Article:
An, L, Yan, Z, Wang, W et al. (2 more authors) (2022) Enhancing Visual Coding Through Collaborative Perception. IEEE Transactions on Cognitive and Developmental Systems. p. 1. ISSN 2379-8920

https://doi.org/10.1109/tcds.2022.3203422

Enhancing Visual Coding Through Collaborative Perception

Lingling An, Zhen Yan, Weizheng Wang, Jian K. Liu, and Keping Yu

***Abstract*—A central challenge facing the nature human-computer interaction involves understanding how neural circuits process visual perceptual information to improve the user's operation ability under complex tasks. Visual coding models aim to explore the biological characteristics of retinal ganglion cells to provide quantitative predictions of responses to a range of visual stimuli. The existing visual coding models lack adaptability in natural and complex scenes. Therefore this paper proposes an enhanced visual coding model through collaborative perception. Our model first extracts the multi-modal spatiotemporal features of the input video to simulate the retinal response characteristics adaptively. Secondly, it uses the basis function to compile the input stimulus into a multi-modal stimulus matrix. Afterward, the upstream and downstream filters reform the stimulus matrix to generate the spike sequence. Experiments show that the proposed model reproduces the physiological characteristics of ganglion cells in the biological retina, leading to the high accuracy, good adaptability, and biological interpretability in comparison with its rivals.**

***Keywords*—Visual Coding, Multi-modal Stimulus, Feature Compilation, Nonlinearity.**

I. Introduction

With the emergence of hybrid intelligence systems, a deep understanding of human perception characteristics is vital to designing natural human-computer interaction. Vision is the most important source of external information input for human beings, which not only provides the input of external environment information but also greatly affects deep-level emotions and cognition. The retina plays an essential role in the human's most important sensory system – vision, which converts complex external light stimulus signals into bioelectrical signals to provide subsequent visual nerves to construct biological vision. The research of signal conversion on the retina can help understand the deep mechanism of human vision, thereby providing a theoretical basis for the realization of artificial vision [1]. Baccus et al. focused on salamander retinal ganglion cells (RGCs) and presented that input visual stimuli transmit nerve impulses to postsynaptic neurons through the polarization/depolarization of membrane potential [2]. The research showed that the ON-/OFF-type bipolar cells in the retina and the biological connection structure on the connection layer where they are located can help achieve preliminary stimulation processing and information compression [3] [4]. Gollisch et al. pointed that the nonlinearities within ganglion cell receptive fields (RFs) are of great importance to building actual neural code of retinal neurons when designing visual prostheses for the eye [5].

Existing visual coding models can be mainly divided into two categories: 1) linear-nonlinear (LN) calculation based and 2) neural networks based. In the first category, the input stimulus and the nonlinear calculation mode determine the final performance of the model. By extracting the temporal and spatial characteristics of visual stimuli, specific nonlinear calculations are used to obtain the activation rate curve of neurons. For example, Carandini et al. presented a linearity and normalization model in simple cells of the macaque primary visual cortex [6], which is designed to account for important response nonlinearities. This LN model extends the linear model to include mutual shunting inhibition among a large number of cortical cells. Pillow et al. proposed a probabilistic spiking model to predict and decode the retinal ganglion cell responses [7], termed as the linear-nonlinear Poisson (LNP) model, which consists of a leaky integrate-and-fire model driven by a stimulus-dependent current, a spike history-dependent current, and a Gaussian noise current. The LNP model can predict the detailed time structure of responses to visual stimuli, capture the interaction between the spiking history and sensory stimulus selectivity, and explain the variability in responses to repeated stimuli. The generalized quadratic model (GQM) provides a high-order approximation to the Wiener-Volterra expansion [8], which fits and trains the 'mean-variance' of the output pulse sequence, resulting in a more powerful and more flexible model for feature space inference. In the retinal coding experiment of artificial white noise, simple stimulus signals based on brightness can artificially realize biological nonlinear calculations in retinal ganglion cells. Moreover, environmental features of complex natural scenes, such as rapid changes in brightness, changes in contrast and brightness, movement of objects, and subtle changes in sharp edges, are necessary for reconstructing the physiological process of the retina [9]–[12]. In view of the different stimulus effects caused by different levels of input features, Park et al. used basis functions with tail effects to encode the original stimulus signal, making the stimulus effect more bio-interpretable [13]. Heitman et al. proposed a

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62072355 and 62072354),the Key Research and Development Program of Shaanxi Province of China (Grant No. 2022KWZ-10) and the Natural Science Foundation of Guangdong Province of China (Grant No. 2022A1515011424). *(Corresponding author: Weizheng Wang)*

Lingling An and Zhen Yan are with the School of Computer Science and Technology, Xidian University, Xi'an, China (email: an.lingling@gmail.com; yanzhen_xdu@163.com).

