HEVS: A Hierarchical Computational Model for Early Stages of the Visual System

Jiuqi Han, Qingqun Kong, Yi Zeng and Hongwei Hao
Institute of Automation, Chinese Academy of Sciences
No. 95, Zhongguancun East Road, Beijing 100190, China
{jiuqi.han,qingqun.kong,yi.zeng,hongwei.hao}@ia.ac.cn

***Abstract*—Early stages of the human visual system consist of retinal cones, retinal ganglion cells(RGC), lateral geniculate nucleus(LGN) and V1. Modeling early visual stages is conducive to reveal the mechanism of visual signal preprocessing and representation inside brain, as well as settle challenges artificial intelligence confronts. However, a majority of previous work often models RGC/LGN or V1 separately, seldom modeling them together hierarchically. In order to be consistent with the biological results, we propose HEVS (a Hierarchical computational model for Early stages of the Visual System), a feedforward neural network composed of three layers, which represent receptor neurons, RGC/LGN and V1 successively. Exactly as the visual system, the proposed locally connected model is derived in the unsupervised scenario on natural images and trained in the bottom-up order. In order to learn the two connection weights among three layers, we formulate two optimization problems based on the reconstruction error and sparse learning. Unlike traditional models on RGC/LGN, we perform weighted similarity measuring as a regular term to simulate the strong correlations among nearby neuron spikes in the same stage. Different from existing researches on modeling V1 neurons from image pixels directly, we transmit the signals represented by the ganglion cells in the second layer to the V1 neurons in the third layer. Moreover, solutions to these objectives are provided as well. Experimental results demonstrate that the characteristics of HEVS are consistent with those of the corresponding biological stages. The results further verify the performance of HEVS on dealing with the de-blurring and de-noising tasks.**

I. Introduction

Over the past decades, the human brain has become a prominent topic of research in both the area of Cognitive Neuroscience (CN) and Artificial Intelligence (AI) [1] [2]. Researchers yearn for the principle of the human visual system so as to endow machines the capacity to recognize variants of objects in complex environment, which is exactly one of the most important goals in Computer Vision (CV). Among stages of the human visual system, early stages enjoy the most prevalence [3] [4], partly because most scholars nowadays approve the vital role of the signal representation (mainly completed in early stages), rather than the objects classification (mainly attained in high-level cortex). According to the structure and function, early stages consist of the retinal cones, retinal ganglion cells (RGC), lateral geniculate nucleus (LGN) and V1 [5] [6] [7], as shown in Fig. 1(a). Before being transmitted to V1 for primary representation, visual signals are processed (mainly de-blurring, de-noising and edge expression) in the retina and LGN [8].

On account of the superiority of these stages over the artificial neural networks (ANNs) on preprocessing and representing signals, numerous approaches [9] [10] have been proposed to find and employ the underlying principles of them, which could be mainly divided into two classes. Some models are built enlightened by RGC/LGN to preprocess the raw images. More attempts are made to imitate V1 for obtaining the shallow representation of signals efficiently [11] [12]. A landmark achievement for modeling these early stages is sparse coding [9] [13], as it successfully predicts the properties of the biological neurons [14] [15] and improves the object recognition performance. However, there are still two key aspects ignored by previous researches. Firstly, visual signals are disposed through the hierarchical structures shown in Fig. 1(d), while RGC/LGN or V1 modeled by previous works directly receives input from raw images as given in Fig. 1(b) and (c). Secondly, the strong correlations and intra-stage connection among nearby neurons in the same layer are reported repeatedly [16]. Nevertheless, to the best of our knowledge, few efforts have been devoted to construct these interactions.

In order to remedy these defects, we propose HEVS (a Hierarchical computational model for Early stages of the Visual System), a biophysically motivated feedforward neural network, as shown in Fig. 1(d) and (e). Our model is composed of three layers (denoted as L1, L2 and L3), corresponding to the retinal cones, RGC/LGN and V1 respectively. Different from previous models on RGC/LGN shown in Fig. 1(b), we train the L2 in HEVS from raw images with additional consideration of the intra-layer connection. Unlike traditional V1 models shown in Fig. 1(c), we obtain the L3 from signals processed by L2. Considering from the following perspectives, HEVS is consistent with the human visual system. Firstly, the hierarchical structure is constructed according to the biological visual system. Secondly, not only the inter-stage, but also the intra-stage interaction of nodes are taken into account. Thirdly, all the undetermined parameters are trained on natural images without supervision. At last, the core goal of HEVS is to maximally and efficiently represent the valid information involved in the raw images.

As shown in Fig. 1(e), we assign each pixel to a node in L1. And, there exists one intra-layer connection matrix ($\widehat{\mathbf{U}}$ in L2) and two inter-layer ones ($\mathbf{U}$ between L1 and L2, $\mathbf{V}$ between L2 and L3). In this paper, $\widehat{\mathbf{U}}$ is assigned in the light of biophysical common sense and the latter ones need to be trained by solving the corresponding constrained optimization problem J_1 and J_2 in sequence. On one hand, J_1 is formulated based on reconstruction error minimization and weighted sparse regularization. Besides, we add the similarity estimation term

978-1-4799-1959-8/15/$31.00

to J_1 inspired by the kernel distance measuring. Thus, nearby neurons are connected heavily and the receptive fields (RFs) of them could be correlated. On the other hand, J_2 is constructed on the sparse coding and least square regression. Following the definition of the problems, we provide effective methods to obtain the solutions to them, whose complexities are analyzed as well. The experimental results are presented to demonstrate the consistence of the properties of HEVS with the ones of the biological stages.

Here we introduce some notations for convenience. Throughout this paper, scalars, matrices and vectors are denoted as small, boldface capital and boldface lowercase letters respectively. $||\mathbf{x}||_p$ is the l_p-norm of $\mathbf{x}$, $trace(\mathbf{X})$ is the trace of $\mathbf{X}$ and $cov(\mathbf{X})$ is the covariance of the columns of $\mathbf{X}$ if $\mathbf{X}$ is square, and $||\mathbf{X}||_r$ is the m_r norm of $\mathbf{X}$. For any vector $\mathbf{x} \in \mathbb{R}^{n\times 1}$ and matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$, the l_p-norm and m_r norm of them are defined as

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{n} |\mathbf{x}_i|^p\right)^{\frac{1}{p}} \quad ||\mathbf{X}||_r = \left(\sum_{i=1}^{n}\sum_{j=1}^{m} |\mathbf{X}_{ij}|^r\right)^{\frac{1}{r}} \tag{1}$$

where $\mathbf{x}_i$ indicates the i-th element of $\mathbf{x}$, $\mathbf{X}_{ij}$ is the entry that occurs in the i-th row and j-th column of $\mathbf{X}$.

