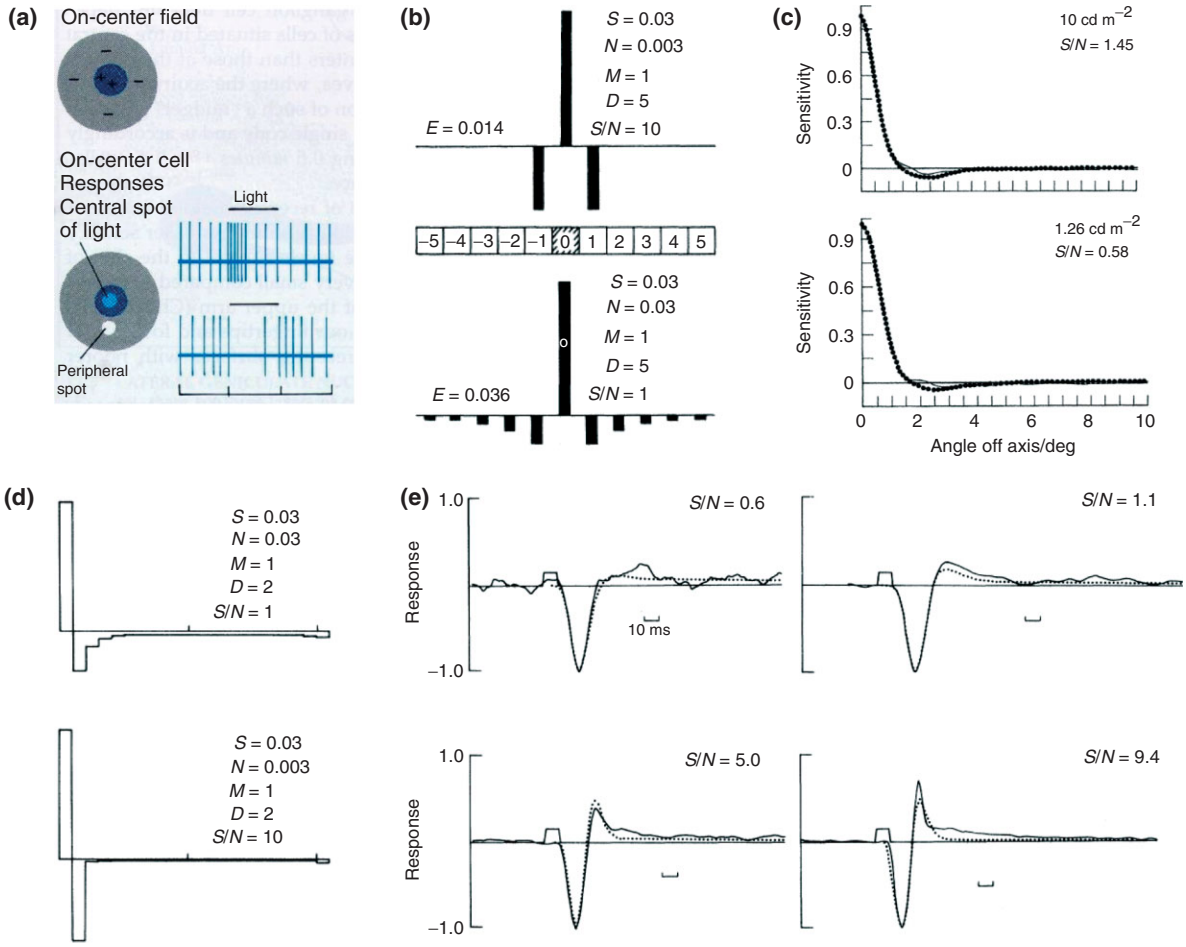

FIGURE 3 | Spatial and temporal predictive coding in the retina. (a) Classic center–surround receptive field found in the retina. (Reprinted with permission from Ref 34. Copyright 1992 Sinauer Assoicates) Compare to the center–surround weighting profile in Figure 1(d). (b) Shape of the receptive field depends on the signal-to-noise ratio (SNR). (Upper) Higher SNR; (Lower) Lower SNR. (c) Comparison of theoretically optimal (dotted curves) with experimentally measured (solid curves) receptive fields of large monopolar cells in the compound eye of the fly. (Upper) SNR = 1.45 at luminance 10 cd m^{-2}. (Lower) SNR = 0.58 at luminance 1.26 cd m^2. Note that the receptive field is more diffuse for the lower SNR as predicted in (b), although the effect is not as pronounced. (d) Effect of SNR on the temporal weight profile. (Upper) Low SNR; (Lower) High SNR. (e) Temporal receptive fields of large monopolar cells in the fly's eye for increasing SNR (top left to bottom right) (Reprinted with permission from Ref 7). Note that the receptive field is inverted compared to (d). As predicted by the theoretical result in (d), the temporal receptive field sharpens as SNR increases.

surround. Hosoya et al.[13] proposed that adaptation to different visual scenes with varying correlational structures should lead to marked changes in predictive coding in retinal ganglion cells. In their experiments, they exposed ganglion cells in the salamander retina to two types of stimuli: a flickering uniform field with perfect positive correlation between all image points (environment A) and a flickering checkerboard pattern with perfect negative correlation between two sets of image regions (environment B) (Figure 4(a) and (b)). They found that after adaptation to environment B (negatively correlated stimuli), the receptive field profile of a typical ganglion cell flattened and the cell became equally sensitive to the two checkerboard regions (Figure 4(d)). The result of this adaptation was that the cell became less sensitive to checkerboard stimuli by a factor of about 0.57 and became *more sensitive to uniform stimuli* by a factor of 1.4 (recall that under normal circumstances, uniform stimuli elicit little or no response from a ganglion cell). These results indicate that retinal ganglion cells rapidly adapt to become less sensitive to stimuli they have been exposed to; in the process, they become more responsive to other stimuli (i.e., novel stimuli), as expected from predictive coding theory.