Weizheng Wang is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (email: weizheng.wang@ieee.org).

Jian K. Liu is with the School of Computing, University of Leeds, UK (email:j.liu9@leeds.ac.uk).

Keping Yu is with the Graduate School of Science and Engineering, Hosei University, Tokyo 184-8584, Japan, and with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan. (email: keping.yu@ieee.org).

generalized linear model (GLM) to encode the responses of primate RGCs to naturalistic visual stimuli [14], which shows that the additional spatial nonlinearities, gain control, and/or peripheral effects are important in the first stage of visual processing. Maheswaranathan et al. presented a computational framework to fit parameters of hierarchical nonlinear models to recordings of ganglion cells in the retina, in which the LN-LN models consist of three cell layers connected by two states of LN processing [15]. Experiments show the LN-LN models can accurately describe the retinal coding and is amenable to biophysical interpretation.

With the rapid development of deep learning, visual coding models through artificial neural networks has attracted much attention [16]. Mcintosh et al. proposed deep learning models of the retinal response to natural scenes [17], which constructs 3-layer deep convolutional neural networks (CNNs) to capture retinal responses to natural image sequences. This model is shown to be more accurate than LN models and GLMs, but it has a weak ability to sense low-order global features and faces the problems of over-fitting and parameter explosion if increasing the network layers. Afterward, a CNN architecture with a sparse readout layer was designed to factorize the spatial and feature dimensions. In addition, multilayer recurrent neural network (RNN) models are proposed to model complex computations of neurons of primate retinal ganglion cell responses [18], which obtain good predictive ability with small amounts of training data by capturing spatial and temporal processing components. However, the models are vulnerable to nonlinear neurons with weak spike-triggered average responses. Sinz et al. developed a deep RNN to predict inferred spiking activity of mouse V1 neurons [19], which can correctly predict the orientation tuning of neurons in responses to artificial noise stimuli. The convolutional RNN [20] used the convolution operation to extract the spatiotemporal features of the input video to improve the performance of the model when dealing with natural scenes. However, the neural network-based visual coding models suffer from the high computational cost and poor interpretability.

In order to improve the adaptability of retina coding (RC) models in complex natural scenes, this paper proposes a multi-modal stimulus non-linear model, termed MSNM. First, our model extracts spatiotemporal features of the visual stimuli in the receptive field (RF) of ganglion cells to obtain the complex exterior stimuli, which is performed by the connection structure of bipolar cells and amacrine cells on the real biological retina. By considering the influence of the luminance intensity of the input video, instantaneous changes in brightness and darkness, and movement of objects, our model then uses the basis function with time distribution characteristics for convolutional coding and constructs a multi-modal stimulus matrix. Thirdly, the stimulus matrix is filtered to obtain the firing rate curve of neurons through a two-layer LN filters with upstream and downstream structures. Finally, the neurons emit pulse spikes according to the firing rate and the Poisson random process. In comparison with the existing models, our model is featured by

1) applicability. The proposed model can encode video signals in different environments due to its adaptive feature selection. For example, in a weakly illuminated scene, the rod cells in the human retina are sensitive to light intensity, while the color-sensitive cone cells have a weakened response. In this case, our model keeps the stimulus input of the luminance intensity and neglects the stimulus corresponding to color channels.
2) flexibility. Convolutional coding of basis functions provides the time-domain reshaping of multi-modal input stimuli and also emphasizes the impact of multiple stimuli on the coding results in different environments. In this way, the stimulus information can be more concentrated in a small-sized stimulus matrix, which speeds up the training and coding speed of our model. In addition, the basis function distribution can be flexibly adjusted according to the parameters such as the duration and offset.
3) interpretability. Our model is a nonlinear retinal coding model with the upstream and downstream structures. In the upstream filtering stage, two independent linear filter subunits are designed to extract the 'excitation-inhibition' information in the stimulus matrix, and the piece-wise nonlinear function performs local optimization to obtain the upstream filtering result. In addition, our model introduces peripheral control inputs, such as historical excitation states. Such a process is similar to the excited state in biological retinal neurons, which makes the model reliable and biologically interpretable.

The remainder of this paper is organized as follows: Section II details the modules in the proposed model including the feature extraction and compilation, nonlinear retinal coding, and parameters estimation. The experimental results are analyzed in Section III. Section IV concludes the paper.