The rest of this paper is organized as follows: in Section 2, we review the related works. We formulate our model and provide our algorithms as well as their complexities analysis in Section 3. Section 4 gives the description about the experimental setup and the results, followed by the conclusion and future work in Section 5.

II. Related Works

Determining what underlying principles might shape visual signal processing in the brain is a central goal of CN. In particular, early stages of the human visual system have attracted many researchers to explore the latent mechanisms of preprocessing and representing signals in RGC, LGN and V1 [17] [18]. The influence of the cone receptor noise on visual signals was affirmed in [8] , which further limits the fidelity when contaminated signals are encoded by populations of RGC. It is widely acknowledged that S. Kuffler firstly detected the center-surround RFs of RGC in [19] and he further pointed that the discharge pattern of RGC could be ON-OFF or OFF-ON. D. Marr and E. Hildreth verified the similarity of the center-surround RF to difference of Gaussian (DoG) and found its function to detect edges [20]. The pioneering work on the RFs of V1 was done by D. Hubel and T. Wiesel [21]. They found that RFs of neurons in cat's striate cortex were in a side-by-side fashion with a central area flanked by antagonistic areas, which could be oriented in a vertical, horizontal or oblique manner. D. Ringach confirmed that the profiles of simple cell RFs were well described by a Gabor function, i.e., a product of a Gaussian envelope and a sinusoid [22]. In [23] [24], Gabor-like features are employed to express images and recognize objects successfully. Besides the works on RFs, D. Arnett [16] observed the correlated spontaneous discharge exhibited by some neighboring RGC. The experimental evidence for sparse coding inside the cortex is presented in [25] and [26].

Meanwhile, great efforts have been made to address the issue of modeling these stages, which could be roughly classified into modeling RGC/LGN and modeling V1. Let's consider the former firstly. In [27], Wiener filtering and robust coding were combined to derive a theoretical framework that took into account of the degradation of input images. E. Doi et al. [10] extended the models in [27] by introducing the spatial locality constraint and showed its prediction of different retinal light adaptations at different eccentricities. M. Cho et al. [28] described the RGC and LGN through the linear-nonlinear (LN) model to clarify the functional role of the RF structure and predict the cross-correlations between ganglion cell spikes. The LN model was also applied in [29] so as to capture both the properties of RF and response of RGC. In addition, a term that estimated the metabolic costs associated with firing spikes was included in its defined problem.

Numerous attempts have been made to model V1 as well [30] [31]. A complete family of localized, oriented and bandpass RFs was developed successfully by means of a coding strategy which could find sparse linear codes for natural scenes [9]. P. King et al. [12] demonstrated that synaptically local plasticity rules were sufficient to learn a sparse image code. Moreover, they also proved mathematically that the key ingredients which made it valid were sparseness and de-correlation. In [32], a spiking network model of separate populations of excitatory and inhibitory neurons was presented.

In order to build the complete model of early visual stages, we attempt to combine and extend the above researches.

III. Proposed Methods

We begin this section by defining two optimization objectives to learn $\mathbf{U}$ and $\mathbf{V}$ respectively. Then, the solutions to these problems are provided, followed by the complexities analysis of these algorithms.

A. Problem formulation

1) Objective to optimize $\mathbf{U}$*:* To simulate the effect of the eye on visual signals, we first generate the observed signals at L1 as

$$\ddot{\mathbf{R}} = \mathbf{H} * \mathbf{R} + \mathbf{n} \tag{2}$$

where $\mathbf{R} \in \mathbb{R}^{K\times N_1}$ is the original signal matrix with K signals, $\ddot{\mathbf{R}}$ is the blurred and noised $\mathbf{R}$, * is referred to convolution or filtering, $\mathbf{H}$ is a linear distortion template imitating the optical blur in vision, and $\mathbf{n} \sim (0, \sigma^2)$ is the white Gaussian noise corresponding to the neural noise. In the remaining parts of this paper, we set σ^2 by measuring the signal-noise ratio (SNR) in dB, namely $10\log_{10}(trace(cov(\mathbf{H} * \mathbf{R}))(N_1\sigma^2)$.

On account of the central goal of early visual stages to retain the useful information involved in signals, we define this objective based on minimizing the reconstruction error, i.e.,

$$\min_{\mathbf{U}} \sum_{k}^{K}\sum_{j}^{N_1}(\mathbf{R}^{\mathbf{U}}_{kj} - \mathbf{R}_{kj})^2 = \min_{\mathbf{U}} \sum_{k}^{K}\sum_{j}^{N_1}(\sum_{i}^{N_2}\mathbf{U}_{ij}\ddot{\mathbf{R}}_{kj} - \mathbf{R}_{kj})^2 \tag{3}$$