Besides spatial correlation, there is also a high degree of chromatic correlation among pixels in natural images. In species with color vision, there are several types of retinal photoreceptors that are sensitive to different wavelengths of light. For example,

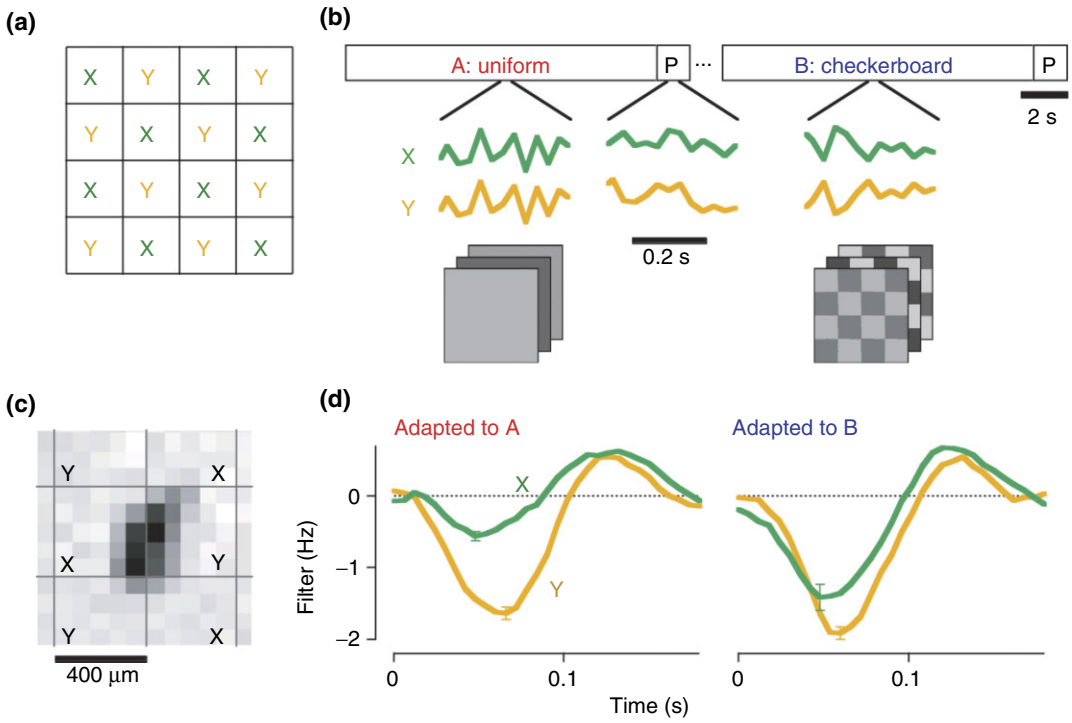

FIGURE 4 | Adaptation to spatial image statistics. (a) Stimulus with two checkerboard regions X and Y. (b) Time-varying stimuli used to test adaptation in retinal ganglion cells. Environment A has perfect positive correlation between all image points, whereas environment B has perfect negative correlation between X and Y regions. An uncorrelated probe stimulus P lasting 1.5 seconds was used to test the cell's spatiotemporal receptive field after adapting to environment A or B for 13.5 seconds. (c) Spatial receptive field of a ganglion cell from salamander retina. (d) Sensitivity of the ganglion cell in response to P for stimulus regions X and Y after adaptation to environment A (left) or B (right) (Reprinted with permission from Ref 13. Copyright 2005 Nature Publishing Group).

humans have three types of photoreceptors, known as S, M, and L for short, medium, and long wavelengths, respectively. The responses of those photoreceptors are often correlated because their spectral sensitivities overlap. Thus the M-cone response can be used to predict the L-cone response, and the L- and M-cone responses can together be used to predict the S-cone response. Correspondingly, one sees color-opponent receptive fields in the retina. For example, one type of color-opponent retinal ganglion cell receives excitatory input from L-cone ('red') receptors in the center and inhibitory input from M-cone ('green') receptors in the surround. Thus, the color-opponent (*red–green*) and *blue–*(*red* + *green*) channels in the retina might reflect predictive coding in the chromatic domain, similar to the predictive coding observed in the spatial and temporal domains.[10]

Predictive Coding in the LGN

Dong, Dan, Atick, and Reid[11,12] have suggested that the LGN performs predictive coding by temporally whitening (decorrelating) the signal from the retina.[9,35,73] As time-varying natural image sequences (or 'movies') exhibit strong positive inter-frame correlations,[3] the intensity value of a particular pixel at time t can be predicted from the weighted sum of intensity values at preceding time points (see the section on *Predictive Coding in Time*):

$$O(t) = \int K(t,t')S(t')dt' = K^*S, \qquad (5)$$

where the temporal kernel $K(t,t')$ is the temporal receptive field of the LGN neuron and $S(t)$ is the input stimulus. The output response $O(t)$ is a linear sum of preceding inputs $S(.)$ with weighting function $K(t,.)$. This is similar to the formulation in Eq. (1) except that the summation is replaced by an integral.

Unlike the section *Predictive Coding in Space* where we derived the optimal linear filter by minimizing the error in the spatial or temporal domain, here we follow Dong and Atick and illustrate an alternative approach to deriving the optimal filter in the frequency domain using the equivalent goal of decorrelation. Suppose the natural movie has a temporal correlation matrix R:

$$R = R(t,t') = \langle S(t)S(t')\rangle \qquad (6)$$

where the brackets denote averaging over different pairs of frames. We would like the output to be decorrelated:

$$\langle O(t)O(t')\rangle = \delta(t,t'), \tag{7}$$

where $\delta(t,t')$ is the Kronecker δ function. We can then solve for the optimal receptive field K by substituting Eq. 5 into Eq. 7 and converting to the frequency domain:

$$K(\omega)R(\omega)K^*(\omega) = 1$$

$$|K(\omega)| = \frac{1}{\sqrt{R(\omega)}}, \tag{8}$$

where ω is the temporal frequency. Here, we assume R and K are time-invariant, i.e., $K(t,t') = K(t-t')$ and $R(t,t') = R(t-t')$.