II. Proposed model

In this section, we detail the proposed multi-modal stimulus nonlinear model, termed MSNM, which physiologically reproduces the information inputs of RGCs with precise spiking output prediction performance. As shown in Fig. 1, the MSNM includes two modules: a) feature extraction and compilation and b) non-linear retinal coding. Given a natural video, the multi-modal spatiotemporal features, such as the luminance intensity, instantaneous changes in brightness and darkness, movement of objects, etc., are firstly extracted and compiled by basis functions to generate the multi-modal stimulus matrix. Secondly, the proposed MSNM employs a pair of LN upstream subunits to filter the stimuli matrix to simulate the excitation/inhibition balance mechanism in visual circuits. Finally, the downstream filter activates the upstream results to obtain the spike sequences.

A. Feature Extraction and Compilation

In the upstream of visual circuits, the photoreceptor cells and transport layer cells preprocess the input video and extract spatiotemporal features. Therefore, how to distinctly stimulate this feature selection mechanism is of great importance to RC models. The LN model simply considers the luminance intensity of pixels in the RF as external inputs of ganglion

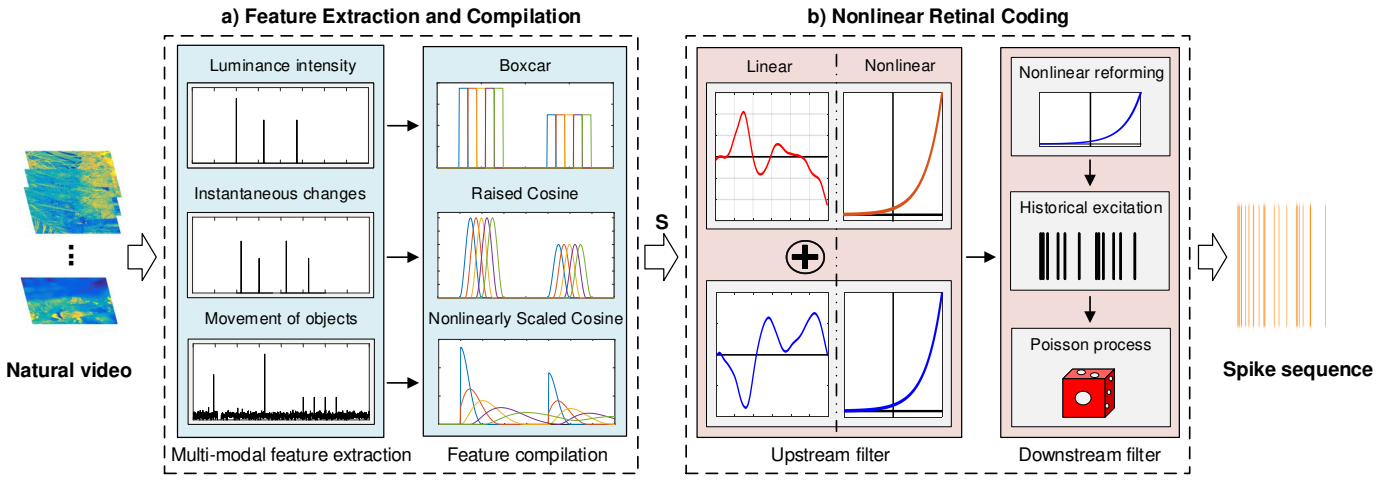


Fig. 1. The framework of the proposed MSNM. It includes two modules: a) feature extraction and compilation and b) nonlinear retinal coding. The $\boldsymbol{S}$ represents the multi-modal stimulus matrix of the natural video after feature extraction and compilation.

cells, which is vulnerable to complex natural scenes due to the lack of effective feature exaction [6]. To obtain high prediction performance, the GQM overemphasizes the mathematical fitting rather than the biological mechanism [8]. In addition, neural network-based RC models employ deep convolution networks to extract features, which is a precise approach but requires much computation cost. Although using large convolution kernels and reducing network depth can alleviate the overfitting and computing cost, the lack of physiological interpretability is a barrier to broad applications.

To target the weakness above of RC models, we propose to extract multi-modal spatiotemporal features from natural videos based on the physiological response mechanism of the retina. Given a natural video with N frames, we first calculate the sum of luminance intensity in tth video frame to construct the luminance intensity feature vector $\mathbf{f}_1 = [f_1(t)]_{t=1}^{N}$ by

$$f_1(t) = \sum_x \sum_y R * V(x, y, t) \tag{1}$$

in which $V(x, y, t)$ is the luminance intensity at the (x, y) in the tth frame, R is the range of RF, the $*$ means the 'center-edge' RF extraction operation. According to our research findings, the instantaneous changes in brightness and darkness can also make ganglion cells generate spikes. Based on this, the instantaneous feature vector $\mathbf{f}_2 = [f_2(t)]_{t=1}^{N}$ is secondly captured by