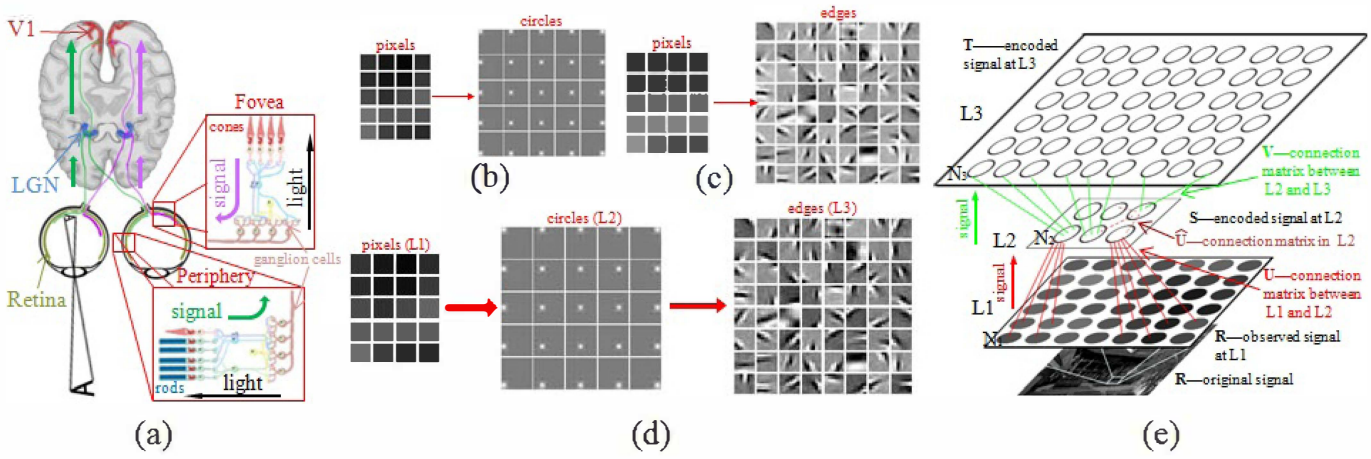

Fig. 1: (a) A brief flowchart showing the visual signal pathway from eyes to V1. The purple and green arrows, representing the right and left visual fields, depict the sequence of signal processing. The magnified details of the retina show the distribution of cone receptors and rod receptors in fovea (only cones) and periphery (concentrated rods and few cones). (b) Traditional models used to learn the center-surround receptive fields to simulate the RGC/LGN without taking into account of the intra-stage connection. (c) Previous models usually obtain the Gabor-like receptive fields from pixels directly, which do not accord with the biological results. (d) We attempt to build a three-layered network to model early stages of the human visual system hierarchically, where L1, L2 and L3 are corresponded to the stage of retinal cones (cones are sensitive in bright light and rods work better in dim light), RGC/LGN (the property and functional role of LGN are similar to that of RGC) and V1. The images are firstly observed by L1, then processed (de-blurred and de-noised) and represented by L2, lastly transmitted to L3. (e) The sketch of HEVS. The numbers of nodes in three layers (roughly shown by the size of the layer) are referred to N_1, N_2, and N_3, which are set according to the biological results. We denote the connection matrices between two neighboring layers as $\mathbf{U} \in \mathbb{R}^{N_2 \times N_1}$ (partly shown in red solid line), $\mathbf{V} \in \mathbb{R}^{N_3 \times N_2}$ (partly shown in green solid line), where the i-th row of $\mathbf{U}$ or $\mathbf{V}$ indicates the RF of i-th node in L2 or L3. The connection matrix with hat, i.e., $\widehat{\mathbf{U}}$ (partly shown in dashed line), is the intra-stage connection weight of nearby nodes in L2.

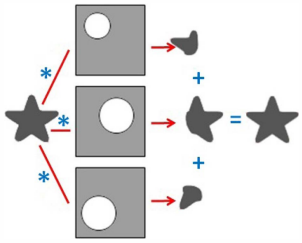

Fig. 2: A brief illustration of Eq.(3), where * indicates convolution, the first column is the observed signal, the second column is the RFs of three sampled nodes in L2, the third column is the signals perceived by them respectively and the last column is the reconstructed signal. Different nodes in L2 could perceive different parts of the observed signal. The overall reconstructed signal is obtained by summing these parts up.

where $\mathbf{R}^{\mathbf{U}}$ is the signals reconstructed by L2. Eq.(3) is equivalent to the following constraint:

$$\sum_{i}^{N_2} \mathbf{U}_{ij} = \frac{\sum_k^K \ddot{\mathbf{R}}_{kj}\mathbf{R}_{kj}}{\sum_k^K \ddot{\mathbf{R}}_{kj}^2} \tag{4}$$

It should be noticed that Eq.(3) is a little different from the traditional definition of reconstruction error. A simple comprehension of it could be found in Fig. 2. We define $\widetilde{\mathbf{U}}$ as the penalty matrix corresponding to the inter-stage synaptical energy cost, where $\widetilde{\mathbf{U}}_{ij}$ is determined by the distance between j-th node in L1 and i-th node in L2. On the basis of the intuition that the connection between nodes far from each other should be weak to save energy, we add the sparse regular term to the optimization objective J_1:

$$J_1(\mathbf{U}) = ||\widetilde{\mathbf{U}} \bullet \mathbf{U}||_2^2 \tag{5}$$

where $\bullet$ is the dot product between two matrices. In this way, far nodes (large values in $\widetilde{\mathbf{U}}$) are endowed with small connection weight (small values in $\mathbf{U}$).

As shown in Fig. 1(e), the intra-stage connection weight of nearby nodes in L2 is defined as $\widehat{\mathbf{U}}$ based on the distance among them. Intuitively, RFs of nearby nodes may have some relations in terms of their inter-stage connections. In this paper, we use the inner product to measure their similarities and this interaction could be added to J_1 as another regular term. Therefore, the overall objective to optimize $\mathbf{U}$ is written as

$$\begin{aligned} &\min_{\mathbf{U}} ||\widetilde{\mathbf{U}} \bullet \mathbf{U}||_2^2 + \alpha \sum_{j}^{N_1} \sum_{i}^{N_2} \sum_{l \neq i}^{N_2} \mathbf{U}_{ij}\mathbf{U}_{lj} \cdot \widehat{\mathbf{U}}_{il} \\ &\text{s.t.} \quad \sum_{i}^{N_2} \mathbf{U}_{ij} = \frac{\sum_k^K \ddot{\mathbf{R}}_{kj}\mathbf{R}_{kj}}{\sum_k^K \ddot{\mathbf{R}}_{kj}^2} \end{aligned} \tag{6}$$

where $\alpha > 0$ is a balance parameter.

2) Objective to optimize $\mathbf{V}$*:* As shown in Fig. 1(c), previous literatures on modeling V1 usually gain the Gabor-like RFs through learning on pixels directly [33]. However, V1 is not connected to the retina and cannot directly cope with the

pixels in fact. Instead, as shown in Fig. 1(a), signals transmitted to V1 is represented by RGC/LGN. Thus, before defining the optimization objective, we express signals $\mathbf{R}$ by all nodes in L2 as $\mathbf{S} \in \mathbb{R}^{K \times N_2}$.

Recalling that the goal of HEVS is to maximally preserve the useful information contained in signals, we also utilize the reconstruction error as the objective function:

$$J_2(\mathbf{T}, \mathbf{V}) = ||\mathbf{S} - \mathbf{T}\mathbf{V}||_2^2 \tag{7}$$

where $\mathbf{T} \in \mathbb{R}^{K \times N_3}$ is the signal represented at L3. Notice that the rows of $\mathbf{T}$ are dynamic vectors that change from one signal to the next.