The receptive field or filter K decorrelates both signal and noise; since $R(\omega) = S^2(\omega) + \eta^2$, where the first term is the signal power and the second term is the noise power. At high temporal frequency, the noise is significant and the filter will amplify noise relative to the signal. To code efficiently in the presence of noise, Dong and Atick[3] suggested a filter that suppresses noise at high frequencies and decorrelates when the signal-to-noise ratio is high. This new filter is the product of the decorrelation filter in Eq. 8 and the optimal noise suppressing filter (also called a Wiener filter) $M = S^2/(S^2+\eta^2) = (R-\eta^2)/R$:

$$|K(\omega)| = \frac{1}{\sqrt{R(\omega)}}\frac{R(\omega)-\eta^2}{R(\omega)}. \tag{9}$$

Dong and Atick[3] measured the power spectrum of natural movies and, assuming temporal white noise, derived $R(\omega)$ as:

$$R(\omega) \approx \frac{1}{\omega^2} + \frac{1}{\omega_c^2}, \tag{10}$$

where ω_c is the noise frequency. Substituting (10) into (9), they obtained the predicted optimal temporal filter of LGN cells:

$$|K(\omega)| = \frac{\omega}{(1+\omega^2/\omega_c^2)^{\frac{3}{2}}}. \tag{11}$$

As shown in Figure 5(a), this predicted optimal filter compares remarkably well with physiological data from the LGN.[37] Figure 5(b) shows the temporal receptive field derived by Dong and Atick from the filter in Figure 5(a) (after placing appropriate constraints such as causality). This receptive field is similar to the one we obtained for our example in Figure 2 and illustrates how a weighted average of past pixel values (negative part of the curve) is subtracted from a weighted average of recent pixel values (positive part of the curve), consistent with the general framework of predictive coding.

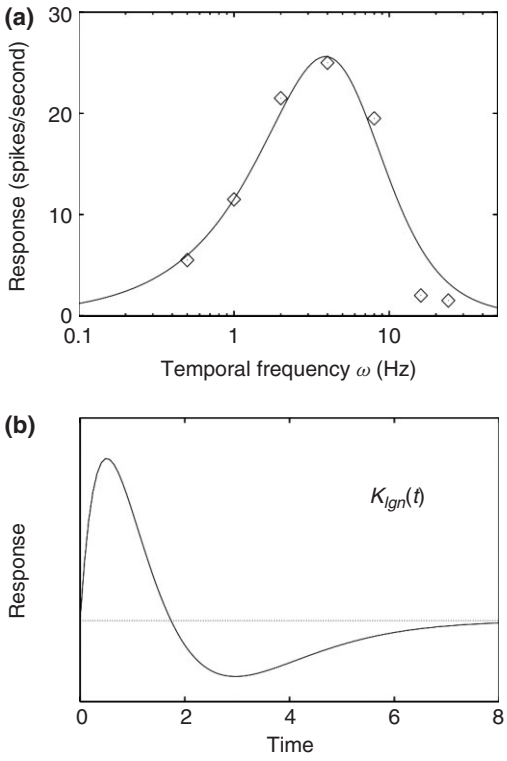

FIGURE 5 | Temporal predictive coding in the lateral geniculate nucleus (LGN). (a) Comparison between theoretically predicted temporal tuning curve (solid curve) and experimental data (points) in the LGN.[37] The predicted curve is generated from Eq.11 with $\omega_c = 5.5$ Hz. (b) Temporal receptive field derived from (a) (Reprinted with permission from Ref 12. Copyright 1995 Informa PLC). Note the similarity to the temporal predictive coding filter in Figure 2(c).

The above analysis assumes that spatial and temporal decorrelations are accomplished separately: the retina is assumed to reduce most of the spatial correlations, whereas the LGN removes much of the temporal redundancy in the input image. Such a model, originally suggested by Dong and Atick,[12] is consistent with the findings that the temporal bandpass filtering done by the retina is essentially flat (i.e., not much temporal decorrelation) and the spatial receptive fields of LGN cells are very similar to those of retina cells (i.e., not much additional spatial processing at the LGN).

Predictive Coding in the Visual Cortex

Predictive coding has also been used to provide explanations for important receptive field properties in the

visual cortex such as oriented receptive fields and contextual effects. It has been used to ascribe an active computational role to the reciprocal connections between the different hierarchically organized cortical areas.

In the predictive coding view of the visual cortex as proposed by Rao and Ballard,[21] the cortex is modeled as a hierarchical network, with higher level units attempting to predict the responses of units in the next lower level via feedback connections (Figure 6(a), lower arrows). The inspiration for the model comes from neurophysiological evidence suggesting that the feedback connections from high level areas play an active role in shaping the tuning properties of lower level areas.[38–40] In the model, lower level units send back the discrepancies between the top-down predictions and the actual activity through feedforward connections[22,23] (Figure 6(a), upper arrows). These discrepancies or residual error signals are then used by the predictive estimator (PE) at each level to correct the higher level estimates of the input signal and generate the next prediction (Figure 6(b)). Lower levels have smaller spatial (and possibly temporal) receptive fields, whereas higher levels have larger receptive fields because a higher level unit predicts and estimates signal properties at a larger scale by combining the responses of several lower level units (three in the example shown in Figure 6(c)). Thus, the effective receptive field size of units at the highest level could span the entire input image.

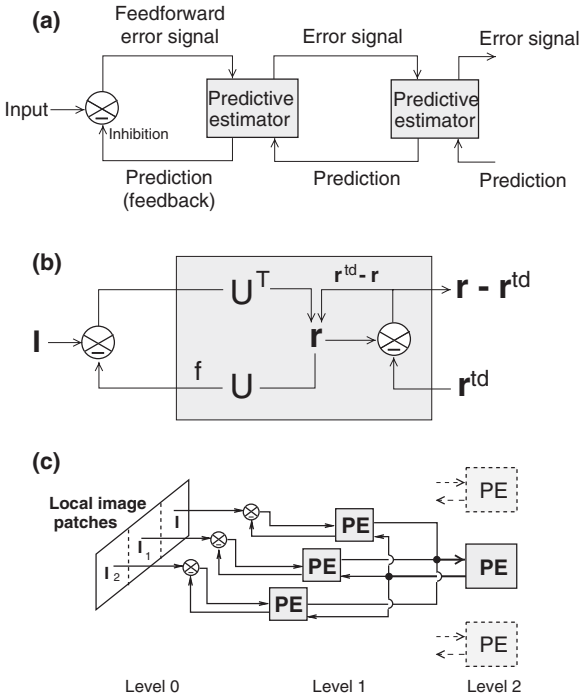

FIGURE 6 | Hierarchical predictive coding model of the visual cortex. (a) General architecture of the hierarchical predictive coding model. Higher level units attempt to predict the responses of units in the next lower level via feedback connections. Lower level units sent back the error between the higher level predictions and the actual activity through feedforward connections. This residual error signal is then used by the predictive estimator (PE) at each level to correct the higher level estimations of input signal. (b) Components of a PE unit. Each unit consists of several kinds of neurons: feedforward neurons encoding the synaptic weights U^T, predictor-estimator neurons maintaining the current estimate $\boldsymbol{r}$ of the input signal, feedback neurons encoding U and carrying the prediction $f(U\boldsymbol{r})$ to lower level, and error-detecting neurons computing the discrepancy $(\boldsymbol{r} - r^{td})$ between the current prediction $\boldsymbol{r}$ and its top-down prediction $\boldsymbol{r}^{td}$ from a yet higher level. (c) An example of three-level hierarchical network. Three image patches at level 0 are processed by three level 1 PE units. The estimates from these three level 1 units are input to a single level 2 unit. This convergence effectively increases the receptive field size of neurons as one ascends the hierarchy. (Reprinted with permission from Ref 21. Copyright 1999 Nature Publishing Group).