$$f_2(t) = f_1(t) - f_1(t-1) \tag{2}$$

Previous research agrees that the response of ganglion cells to visual stimuli is not the simple summation of luminance intensity [21]. Local features or variations in contrast or colors can alter the excitation state of ganglion cells significantly. Therefore, both brightness of the central area and object movement is collected and normalized to build the movement feature vector $\mathbf{f}_3 = [f_3(t)]_{t=1}^{N}$ by

$$f_3(t) = \sum_x \sum_y (V'(x, y, t) - V'(x, y, t-1)) \tag{3}$$

$$V'(x,y,z) = \iiint \frac{1}{\sqrt{(2\pi)^3 |\sum|}} exp(-\frac{1}{2}(V(x,y,z) - \boldsymbol{\mu})^\top \sum^{-1} (V(x,y,z) - \boldsymbol{\mu}))dxdydz \tag{4}$$

where $V'(x, y, t)$ means the result when the tth video frame is processed by 3-D Gaussian blur, μ is the mean vector, and $\sum$ is the covariance matrix.

After the feature extraction, the proposed MSNM compiles the input stimuli into multi-modal stimulus matrix $\mathbf{S} = [\mathbf{S}_1\ \mathbf{S}_2\ \mathbf{S}_3]$ by flexibly selecting basis functions, in which the kth basis matrix, $\mathbf{S}_k$, is calculated by

$$\mathbf{S}_k = \mathbf{f}_k \otimes \mathbf{B}_k, \quad k = 1, 2, 3. \tag{5}$$

Here, the $\mathbf{f}_k$ is the kth feature vector, the $\mathbf{B}_k$ means the corresponding basis function, and the $\otimes$ performs the 2-D convolution operation. It can be seen that the compilation process can transform the simple stimuli vector into a 2-D matrix, including both instantaneous stimulus and persistent effect by controlling the parameters of $\mathbf{B}_k$, which helps to stimulate the biological process of visual circuits on the retina.

The response of neurons to a single input stimulus can be considered a cluster of a series of physiological processes. Receiving an external stimulus, the object neuron will generate both a rapid response and some weak but persistent change in membrane potential, which demands basis functions to compile input stimuli into a distribution that contains variable temporal features. For example, the basis function 'Boxcar' can transform an afferent stimulus into a continuous equivalent stimulus effect with a specific time delay [22], defined by

$$\boldsymbol{B}^{\eta}(\varepsilon) = \begin{cases} c & \text{for } \eta = 1 \text{ and } \varepsilon_0 < \varepsilon < \varepsilon_1 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where c is a constant, ε_0 and ε_1 are the upper and lower bounds of the interval, respectively. The 'Raised Cosine' function is based on 'Boxcar' and takes the subtle temporal effect of external stimuli into account [23]. The first order raised cosine function over a finite interval, $-T < x < T$, is defined as

$$\boldsymbol{B}^{1}(x) = \begin{cases} \frac{1}{2T}[1 + \cos\frac{\pi}{T}x] & -T < x < T \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The 'Nonlinearly Scaled Cosine (NSC)' function is defined by

$$\boldsymbol{B}^2(x) = \frac{1}{2}(\cos[mid(-\pi, \frac{lnx-\beta}{2\beta}\pi, \pi)] + 1) \tag{8}$$

where mid function means to take the middle value of the three, and the β is a constant greater than zero [24]. It is noted that different basis functions can be employed to achieve the sustained effect of stimulus for different application situations.

B. Nonlinear Retinal Coding

The nonlinear computation in RGCs is the key component of retinal coding models. Inspired by the nonlinear input model (NIM) [25], we utilize a cascade LN scheme with upstream and downstream structures to simulate the biological computation of RGCs. In the upstream part, a pair of parallel LN subunits is first used to filter the multi-modal stimulus matrix $\boldsymbol{S}$.

$$G(t) = \sum_{i=1}^{2} g_i(t) = \sum_{i=1}^{2} \omega_i y_i(\boldsymbol{S} \cdot \boldsymbol{k_i}) \tag{9}$$

in which the $G(t)$ is the generation function that represents the response of neurons to the input stimulus. The $\boldsymbol{k_i}$ is a linear filter that transforms the input stimulus matrix into the optimal solution in the corresponding parametric space. The $y(\cdot)$ is a nonlinear calculation that rectifies the liner filter result to further find the optimal nonlinear solution in a local field. The ω_i is an excitation/inhibition scale term and determines whether each subunit has an 'excitatory' or 'inhibitory' effect on the neuron. When the neuron generates a selective response by excitation or inhibition subunit, the generation function $G(t)$ with high values will generate firing spikes. If the results of the linear and nonlinear filters are offset, the neuron keeps rest.