To alleviate the burden of higher stages on tackling innumerable signals in a short time, signals are always represented as a sparse linear combination of neurons. Therefore, we employ the L_1 penalty sparse regularization on the rows of $\mathbf{T}$ to J_2 inspired by [13]:

$$J_2(\mathbf{T}, \mathbf{V}) = ||\mathbf{S} - \mathbf{T}\mathbf{V}||_2^2 + \beta \sum_i^K ||\mathbf{t}_i||_1 \tag{8}$$

where $\beta > 0$ is the balance parameter and $\mathbf{t}_i$ is the i-th row of $\mathbf{T}$.

Moreover, each nodes in L3 is desired to connect with as few nodes in L2 as possible, which could be gained by the sparse constraint on the rows of $\mathbf{V}$:

$$\begin{aligned} J_2(\mathbf{T}, \mathbf{V}) =& ||\mathbf{S} - \mathbf{T}\mathbf{V}||_2^2 + \beta \sum_i^K ||\mathbf{t}_i||_1 \\ \text{s.t.} \quad & \sum_j^{N_3} ||\mathbf{v}_j||_2^2 \le \mathbf{c}_j \end{aligned} \tag{9}$$

or

$$J_2(\mathbf{T}, \mathbf{V}) = ||\mathbf{S} - \mathbf{T}\mathbf{V}||_2^2 + \beta \sum_i^K ||\mathbf{t}_i||_1 + \sum_j^{N_3} \boldsymbol{\gamma}_j ||\mathbf{v}_j||_2^2 \tag{10}$$

where $\mathbf{c}$ is a constant vector, $\boldsymbol{\gamma} \in \mathbb{R}^{N_3 \times 1}$ is a balance vector and $\mathbf{v}_j$ is a row of $\mathbf{V}$.

Hence, the whole optimization problem to find the optimal $\mathbf{V}$ is stated by

$$\min_{\mathbf{T},\mathbf{V}} ||\mathbf{S} - \mathbf{T}\mathbf{V}||_2^2 + \beta \sum_i^K ||\mathbf{t}_i||_1 + \sum_j^{N_3} \boldsymbol{\gamma}_j ||\mathbf{v}_j||_2^2 \tag{11}$$

B. Solutions to the problems

1) Closed solution to Eq.(6): Besed on the assumption that the nodes in L1 are independent, Eq.(6) could be transformed to the following.

$$\begin{aligned} J_1(\mathbf{u}) &= ||\tilde{\mathbf{u}} \bullet \mathbf{u}||_2^2 + \alpha \sum_i^{N_2} \sum_{l \neq i}^{N_2} \mathbf{u}_i \mathbf{u}_l \cdot \widehat{\mathbf{U}}_{il} \\ &= \sum_i^{N_2} \tilde{\mathbf{u}}_i^2 \mathbf{u}_i^2 + \alpha \sum_i^{N_2} \sum_l^{N_2} \mathbf{u}_i \mathbf{u}_l \cdot \widehat{\mathbf{U}}_{il} \\ &= \sum_i^{N_2} \sum_l^{N_2} \mathbf{u}_i \mathbf{u}_l \boldsymbol{\Lambda} \\ &= trace(\boldsymbol{\Lambda} \mathbf{u} \mathbf{u}^T) \end{aligned} \tag{12}$$

where $\mathbf{u}$ is a column of $\mathbf{U}$, $\tilde{\mathbf{u}}$ is the corresponding column of $\widetilde{\mathbf{U}}$. $\boldsymbol{\Lambda} \in \mathbb{R}^{N_2 \times N_2}$ is an auxiliary matrix whose diagonal elements $\boldsymbol{\Lambda}_{ii}$ are $\tilde{\mathbf{u}}_i^2$ and the off-diagonal elements $\boldsymbol{\Lambda}_{il} = \alpha \widehat{\mathbf{U}}_{il}$. It should be noticed that the fact $\widehat{\mathbf{U}}_{ii} = 0$ is used to get the second equation.

Considering Eq.(4), we can rewrite the objective to optimize $\mathbf{U}$ as

$$J_1(\mathbf{u}) = trace(\boldsymbol{\Lambda} \mathbf{u} \mathbf{u}^T), \qquad \text{s.t.} \quad \mathbf{1}^T \mathbf{u} = C, \tag{13}$$

or

$$J_1(\mathbf{u}) = trace(\boldsymbol{\Lambda} \mathbf{u} \mathbf{u}^T) + \lambda ||\mathbf{1}^T \mathbf{u} - C||_2^2 \tag{14}$$

where $\mathbf{1}$ is the column vector with all elements 1, $C = \frac{\sum_k^K \ddot{\mathbf{r}}_k \mathbf{r}_k}{\sum_k^K \ddot{\mathbf{r}}_k^2}$, $\mathbf{r}$ and $\ddot{\mathbf{r}}$ are the corresponding columns of $\mathbf{R}$ and $\ddot{\mathbf{R}}$, λ (we fix λ to 10^8 in this paper) is the penalty parameter to guarantee the satisfaction of Eq.(4).

Thus, we gain

$$\frac{\partial J_1}{\partial \mathbf{u}} = 2\boldsymbol{\Lambda}\mathbf{u} + 2\lambda(\mathbf{1}\mathbf{1}^T\mathbf{u} - \mathbf{1}C) \tag{15}$$

By setting $\partial J_1 / \partial \mathbf{u} = 0$, we derive the closed optimal solution as

$$\mathbf{u} = \lambda(\boldsymbol{\Lambda} + \lambda \mathbf{1}\mathbf{1}^T)^{-1} \mathbf{1} C \tag{16}$$

2) Solution to Eq.(11): It is natural to exploit the iterative algorithm to find the optimal $\mathbf{T}$ and $\mathbf{V}$ since these two matrices are both undetermined.

We may as well try to find the optimal $\mathbf{T}$ first. Taking the independence of several signals into account, we write $\mathbf{T}$ in rows and take one of them as an example.

Eq.(11) could be written as

$$J_2(\mathbf{t}) = ||\mathbf{s} - \mathbf{t}\mathbf{V}||_2^2 + \beta ||\mathbf{t}||_1 \tag{17}$$

where $\mathbf{t}$ is an arbitrary row of $\mathbf{T}$ and $\mathbf{s}$ is the corresponding row of $\mathbf{S}$.

As a large number of algorithms [34] [35] have been proposed to the solve the problem shown in Eq.(17), we exploited the existing feature-sign search algorithm [13] to find the solution to Eq.(17).