By assuming a probabilistic hierarchical generative model for images (Supporting Information), Rao and Ballard derived the dynamics of the hierarchical network and learning rules for the synaptic connections between two levels, allowing them to model interactions in the LGN–V1–V2 feedback loop[21] (similar models have since been suggested for the LGN–V1 circuit[41,24] and the middle temporal (MT)–MST circuit[24]). Specifically, the activity in a lower area, for example V1, is represented by a vector of firing rates $\boldsymbol{r}$. The output of a higher area, for example V2, conveys a 'top-down' prediction $\boldsymbol{r}^{td}$ of the expected neural responses in the lower area. As shown in Figure 6(b), the feedforward input from the lower to the higher area carries the residual error $\boldsymbol{r} - \boldsymbol{r}^{td}$ which is used by higher area neurons to correct its local estimate and generate a new prediction $\boldsymbol{r}^{td}$ (see Supporting Information for details). The bottom-up error $U^T(\boldsymbol{I} - f(U\boldsymbol{r}))$ and the top-down prediction error $(\boldsymbol{r}^{td} - \boldsymbol{r})$ are weighted by the inverse of their corresponding noise variances: the larger the noise variance in the source of information, the smaller the weight given to that source, consistent with the principle of Kalman filtering.[33,42] This dynamics, along with the associated learning rule for the synaptic weights $\boldsymbol{U}$ at each level, can be shown to maximize the posterior probability of the observed input data (which is equivalent to the MDL principle; see Supporting Information). The feedforward weight vectors (rows of $\boldsymbol{U}^T$) can be shown to effectively determine the receptive fields of the feedforward model neurons.[16,29] When trained on natural images, these weight vectors resemble oriented filters or Gabor wavelets[43,16,29] which have been used to model the receptive fields of simple cells in V1, while the higher level synaptic weights represent more complex features[21] (Figure 7).

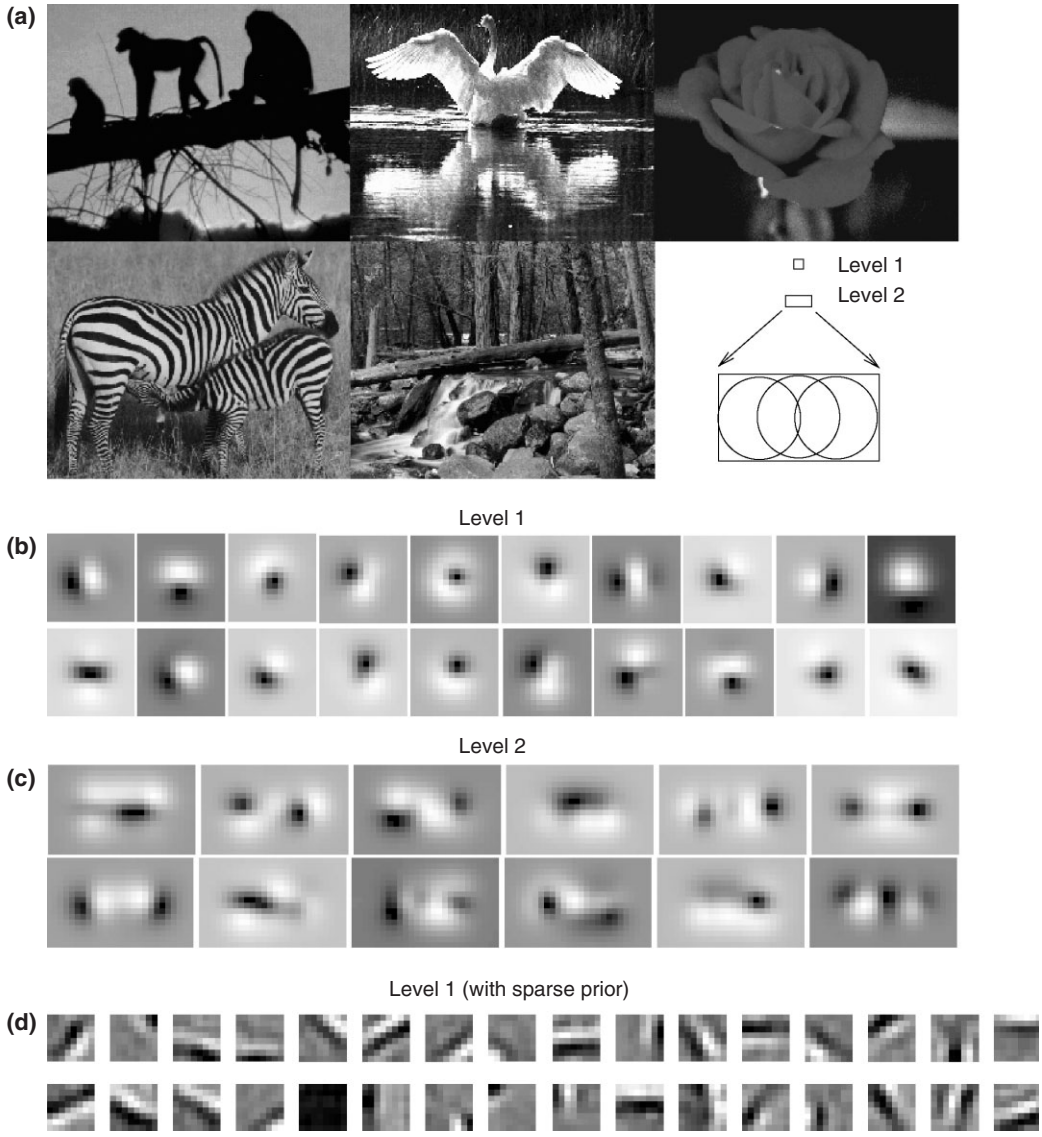

FIGURE 7 | Receptive field properties of feedforward neurons in the hierarchical predictive coding model. (a) Natural images used for training the hierarchical model. Several thousand natural image patches were extracted from these five images. (b) Feedforward synaptic weights of level 1 neurons learned from the natural images in (a) using a Gaussian prior. These synaptic weights determine the receptive field properties of the feedforward neurons. (c) Feedforward synaptic weights of level 2 neurons. These weights resemble various combinations of the synaptic weights in level 1. (d) Localized feedforward synaptic weights (rows in basis matrix U^T) learned by using a sigmoidal nonlinear generative model and a sparse kurtotic prior distribution. Values can be zero (always represented by the same gray level), negative (inhibitory, black regions) and positive (excitatory, bright regions). (Reprinted with permission from Ref 21. Copyright 1999 Nature Publishing Group).