After getting the response of neurons to the input stimulus, an activation function is necessary to reform the $G(t)$ into the firing rate $r(t)$ of the RGCs. The spiking of RGCs is sparse which means it keeps rest in most of the time. Thus, the $G(t)$ obtained from the upstream filters needs further nonlinear calculation reforming.

$$r(t) = \alpha\, log[1 + exp(\beta(G(t) + \boldsymbol{h} \cdot \boldsymbol{x} - \gamma))] \tag{10}$$

where the $\boldsymbol{h}$ is the linear operation on extra external input $\boldsymbol{x}$ such as historical excitation state. The α, β, and γ determine the scale, shape, and offset of the spiking nonlinearity respectively. A stochastic Poisson process is finally conducted to generate the spike sequence.

C. Parameters Estimation

The proposed MSNM can be described as a quadratic linear-nonlinear model, including an upstream parallel LN module and a downstream spiking nonlinear computation. The whole process can be expressed by

$$r(t) = F[\sum_i \omega_i y_i(\boldsymbol{S} \cdot \boldsymbol{k_i}) + \boldsymbol{h} \cdot \boldsymbol{x}] \tag{11}$$

The parameter estimation of the MSNM mainly includes the linear filter $\boldsymbol{k_i}$, the subunit nonlinear rectification function $y_i(\cdot)$ and the downstream nonlinear function $F[\cdot]$. Assuming that the spiking process is a conditionally inhomogeneous Poisson process with rate $r(t) = F(\boldsymbol{k}, \boldsymbol{S}, y(\cdot), \alpha, \beta, \gamma)$, according to the general point-process theory [26], the log-likelihood LL of the observed spike sequences $Q(t)$ is defined as

$$LL \sim \int Q(t) log(F(\boldsymbol{k}, \boldsymbol{S}, y_i(\cdot), \theta)) - F(\boldsymbol{k}, \boldsymbol{S}, y_i(\cdot), \theta) dt \tag{12}$$

in which the θ is the $\{\alpha,\beta,\gamma\}$ for short. To optimize the parameters to maximize the log-likelihood LL, we utilize the block coordinate ascent to estimate the parameters respectively [27]. When optimizing one set of parameters, the other two parameters are fixed. The optimization can be conducted alternately to obtain the global solution as long as the$y_i(\cdot)$ is piece-wise linear and $F[\cdot]$ is concave.

The estimation of the linear filter $\boldsymbol{k_i}$ is accomplished by calculating the gradient of log-likelihood gradient with respect to the $\boldsymbol{k_i}$ by

$$\frac{\partial LL}{\partial k_{i,j}} = \sum_t (\frac{Q(t)}{r(t)} - 1) F'[G(t)] \omega_i y_i'(g_i(t)) \boldsymbol{S}_j(t). \tag{13}$$

Here, $F'[\cdot]$ and $y_i'(\cdot)$ are the derivatives of $F[\cdot]$ and $y_i(\cdot)$ with respect to their arguments, and $\boldsymbol{S}_j(t)$ is the jth column of the multi-modal stimulus matrix at the time t. In addition, a L2 norm constrained is employed to prevent the fragility caused by minimum singular value in $\boldsymbol{S}$.

The upstream nonlinear function $y(\cdot)$ is often defined as parametric functions such as the rectified-linear or regressive quadratic function. Here we provide a non-parametric scheme to construct the function $y(\cdot)$. Firstly, the $y_i(\cdot)$ is initialized to be a non-zero linear function as

$$y(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise} \end{cases}. \tag{14}$$

The filter $\boldsymbol{k_i}$ is estimated based on this non-zero linear function. Afterward, the upstream rectification functions $y(\cdot)$ are added nonlinearity by decomposing the $y(\cdot)$ into a piece-wise linear basis function, $y_i(t) = \sum_j a_{ij}\varphi_j(t)$ when keeping the $\boldsymbol{k}_i$ fixed. These basis functions $\varphi(x)$ are given by

$$\varphi(x) = \begin{cases} \frac{x - x_{k-1}}{x_k - x_{k-1}} & \text{if } x \in [x_{k-1}, x_k] \\ \frac{x_{k+1} - x}{x_{k+1} - x_k} & \text{if } x \in [x_k, x_{k+1}] \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

in which the x_k is a set of grid points which is selected from the n-quantiles of $g_i(t)$.

The downstream nonlinear parameters thus can be estimated by the log-likelihood gradient to $\{\alpha,\beta,\gamma\}$ respectively when the upstream filters are optimized and fixed in this iteration.

In summary, the proposed multi-modal stimulus non-linear model achieves high flexibility and expansibility. In the stage of feature extraction and compilation, the MSNM can provide good adaptation to different application situations by adjusting the types of input stimuli and basis functions. In the stage of nonlinear retinal coding, two upstream filter subunits can be added to simulate complex physiological mechanisms in RGCs and gain excellent prediction performance. Moreover, any non-linear monotonous function can be employed as the

downstream spiking function as long as it is concave in the parameter space.