Given fixed coefficients matrix $\mathbf{T}$, the optimization problem in Eq.(11) over connection matrix $\mathbf{V}$ reduces to the least squares problem with quadratic regularization.

$$\min_{\mathbf{V},\boldsymbol{\gamma}} ||\mathbf{S} - \mathbf{T}\mathbf{V}||_2^2 + \sum_j^{N_3} \boldsymbol{\gamma}_j ||\mathbf{v}_j||_2^2 \tag{18}$$

Generally, gradient descent is utilized to solve this kind of constrained problems. Nevertheless, we turn to another approach, i.e., Langrange dual, stimulated by [13] on account of the huge time cost of gradient descent.

When $\mathbf{V}$ is fixed, the objective to optimize $\boldsymbol{\gamma}$ could be rewritten as

$$\begin{aligned} J_{22}(\boldsymbol{\gamma}) &= \min_{\mathbf{V}} J_2(\mathbf{V}, \boldsymbol{\gamma}) \\ &= trace(\mathbf{S}^T\mathbf{S} - \mathbf{S}\mathbf{T}^T(\mathbf{T}\mathbf{T}^T + \boldsymbol{\Gamma})^{-1}(\mathbf{S}\mathbf{T}^T)^T) \end{aligned} \tag{19}$$

TABLE I: Hyperparameter settings for fovea and periphery aspect.

	Fovea	Periphery
Patch size(N)	11	25
nodes in L1(N1)	121	625
nodes in L2(N2)	121	25
N1:N2	1:1	25:1
blur matrix $\mathbf{H}$	circular averaging filter of size 21×21	none
noise $\mathbf{n}$	1dB	20dB

where $\mathbf{\Gamma}$ is a squared matrix with diagonal elements γ and other elements zero.

Thus, we could calculate the gradient and Hessian matrix of $J_{22}(\boldsymbol{\gamma})$ in the following:

$$\frac{\partial J_{22}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i} = ||\mathbf{S}\mathbf{T}^T(\mathbf{T}\mathbf{T}^T+\mathbf{\Gamma})^{-1}\mathbf{e}_i||_2^2 \tag{20}$$

where $\mathbf{e}_i$ is a vector with the i-th element 1 and other 0.

$$\begin{aligned}\frac{\partial^2 J_{22}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i\boldsymbol{\gamma}_j} &= \\ &-2[((\mathbf{T}\mathbf{T}^T+\mathbf{\Gamma})^{-1}(\mathbf{S}\mathbf{T}^T)^T\mathbf{S}\mathbf{T}^T(\mathbf{T}\mathbf{T}^T+\mathbf{\Gamma})^{-1}) \\ &\bullet(\mathbf{T}\mathbf{T}^T+\mathbf{\Gamma})]_{ij}\end{aligned} \tag{21}$$

After obtaining $\boldsymbol{\gamma}$ using conjugate gradient descent, we could find the optimal $\mathbf{V}$ by

$$\mathbf{V} = (\mathbf{S}\mathbf{T}^T)(\mathbf{T}\mathbf{T}^T+\mathbf{\Gamma})^{-1} \tag{22}$$

C. Complexities analysis

Finallywe analyze the time complexity of the proposed algorithm to solve Eq.(6), since the cost of the algorithm given in **Section** III-B2 could be easily gained by taking into account of the cost of subalgorithms used in it. In the first algorithm, we need $O(N_2^2)$ to construct $\mathbf{\Lambda}$ and computing the inverse of $\mathbf{\Lambda}+\lambda\mathbf{1}\mathbf{1}^T$ needs $O(N_2^3)$. In additional, the cost for computing $\mathbf{u}$ is $O(N_2^2)$ and there are totally N_1 columns in $\mathbf{U}$. Thus, the complexity of calculating $\mathbf{U}$ is $O(N_2^3+N_1N_2^2+N_2^2)$.

IV. Experiments

A. Experimental setup

In experiment 1, we predict the RFs of ganglion cells in fovea and periphery. In order to accelerate the learning process, we train our model on $K = 100$ patches of size $N \times N$ extracted from two randomly selected natural images, as given in Fig. 3. Because of the difference of properties between fovea and periphery, we define two sets of hyperparameters as given in Table.I to simulate these two aspects between L1 and L2. Notice that the ratio of N1:N2 is set according to the biological results (about 1:2 in fovea [36] and 30:1 in periphery [37]).

To design the intra-stage connection weight $\widehat{\mathbf{U}}$ in L2 and penalty matrix of inter-stage connection $\widetilde{\mathbf{U}}$ between L1 and L2, we locate the nodes in L2 uniformly in the L1 plane and based on which the radial basis function (RBF) is used to determine the distance among them.

With regard to the second experiment, we check the performance of our model on tackling the de-blurring and de-noising tasks, which is measured by the mean squared error (MSE) as follows:

Image 1 (965×686) Image 2 (714×753)

Fig. 3: The original images used in our experiment and the size of them are given in the bracket.

$$MSE = \frac{1}{H\cdot W}\sum_i^H\sum_j^W(I(i,j)-\hat{I}(i,j))^2 \tag{23}$$

where H and W is the height and width of images respectively, $I(i,j)$ and $\hat{I}(i,j)$ is the pixel value at the point (i,j) of images before and after reconstruction.

The patches used in the last experiment are from the signals expressed by periphery nodes. That is to say, the dimension of signals transmitted to L3 is 625. We fix $N_3 = 36\times N_2 = 900$ according to the fact that the ratio among the number of neurons for V1, LGN and RGC is approximately 40:1:1 [38]. In addition, we extract some patches from both $\mathbf{S}$ and $\mathbf{R}$ in each iteration updating $\mathbf{T}$ and $\mathbf{V}$ in order to speed up the learning process.

In the experiments, we set $\alpha = 0.08$ for fovea and $\alpha = 0.06$ for periphery. β is fixed as 0.4 according to [13] in this paper.

B. Experimental results

1) Results of experiment 1: We check the properties of the neurons simulating the ganglion cells in the fovea and periphery respectively in this experiment. The results are shown in Fig. 4.

Firstly, we observe that the size of neurons' RFs in the fovea is much smaller than that in the periphery. This result is consistent with the biological result [39] [40]. The reason primarily lies on the different structures of the receptor-RGC pathway in the fovea and periphery, which are corresponding to their functional roles. In the fovea, the ratio of the receptor (mainly cone) to RGC is near 1:1 to guarantee the resolution [41], i.e., to respond to small details of visual patterns. Whereas, in the periphery, this ratio is many-to-one to meet the requirement of the sensitivity of vision [42], that is, to simply detect the presence of a lightness change.