An important attribute of the hierarchical predictive coding model is that it provides a functional explanation for extraclassical receptive field effects (also called contextual effects) in the visual cortex. Such contextual effects have been reported in several cortical areas including V1,[14,44] V2,[45,46] V4,[47] and MT.[48] For many neurons in these areas, when the properties of a stimulus in the center, such as orientation, velocity, or direction of motion, match those in the surrounding regions, the responses are suppressed or eliminated compared to when the same stimulus is shown alone in the center. Figure 8(a) (solid line) shows an example of such an effect, known as 'end-stopping', in a complex cell in layer 2/3 of cat primary visual cortex. A vigorous response is suppressed as the length of an optimally oriented bar grows beyond the classical receptive field.

Rao and Ballard[21] postulated that extraclassical receptive field effects in the visual cortex may result from predictive coding of natural images but

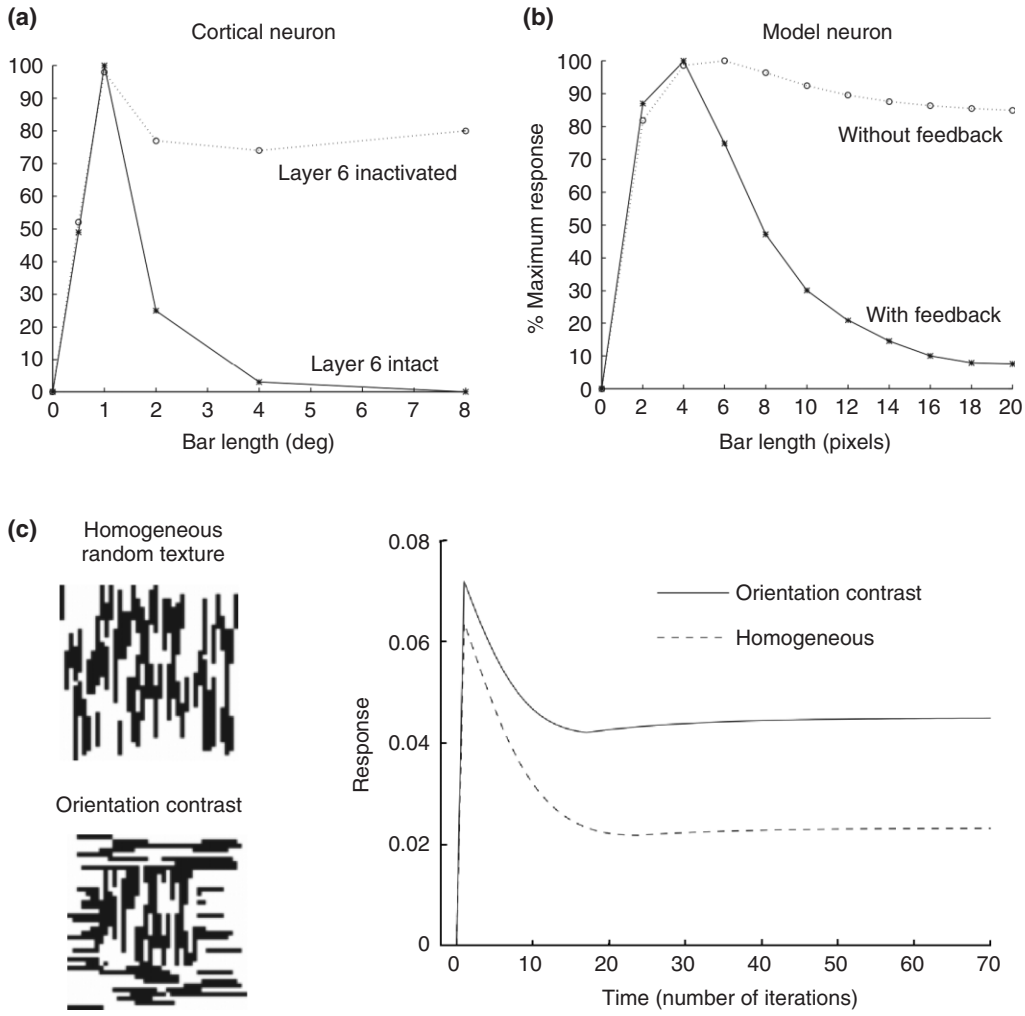

FIGURE 8 | Extraclassical receptive field effects in the hierarchical predictive coding model. (a) End-stopping in a layer 2/3 complex cell in cat striate cortex. Tuning curves are shown for inactivation of layer 6 (dotted curve) and for the control case (solid curve). (b) Tuning curve of lower level model neuron after inactivation of feedback from higher level (dotted curve) and for the control case (solid curve). (c) Extraclassical receptive field effect (contextual modulation). Responses of an error-detecting model neuron for oriented texture stimuli with center and surround are the same (dotted line) versus different (solid line) orientations. (Reprinted with permission from Ref 21. Copyright 1999 Nature Publishing Group).

along more complex dimensions than pixel intensities as is the case in the retina and LGN. In particular, when the stimulus properties in the center match those in the surround, the responses from higher level neurons (which process a larger region) can predict the response of the central neuron, resulting in small residual errors. Although the neurons in layer 2/3 are the ones that send feedforward connections to a higher area, these would correspond to the error-detecting neurons in the model and thus can be expected to show end-stopping and other extraclassical effects.