III. Experimental results and analysis

A. Experimental Configuration

In order to verify the performance of the proposed MSNM, this section utilizes the log-likelihood LL, cross-correlation [28], skewness [29], and Clayton copula function [30] as the metrics for evaluation. The LL can accurately capture the subtle changes in the excitation probability of biological neurons and estimate the goodness-of-fit.

The cross-correlation function describes the degree of correlation between the values of random signals $X(t)$, $Y(t)$ at any two different times t_1 and t_2. The cross-correlation function $P_{XY}(t,\tau)$ is given by

$$P_{XY}(t,\tau) = E[Y(t+\tau)X(t)], \tag{16}$$

where the $E(\cdot)$ means the expected value of the product of $Y(t+\tau)$ and $X(t)$, and the lag τ is positive.

The skewness is the number of features that characterize the degree of asymmetry of a probability distribution density curve relative to the mean. Intuitively, it is the relative length of the tail of the density function curve. The definition of skewness is given by

$$d_3 = \sqrt{\frac{m_3^2}{m_2^3}} \tag{17}$$

where

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r, \quad r = 1, 2, \ldots \tag{18}$$

is the rth central moment and $\bar{x}$ is the arithmetic mean of n real numbers $x_i (i = 1, 2, \ldots, n)$.

The Clayton copula function is one of the most popular parametric families of copulas, which can effectively describe the nonlinear tail dependence between variables, defined as:

$$C^C(u,v) = \left[u^{-\zeta} + v^{-\zeta} - 1\right]^{-\frac{1}{\zeta}}, \quad \zeta > 0 \tag{19}$$

where C is a copula function, u and v are two arbitrary probability distribution functions, and the ζ is a constant greater than zero.

B. Physiological Feature Extraction and Compilation

In the mammalian retina, the light stimulus signal is processed by photoreceptor cells, transport layer cells, and ganglion cells and then converted into a time series of spikes carried by neurotransmitters. Some simple retinal coding models think that the visual coding of ganglion cells exclusively depends on the light luminance input. Some recent models introduce additional control information gained from physiological functions of neurons [14]. But the visual processing in the retina goes beyond these simple hypotheses in the following way.

First of all, the photoreceptors collectively refer to rod cells and cone cells, which transmit bioelectrical signals to ganglion cells through bipolar cells and amacrine cells in the transport layer, respectively. The two photoreceptors make the retina form a typical 'center-edge' sensitivity. Secondly, the biological retina also has a significant ability to respond to the input visual stimuli instantaneously [21]. When the input video contains severe or rapid vibration in luminance intensity, the ganglion cells also produce high-frequency spiking signals. In addition, the intrinsic attributes of input images, such as contrast, can also profoundly affect the coding state of RGCs and provide additional linear space for the visual coding. Thirdly, besides the hierarchical structure of the photoreceptor, transport, and coding layers, the visual neural circuits in the retina contain additional control mechanisms within the horizontal connection. Therefore, it is very necessary and reasonable to extract the rich features of the input video to show the physiological characteristics of retina coding.

In this paper, the proposed MSNM extracts the multi-modal features (MMF) through collaborative perception and compiles the feature vectors by flexible basis functions, leading to higher prediction performance. By contrast, the stimulus matrix is constructed by applying a simple delay to the input luminance intensity in the NIM and is directly fed to the coding module without the feature compilation. Fig. 2(a) shows the cross-correlation values vs. the log-likelihood LL of three RC models to demonstrate the influence of introducing multi-modal features on performance. Here, we create the NIM-MMF by adding the features, such as instantaneous changes and movement of objects used in the MSNM, to the NIM. According to the results, adding additional stimulus features can improve the prediction performance with greater LL and cross-correlation. Since the proposed MSNM further compiles the extracted MMF by the basis functions, it achieves the highest values of LL and cross-correlation. Moreover, to show the influence of feature compilation on the performance, Fig. 2(b) compares the cross-correlation vs. the log-likelihood LL of three RC models. In particular, the NIM compiled represents the model in which the basis function compiles the luminance intensity in the NIM. It can be seen that the prediction result is promoted when the afferent light luminance is compiled by the basis function, as shown in the results of the NIM and NIM compiled.