Secondly, it is clear that the center-surround RFs are consistent with the RFs of RGC and neurons in LGN [19] [21] [43]. The RFs of center-ON and surround-OFF, which could be regarded as a band-pass filter and difference of Gaussian (DoG) template, have been demonstrated to be useful to sharpen the contrast information of the input signal [20].

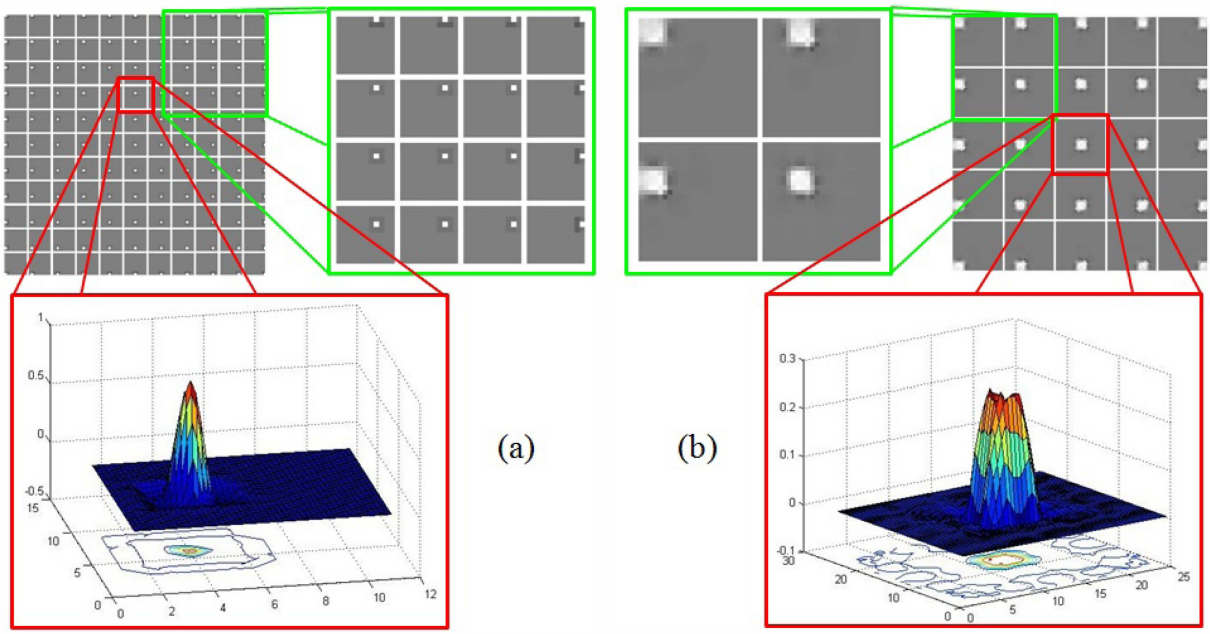

Fig. 4: Predicting different RFs of nodes in L2 simulating the RGC in fovea and periphery respectively. (a) RFs of nodes in fovea. (b) RFs of nodes in periphery. The two 3-D figures in the second row clearly depict the DoG-like RFs of the nodes in L2. The values in $\mathbf{U}$ are normalized to display.

2) Results of experiment 2: The capability of neurons in fovea and periphery to de-blur and de-noise degenerated images is tested in this part. From Fig. 5, we could find that the functional role of L2 in HEVS is similar to RGC and LGN. Especially when signals are contaminated by loud noises, our model is able to decrease the MSE by 89.7% at least and 96.5% at most, as given in Table.II and Table.III. The results further confirm the stronger power of fovea neurons on de-blurring and de-noising than of periphery neurons, which could be seen from the best performance marked bold.

We could observe another interesting result from this experiment. The statistics indicate that the effect of noise is much heavier than that of blur in reconstruction and our model is less beneficial for de-blurring than de-noising. Take Table.III as an example, the MSE of observed signals rises from 0.48 + 0.81 = 1.29 to 11.80 + 12.13 = 23.93 (1752%) with SNR reduces from 16dB to 2dB, and it rises from 22.7 to 24.02 (only 5.73%) with the blur matrix size increases from 3 pixels to 21 pixels. In response to this, the reconstruction MSE rises just from 0.25 + 0.36 = 0.61 to 0.41 + 0.49 = 0.9 (47.5% compared with 1752%) in fovea and from 0.44 + 0.50 = 0.94 to 0.46 + 0.50 = 0.96 (2.13% compared to 1752%) in periphery when the SNR decreases sharply. Conversely, the reconstruction MSE increases from 1.29 to 1.65 (27.9%, several times 5.73%) in fovea and from 1.8 to 2.02 (12.2%, 2 times 5.73%).

3) Results of experiment 3: In the last experiment, we display the sampled (100 of 900 totally) RFs of the nodes in L3 trained in **Section** III-B2. The main result is that most RFs in Fig. 6 are Gabor-like, which is strongly consistent with the biological results [21] [22]. Moreover, we could also find that the RFs satisfy the necessary characteristics of mammalian primary visual cortex, i.e., being spatially localized, oriented and bandpass [44] [45].

TABLE II: The performance (%) of our model on de-blurring and de-noising on Image 1. $S_{\mathbf{H}}$ is the size (pixel) of the blur squared matrix $\mathbf{H}$ and the first row indicates the SNR measured in dB. Smaller MSE indicates the better performance of our model. The best MSE is marked bold.

$S_{\mathbf{H}} = 3$	2 dB	4 dB	8 dB	16 dB	sum
Observed signals	12.27	7.79	3.18	**0.61**	23.85
reconstructed by fovea nodes	**1.18**	**1.03**	**0.98**	0.90	**4.09**
reconstructed by periphery nodes	1.30	1.29	1.28	1.27	5.14
$S_{\mathbf{H}} = 21$	2 dB	4 dB	8 dB	16 dB	sum
Observed signals	13.24	8.74	4.15	1.58	27.71
reconstructed by fovea nodes	**1.36**	1.43	**1.17**	**1.13**	**5.09**
reconstructed by periphery nodes	1.44	**1.36**	1.30	1.30	5.4

TABLE III: The performance (%) of our model on de-blurring and de-noising on Image 2. $S_{\mathbf{H}}$ is the size (pixel) of the blur squared matrix $\mathbf{H}$ and the first row indicates the SNR measured in dB. Smaller MSE indicates the better performance of our model. The best MSE is marked bold.