To test this hypothesis, a hierarchical predictive coding model network was first trained on natural images and then exposed to oriented bars of various lengths. The error-detecting neurons at the lower level displayed end-stopping: their responses were suppressed when the bar extended beyond the classical receptive field (Figure 8(b), solid curve) because the prediction from the higher level was more accurate for the longer bar than the shorter bar. Analogous to the case of spatial predictive coding in the retina, the longer bar provides the necessary context for the network to predict the bar in the center. The feedback predictions from the higher area become progressively more accurate with longer bars, bringing the prediction errors closer to zero. As shown in dotted curve of Figure 8(b), elimination of predictive feedback caused the error-detecting neurons to continue to respond robustly to longer bars, similar to neural responses in the cat visual cortex when layer 6 was inactivated, possibly removing feedback from V2 (Figure 8(a), dotted

curve). The error-detecting neurons in the predictive coding model also exhibited other contextual effects: for example, when presented with an oriented texture stimulus in the center and a texture stimulus of a different orientation in the surround,[49] the error-detecting neurons developed a large positive difference over time (Figure 8(c)), resembling contextual effects observed in V1 neurons.[49] Other V1 response properties, such as cross-orientation suppression and orientation contrast facilitation, can also be explained within the predictive coding framework.[50,51]

Recent imaging studies indicate that activity increases in higher object-processing areas result in concurrent reduction of responses in lower areas such as V1,[52,53] consistent with the predictions of the hierarchical predictive coding model. The hierarchical model, however, does not rule out the possibility that predictive inhibitory feedback may also arise from local recurrent feedback connections[16,54] in a manner similar to the retina. In fact, the dynamics of the network can be rewritten to replace the feedback connections from a higher area with lateral inhibition.[29]

Predictive coding can also be used to model motion processing in the visual cortex. Neurons in area MST, which are tuned to optic flow such as planar, radial, and circular motion, receive inputs from MT area, where neurons code for local visual motion (magnitude and direction).[19,20] Jehee et al.[24] applied the hierarchical predictive coding model to explain receptive field properties in MST. Visual motion inputs extracted from natural image sequences, resembling MT inputs to MST, were used to train the model. After training, the model developed preferred responses to translation and expansion, which are components of optic flow, similar to MST neurons (Figure 9).

The predictive coding models described above assume a strict underlying hierarchy. However, the basic idea can be extended to allow more complicated graph topologies as well. For example, Rao and Ballard[55] suggested a predictive coding model that achieves visual invariance by factoring an input image into two representations, one representing object-specific properties and the other representing spatial transformations. These representations are maintained in two separate networks, an 'object' network and a 'transformation' network. Transformations such as translations of the object are predicted by the transformation network, allowing the object representation to remain stable (see Refs 55–57 for details). The two networks in such a predictive coding model are reminiscent of the dichotomy between the dorsal and ventral pathways in the primate visual cortex.

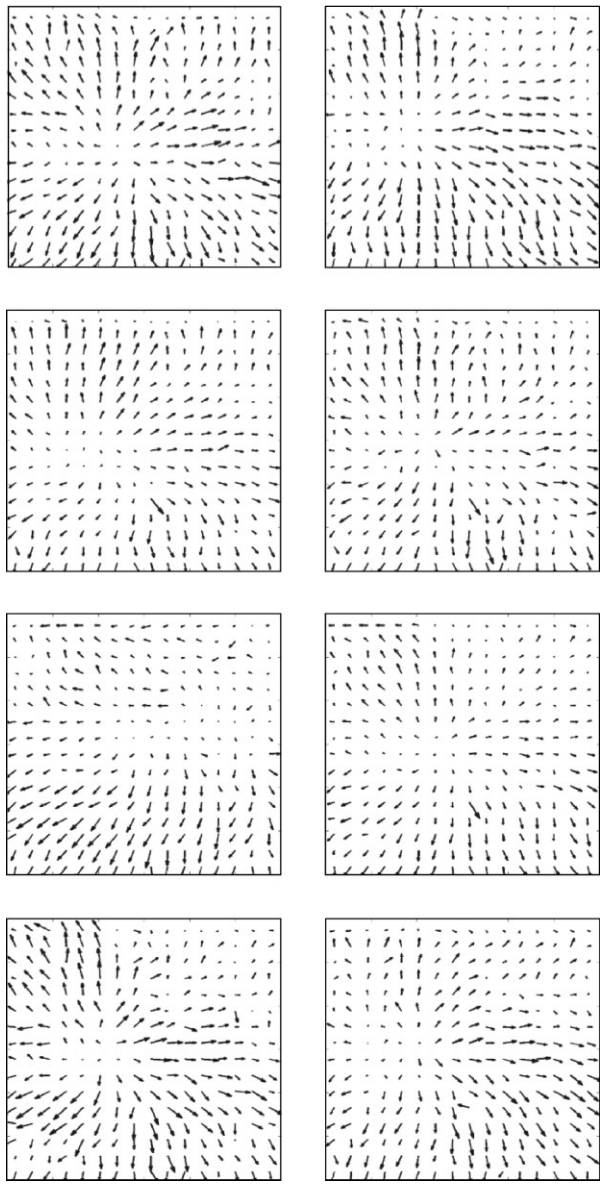

FIGURE 9 | Learned feedforward receptive fields in a predictive coding model of the middle temporal–medial superior temporal area (MT–MST) circuit. Receptive fields show preferred responses to translation and expansion similar to MST neurons. (Reprinted with permission from Ref 24. Copyright 2006 Elsevier).

Predictive Coding in Other Areas

The principle of predictive coding has also been applied to the auditory system[58,59] hippocampus,[60] ventral midbrain,[61] frontal cortex,[62] and many other brain areas.[50,53,63–67] For example, in the auditory system, predictive coding offers a way to efficiently encode the input sound signal $s(t)$ using a linear prediction of the form $s(t) = \Sigma_i r_i u_i(t)$, where r_i is the firing rate of neuron i and $u_i(t)$ are a set of synaptic weights (or basis functions). Smith and Lewicki[58]

showed a striking similarity between the optimal weights $u_i(t)$ learned from natural sounds and the impulse response function of cat auditory nerve fibers. Similarly, Mehta[60] has posited a predictive coding mechanism for the hippocampus based on the observation that the spatial receptive fields in the rat hippocampus undergo significant and rapid anticipatory changes as the rat repeatedly traverses a track. In summary, although the detailed neural representation may vary from area to area, the principle of predictive coding provides a unifying functional explanation for a variety of neural phenomena by assuming that the brain actively predicts the hidden causes of incoming sensory information.