C. Flexibility and Biological Interpretability

In our model, different combinations of features and parameters determine the stimulus matrix, making the MSNM adaptable to complex nature scenes flexibly. Fig. 3 shows the distributions of singular values for different basis functions, including the 'Boxcar', 'Raised Cosine', and NSCs with different nonlinear offsets. To be specific, the NSC-i means the 'Nonlinearly Scaled Cosine' basis function with the nonlinear offset is i. By adjusting the parameters such as the duration and the offsets, the basis function has a distinctive distribution of its singular values. For example, the singular values of the NSC-1 concentrate more on the top than the ones of the NSC-50. It means the stimulus matrix can be compiled by the basis function smaller in size, but the performance remains good, which is of great importance in those tasks for fast coding needs. In this way, our model achieves good flexibility in comparison with its rivals.

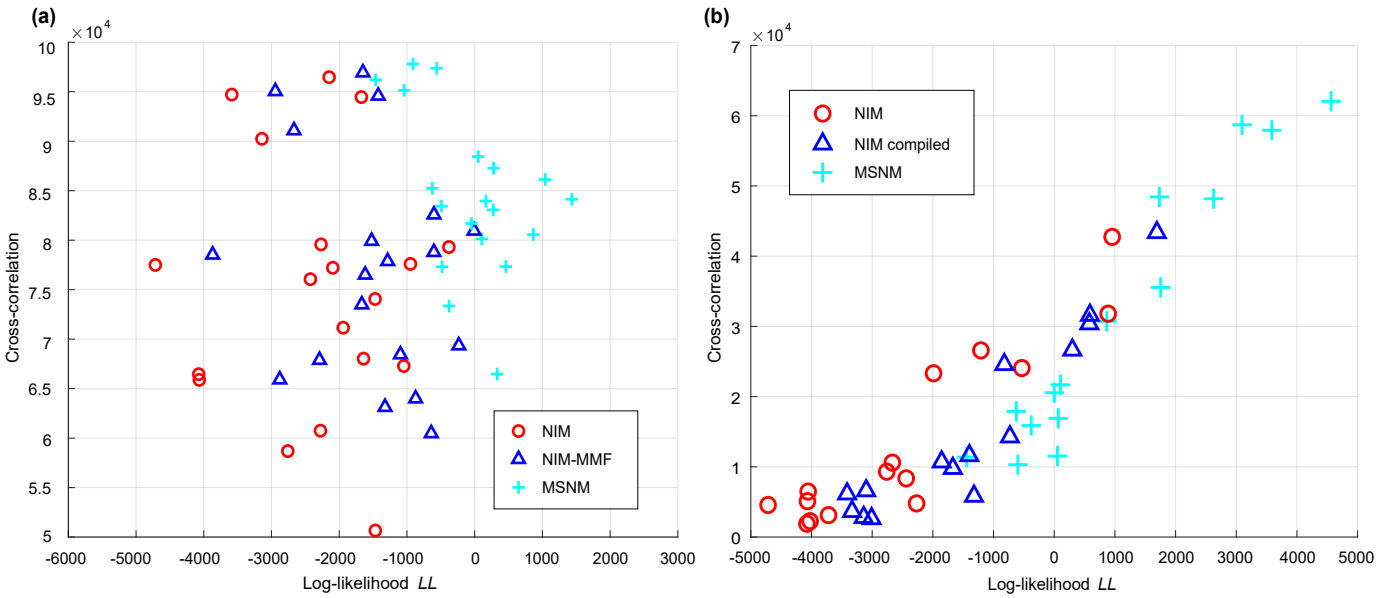


Fig. 2. The comparison results of the cross-correlation vs. log-likelihood LL. The higher the LL is, the better the prediction performance is. And the higher cross-correlation means the RC model highly fits the natural responses of the retina. For the clearness, we plot the results of 19 neurons and 15 neurons in (a) and (b), respectively.

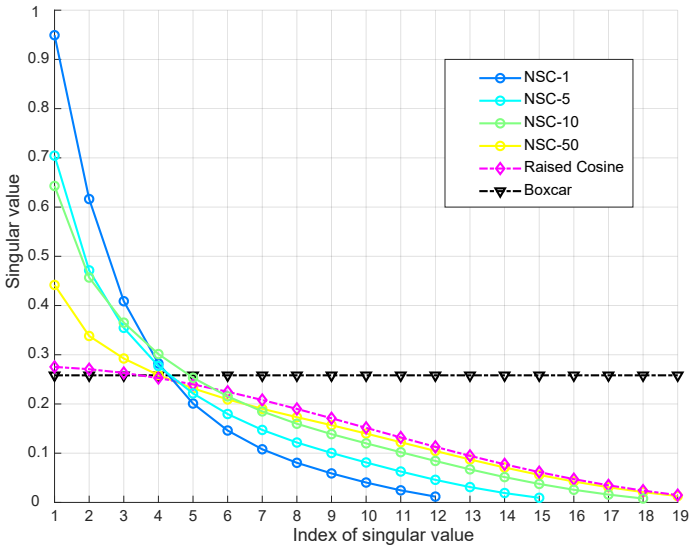


Fig. 3. Singular value distribution for different basis functions. The NSC-i means the 'Nonlinearly Scaled Cosine' basis function with the nonlinear offset is i. The duration of all basis functions is 100.