$S_{\mathbf{H}} = 3$	2 dB	4 dB	8 dB	16 dB	sum
Observed signals	11.80	7.45	2.97	0.48	22.7
reconstructed by fovea nodes	**0.41**	**0.34**	**0.29**	**0.25**	**1.29**
reconstructed by periphery nodes	0.46	0.45	0.45	0.44	1.8
$S_{\mathbf{H}} = 21$	2 dB	4 dB	8 dB	16 dB	sum
Observed signals	12.13	7.78	3.30	0.81	24.02
reconstructed by fovea nodes	**0.49**	**0.42**	**0.38**	**0.36**	**1.65**
reconstructed by periphery nodes	0.50	0.51	0.51	0.50	2.02

V. Conclusion

In this paper, we present a novel three-layered feedforward neural network, namely HEVS, to model the early stages of the human visual system. Different from previous literatures which often build the models of these stages separately, we

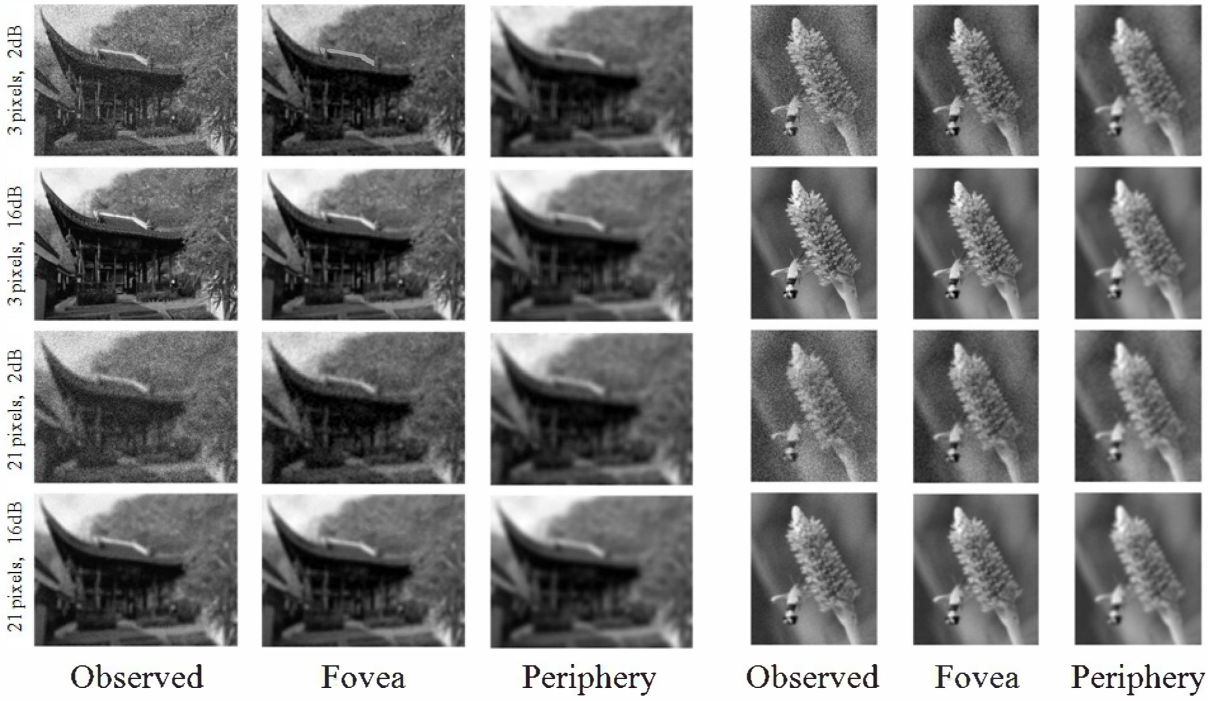

Fig. 5: Images reconstruction examples. We compare the performance of nodes at different eccentricities on de-blurring and de-noising Image 1 and Image 2. For each image, the first column is the observed signals which are generated from degrading the original images with blur and noise. The second and third columns are the reconstructed images by nodes in fovea and periphery respectively. The blur matrix size (pixel) and intensity of the SNR (dB) vary in four rows, which are given in the left.

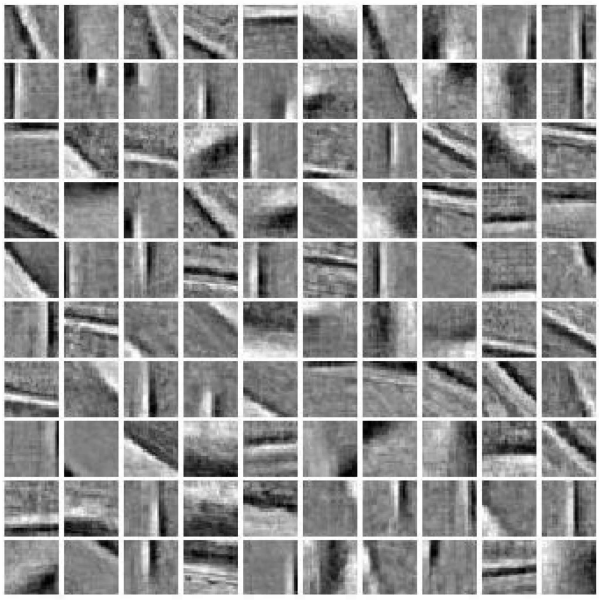

Fig. 6: Sampled RFs of the nodes in L3. The values in $\mathbf{V}$ are normalized to display.

simulate the retinal cones, retinal ganglion cells (RGC), lateral geniculate nucleus (LGN) and V1 as a unified hierarchy. We train HEVS from the bottom layer to the top layer on unlabeled natural images. An important progress of HEVS is that we simulate the strong correlation among neurons in the same stage. Another improvement of HEVS is the weighted sparse regularization on account of the synaptical energy cost. Besides the solutions provided to the optimization problems, three experimental results are shown to demonstrate that the properties of HEVS are consistent with those of the early vision stages. In addition, HEVS shows good performances in the de-blurring and de-noising tasks. The proposed HEVS can be considered as an advance for modeling the human early visual systems. Moreover, its outputs could be input into the deep neural networks which simulate the higher visual cortex, including V2, V4 and IT et., al.