Recent Developments in Predictive Coding

Spratling[50] has recently shown that predictive coding can be implemented using a particular form of biased competition in which neurons compete to receive inputs. In such a predictive-coding/biased-competition (PC/BC) model, the residual error is computed via lateral inhibitory connections. Feedforward and feedback connections between different areas of the brain can be both excitatory, instead of inhibitory cortical feedback as in Rao and Ballard.[21] The PC/BC model has also been shown to account for visual attention as well as extraclassical properties such cross-orientation suppression and orientation contrast facilitation.[51]

Friston et al.[68] have explored how neural dynamics can be understood in terms of prediction errors and have shown how hidden causes in a hierarchical dynamical model of the world can be estimated by optimizing free-energy. Their model can account for perceptual inference[68] as well as complex cognitive phenomena such as, decision making and motor control.[69,70]

CONCLUSION

Predictive coding provides a unifying principle for understanding the receptive field properties and neuroanatomical features of the mammalian brain. Predictive coding models characterize the function of the cortex as learning an efficient internal representation $\boldsymbol{r}$ of incoming sensory signals $\boldsymbol{I}$ by minimizing prediction errors subject to particular constraints on $\boldsymbol{r}$. These constraints, which perform 'regularization', can take the form of a penalty on the length of $\boldsymbol{r}$ (MDL principle), sparseness of $\boldsymbol{r}$, or statistical independence of $\boldsymbol{r}$. By transmitting only the unpredicted parts of an signal to the next level, predictive coding reduces the dynamic range needed to code for the incoming signal, allowing the signal to be efficiently transmitted along neural pathways with limited dynamic range. Predictive coding also ascribes a prominent computational role to feedback connections between cortical areas, positing that these connections convey predictions of expected neural activity from higher to lower levels. The feedforward connections are assumed to convey the residual prediction errors.

There exists strong experimental evidence for predictive coding in the early visual system.[7,9,13] Higher up in the visual pathway, both classical[16,30] and extraclassical[21] receptive field properties in V1 as well as receptive fields in MST[24] have been explained using predictive coding. However, it is not yet clear from neurophysiological experiments whether feedback connections indeed carry predictions and feedforward connections the residual errors, although results from neuroimaging studies[52,53] appear to be consistent with the residual error detection hypothesis. Many cognitive phenomena such as binocular rivalry,[67] mismatch negativity,[71] and repetition suppression[72] can be explained within the context of predictive coding. Finally, predictive coding has also proved useful in understanding *N*-methyl-D-aspartate (NMDA)-dependent plasticity.[60] Taken together, these examples suggest that predictive coding may be a general computational strategy employed by the brain.

REFERENCES

1. Field DJ. Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am* 1987, 4:2379–2394.
2. Ruderman DL, Bialek W. Statistics of natural images: Scaling in the woods. *Phys Rev Lett* 1994, 73: 814–817.
3. Dong D, Atick J. Statistics of natural time-varying images. *Network: Comput Neural Sys* 1995, 6: 345–358.
4. Attneave F. Some informational aspects of visual perception. *Psychol Rev* 1954, 61:183–193.
5. MacKay DM. The epistemological problem for automata. In: *Automata Studies*. NJ: Princeton University Press; 1956, 235–251.
6. Barlow HB. Possible principles underlying the transformation of sensory messages sensory. In: *Sensory Communication*. MA: MIT Press; 1961, 217–234.

7. Srinivasan MV, Laughlin SB, Dubs A. Predictive coding: a fresh view of inhibition in the retina. *Proc R Soc London Ser B* 1982, 216:427–459.
8. Meister M, Berry MJ II. The neural code of the retina. *Neuron* 1999, 22:435–450.
9. Atick J. Could information theory provide an ecological theory of sensory processing? *Network: Comput Neural Sys* 1992, 3:213–251.
10. Buchsbaum G, Gottschalk A. Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proc R Soc Lond B Biol Sci* 1983, 220:89–113.
11. Dan Y, Atick JJ, Reid RC. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J Neurosci* 1996, 16: 3351–3362.
12. Dong D, Atick J. Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network Comput Neural Sys* 1995, 16: 159–178.
13. Hosoya T, Baccus S, Meister M. Dynamic predictive coding by the retina. *Nature* 2005, 436:71–77.
14. Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. *J Physiol* 1968, 195:215–243.
15. Jones JP, Palmer LA. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol* 1987, 58: 1233–1258.
16. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 1996, 381:607–609.
17. Pasupathy A, Connor C. Responses to contour features in macaque area v4. *J Neurophysiol* 1999, 82: 2490–2502.
18. Hegde J, Van Essen DC. Selectivity for complex shapes in primate visual area v2. *J Neurosci* 2000, 20: (RC61):1–6.
19. Maunsell JH, Van Essen DC. Functional properties of neurons in middle temporal area of the macaque monkey. i. selectivity for stimulus direction, speed and orientation. *J Neurophysiol* 1983, 49:1127–1147.
20. Allbright TD. Direction and orientation selectivity of neurons in visual area MT of the macaque. *J Neurophysiol* 1984, 52:1106–1130.
21. Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 1999, 2: 79–87.
22. Mumford D. On the computational architecture of the neocortex: the role of cortico-cortical loops. *Biol Cybern* 1992, 66:241–251.
23. Pece AEC. Redundancy reduction of a Gabor representation: a possible computational role for feedback from primary visual cortex to lateral geniculate nucleus. *Artificial Neural Networks 2*. Amsterdam: Elsevier Science; 1992, 865–868.
24. Jehee J, Rothkopf C, Beck J, Ballard D. Learning receptive fields using predictive feedback. *J Physiol Paris* 2006, 100:125–132.
25. Dayan P, Hinton GE, Neal RM, Zemel RS. The Helmholtz machine. *Neural Comput* 1995, 7:889–904.
26. Luettgen MR, Willsky AS. Likelihood calculation for a class of multiscale stochastic models, with application to texture discrimination. *IEEE Trans Image Process* 1995, 4:194–207.
27. Rissanen J. *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific Publishing; 1999.
28. Schwabe L, Obermayer K. Modeling the adaptive visual system: a survey of principled approaches. *Neural Networks* 2003, 16:1357–1371.
29. Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res* 1997, 37:3311–3325.
30. Bell AJ, Sejnowski TJ. The ''independent components'' of natural scenes are edge filters. *Vision Res* 1997, 37: 3327–3338.
31. Olshausen BA. *Learning Linear, Sparse, Factorial Codes. A.I. Memo 1580*. Massachusetts: Institute of Technology; 1996.
32. van Hateren JH, Ruderman DL. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc R Soc Lond B* 1998, 265:2315–2320.
33. Rao RPN. An optimal estimation approach to visual perception and learning. *Vision Res* 1999, 39:1963–89.
34. Nicholls JG, Martin AR, Wallace BG. *From Neuron to Brain: A Cellular and Molecular Approach to the Function of the Nervous System*. 3rd ed. MA: Sinauer Associates; 1992.
35. Barlow HB, Foldiák P. Adaptation and decorrelation in the cortex. In: Miall C, Durbin RM, Mitchison GJ, eds. *The Computing Neuron*. Wokingham: Addison-Wesley; 1989, 54–72.
36. Barlow HB. A theory about the functional role and synaptic mechanism of visual aftereffects. In: Blakemore C, ed. *Vision: Coding and Efficiency*. Cambridge: Cambridge University Press; 1990, 363–375.
37. Saul AB, Humphrey AL. Spatial and temporal response properties of lagged and nonlagged cells in cat lateral geniculate nucleus. *J Neurophysiol* 1990, 64:206–224.
38. Jones JP, Palmer LA. Effect of cooling area 18 on striate cortex cells in the squirrel monkey. *J Neurophysiol* 1982, 48:38–48.
39. Murphy PC, Sillito AM. Corticofugal feedback influences the generation of length tuning in the visual pathway. *Nature* 1987, 329:727–729.
40. Mignard M, Malpeli JG. Paths of information flow through visual cortex. *Science* 1991, 93:1249–1251.