Several limitations in this paper could be addressed in future work. The first one lies in modeling different kinds of cones. A complete model of different cones is desired to handling the color information. Another one is the inclusion of temporal information, which could be useful to explain the transformation of RFs of neurons.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and thank the editors for their fruitful work. This work is supported by the National Natural Science Foundation of China (No.61403375) and the Hundred Talents Program of Chinese Academy of Sciences (No.Y3S4011D31).

REFERENCES

[1] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. Rodriguez-Sanchez, and L. Wiskott, "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?" *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1847–1871, Aug 2013.

[2] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 3, pp. 411–426, March 2007.

[3] D. L. Ruderman, "Designing receptive fields for highest fidelity," *Network: Computation in Neural Systems*, vol. 5, no. 2, pp. 147–155, 1994.

[4] J. van Hateren, "A theory of maximizing sensory information," *Biological Cybernetics*, vol. 68, no. 1, pp. 23–29, 1992.

[5] M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust, "Do we know what the early visual system does?" *Journal of Neuroscience*, no. 46, pp. 10 577–10 597, 2005.

[6] H. Shan, "Computational models of early visual processing layers," *UC San Diego Electronic Theses and Dissertations*, 2010.

[7] J. L. Zylberberg, "From scenes to spikes: understanding vision from the outside in," *UC Berkeley Electronic Theses and Dissertations*, 2012.

[8] P. Ala-Laurila, M. Greschner, J. Chichilnisky, E., and F. Rieke, "Cone photoreceptor contributions to noise and correlations in the retinal output," *Nat Neurosci*, vol. 14, no. 10, pp. 1309–1316, 2011.

[9] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[10] E. Doi and M. S. Lewicki, "A simple model of optimal population coding for sensory systems," *PLoS Computational Biology*, vol. 10, no. 8, 2014.

[11] B. Haider, M. R. Krause, A. Duque, Y. Yu, J. Touryan, J. A. Mazer, and D. A. McCormick, "Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation," *Neuron*, vol. 65, pp. 107–121, 2010.

[12] P. D. King, J. Zylberberg, and M. R. DeWeese, "Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of v1," *The Journal of Neuroscience*, vol. 33, no. 13, pp. 5475–5485, 2013.

[13] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *In NIPS*. NIPS, 2007, pp. 801–808.

[14] J. Atick and A. Redlich, "What does the retina know about natural scenes?" *Neural Computation*, vol. 4, no. 2, pp. 196–210, March 1992.

[15] M. DeWeese, "Optimization principles for the neural code." in *NIPS*. MIT Press, 1995, pp. 281–287.

[16] D. Arnett, "Statistical dependence between neighboring retinal ganglion cells in goldfish," *Experimental Brain Research*, vol. 32, no. 1, pp. 49–53, 1978.

[17] B. G. Cleland, "Brisk and sluggish concentrically organized ganglion cells in the cat's retina," *The Journal of Physiology*, vol. 240, pp. 421–456, 1974.

[18] H. B. Barlow, "Summation and inhibition in the frogs retina," *The Journal of Physiology*, vol. 119, p. 69C88, 1953.

[19] S. W. Kuffler, "Discharge patterns and functional organization of mammalian retina," *Journal of Neurophysiology*, vol. 16, no. 1, pp. 37–68, 1953.

[20] D. Marr and E. Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.

[21] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[22] D. L. Ringach, "Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex," *Journal of Neurophysiology*, vol. 88, no. 1, pp. 455–463, 2002.

[23] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[25] B. Haider, M. R. Krause, A. Duque, Y. Yu, J. Touryan, J. A. Mazer, and D. A. McCormick, "Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation," *Neuron*, vol. 65, no. 1, pp. 107–121, 2010.

[26] E. Vinje, W. and L. Gallant, J., "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, pp. 1273–1276, 2000.

[27] E. Doi and M. S. Lewicki, "A theory of retinal population coding." in *NIPS*, 2006, pp. 353–360.

[28] M. W. Cho and M. Y. Choi, "A model for the receptive field of retinal ganglion cells." *Neural Networks*, vol. 49, pp. 51–58, 2014.

[29] Y. Karklin and E. P. Simoncelli, "Efficient coding of natural images with a population of noisy linear-nonlinear neurons." in *NIPS*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, Eds., 2011, pp. 999–1007.

[30] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A*, vol. 4, pp. 2379–2394, 1987.

[31] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[32] J. Zylberberg, J. T. Murphy, and M. R. DeWeese, "A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields," *PLoS Comput Biol*, vol. 7, no. 10, pp. 1–12, 2011.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." 2012, pp. 1106–1114.

[34] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–451, 2004.

[35] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, vol. 58, pp. 267–288, 1996.

[36] K. M. AHMAD, K. KLUG, S. HERR, P. STERLING, and S. SCHEIN, "Cell density ratios in a foveal patch in macaque retina," *Visual Neuroscience*, vol. 20, pp. 189–209, 3 2003.

[37] R. H. Steinberg, M. Reid, and P. L. Lacy, "The distribution of rods and cones in the retina of the cat (felis domesticus)," *The Journal of Comparative Neurology*, vol. 148, no. 2, pp. 229–248, 1973.

[38] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, Inc., 1995.

[39] H. B. Barlow, R. Fitzhugh, and S. W. Kuffler, "Change of organization in the receptive fields of the cat's retina during dark adaptation," *The Journal of Physiology*, vol. 137, no. 3, pp. 338–354, Aug 1957.

[40] R. Shapley and C. Enroth-Cugell, "Visual adaptation and retinal gain controls," *Progress in Retinal Research*, vol. 3, p. 263C346, 1984.

[41] E. A. Rossi and A. Roorda, "The relationship between visual resolution and cone spacing in the human fovea," *Nature Neuroscience*, vol. 13, p. 156C157, 2010.

[42] S. J. Williamson and H. Z. Cummins, *Light and Color in Nature and Art*. Wiley, 1983.

[43] C. Enroth-Cugell and J. G. Robson, "The contrast sensitivity of retinal ganglion cells of the cat," *The Journal of Physiology*, vol. 187, no. 3, pp. 517–552, 1966.

[44] R. L. D. Valois, D. G. Albrecht, and L. G. Thorell, "Spatial frequency selectivity of cells in macaque visual cortex," *Vision Research*, vol. 22, no. 5, pp. 545 – 559, 1982.

[45] A. J. Parker and M. J. Hawken, "Two-dimensional spatial structure of receptive fields in monkey striate cortex," *Journal of the Optical Society of America A*, no. 4, pp. 598–605, 1988.