41. Zhang Z, Ballard D. A single spike model of predictive coding. *Neurocomputing* 2004, 58–60:165–171.
42. Rao RPN, Ballard DH. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Comput* 1997, 9:721–763.
43. Dobbins A, Zucker SW, Cynader MS. Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature* 1987, 329:438–441.
44. Bolz J, Gilbert CD. Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature* 1986, 320:362–365.
45. Hubel DH, Wiesel TN. Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *J Neurophysiol* 1965, 28:229–289.
46. Hubel DH, Livingstone MS. Receptive fields and functional architecture of monkey striate cortex. *J Neurosci* 1987, 7:3378–3415.
47. Desimone R, Schein SJ. Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form. *J Neurophysiol* 1987, 57:835–868.
48. Allman J, Miezin F, McGuinness E. Stimulus specific response from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu Rev Neurosci* 1985, 8: 407–429.
49. Zipser K, Lamme VAF, Schiller PH. Contextual modulation in primary visual cortex. *J Neuroscience* 1996, 16:7376–7389.
50. Spratling M. Predictive coding as a model of biased competition in visual attention. *Vision Res* 2008, 48:1391–1408.
51. Spratling MW. Predictive coding as a model of response properties in cortical area V1. *J Neurosci* 2010 March 3, 30:3531–43.
52. Murray S, Kersten D, Olshausen B, Schrater P, Woods D. Shape perception reduces activity in human primary visual cortex. *PNAS* 2002, 99:15164–15169.
53. Murray S, Scott O, Schrater P, Kersten D. Perceptual grouping and the interactions between visual cortical areas. *Neural Networks* 2004, 17:695–705.
54. Deco G, Schürmann B. Predictive coding in the visual cortex by a recurrent network with gabor receptive fields. *Neural Process Lett* 2001, 14:107–114.
55. Rao RPN, Ballard DH. Development of localized oriented receptive fields by learning a translation-invariant code for natural images. *Network: Comput Neural Sys* 1998, 9:219–234.
56. Grimes DB, Rao RPN. Bilinear sparse coding for invariant vision. *Neural Comput* 2005, 17:47–73.
57. Miao X, Rao RPN. Learning the Lie groups of visual invariance. *Neural Comput* 2007, 19:2665–2693.
58. Smith EC, Lewicki MS. Efficient auditory coding. *Nature* 2006, 439:978–992.
59. Vuust P, Ostergaard KJ, Pallesen L, Bailey C, Roepstorff A. Predictive coding of music–brain responses to rhythmic incongruity. *Cortex* 2009, 45:80–92.
60. Mehta M. Neuronal dynamics of predictive coding. *Neuroscientist* 2001, 7:490–495.
61. O'Doherty J, Buchanan T, Seymour B, Dolan R. Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron* 2006, 49:157–166.
62. Summerfield C, Egner T, Greene M, Koechlin E, Mangels J, Hirsch J. Predictive codes for forthcoming perception in the frontal cortex. *Science* 2006 November, 314:1311–1314.
63. Rao RPN, Sejnowski TJ. Predictive coding, cortical feedback, and spike-timing dependent plasticity. In: Rao RPN, Olshausen B, Lewicki MS, eds. *Probabilistic Models of the Brain.* MA: MIT Press; 2002; 297–315.
64. Rao RPN, Ballard DH. Probabilistic models of attention based on iconic representations and predictive coding. In: Itti L, Rees G, Tsotsos J, eds. *Neurobiology of Attention.* MO: Academic Press; 2004.
65. Hamker FH. Modeling feature-based attention as an active top-down inference process. *BioSystems* 2006, 86:91–99.
66. Kilner J, Friston K, Frith C. Predictive coding: an account of the mirror neuron system. *Cognitive Processing* 2007 September, 8:159–166.
67. Hohwy J, Roepstorff A, Friston K. Predictive coding explains binocular rivalry: An epistemological review. *Cognition* 2008, 108:687.
68. Friston KJ, Kiebel S. Predictive coding under the free-energy principle. Phil. *Trans R Soc B* 2009, 364: 1211–1221.
69. Hesselmann G, Sadaghiani S, Friston KJ, Kleinschmidt A. Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PLoS ONE* 2010, 5:e9926.
70. Friston KJ, Daunizeau J, Kilner J, Kiebel SJ. Action and behavior: a free-energy formulation. *Biol Cybern* 2010, 102:227–60.
71. Jääskeläinen IP, Ahveninen J, Bonmassar G, Dale AM, Ilmoniemi RJ, Levanen S, Lin FH, May P, Melcher J, Stufflebeam S, et al. Human posterior auditory cortex gates novel sounds to consciousness. *Proc Natl Acad Sci* 2004, 101:6809–6814.
72. Desimone R. Neural mechanisms for visual memory and their role in attention. *Proc Natl Acad Sci* 1996, 93:13494–13499.
73. Kaplan E, Mukherjee P, Shapley R. Information filtering in the lateral geniculate nucleus. In: Shapley R, Lam DMK, eds. *Contrast Sensitivity*. Cambridge, MA: The MIT Press; 1993, 183–200.