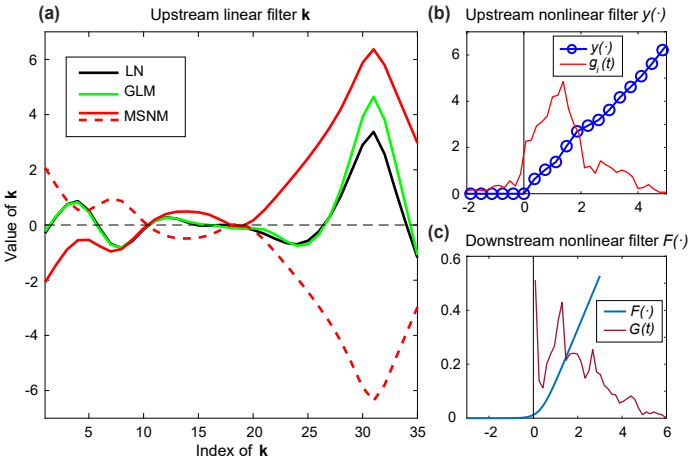


Fig. 4. Filters for different RC models. The (a) shows the upstream linear filter $\boldsymbol{k}$, the horizontal axis of (a) means the index of $\boldsymbol{k}$, and the vertical axis is its value. The (b) shows the upstream nonlinear filter $y(\cdot)$ and the histogram of $g_i(t)$. The (c) is the downstream nonlinear filter and the histogram of the generation function $G(t)$.

In the stage of nonlinear retina coding, our MSNM designs an upstream linear filter with two parallel subunits, a piece-wise nonlinear filter, and a downstream nonlinear filter for activating the generation function. Fig. 4 and Fig. 5 show the upstream and downstream filters for different RC models. As shown in Fig. 4 (a), the upstream parallel filter subunits in our MSNM capture the excitation and inhibition information from the stimulus matrix, which better fits the biological characteristics of the retina coding.

The GQM is a second-order model generated by the spike-triggered average (STA) and spike-triggered covariance (STC) statistics. A linear filter and a nonlinear quadratic filter provide an approximate decomposition like in Taylor Formula. As shown in Fig. 5 (b), the fixed quadratic nonlinear filter in the GQM results in the bias between its linear filter and the actual retina filtering when a good mathematical fitting is needed. By contrast, our upstream nonlinear filter is piece-wise nonlinear, as shown in Fig. 4 (b), and the parameter optimization of $\boldsymbol{k}$, $y(\cdot)$ and $F(\cdot)$ can achieve a satisfactory solution based on biological mechanisms of the retina coding. Given a biological filter of RGCs, Fig. 6 further compares the upstream linear filters $\boldsymbol{k}$ in the GQM and MSNM. It can be noticed that our MSNM outperforms the GQM in terms of fitting results, which indicates the MSNM can better restore the excitation/inhibition filtering of the RGCs. In addition, our MSNM can degenerate to other RC models. For instance, the MSNM can be regarded as the GQM when the upstream nonlinear filter $y(\cdot)$ is set as a

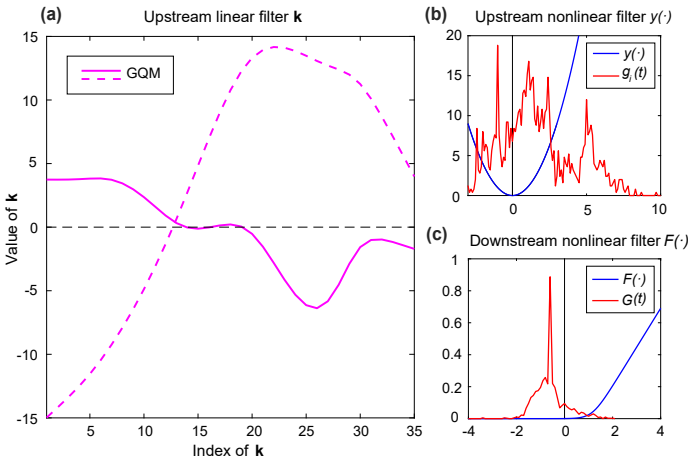


Fig. 5. Filters of the GQM. The (a) shows the upstream linear filter $\boldsymbol{k}$, the horizontal axis of (a) means the index of $\boldsymbol{k}$, and the vertical axis is its value. The (b) shows the upstream nonlinear filter $y(\cdot)$ and the histogram of $g_i(t)$. The (c) is the downstream nonlinear filter and the histogram of the generation function $G(t)$.

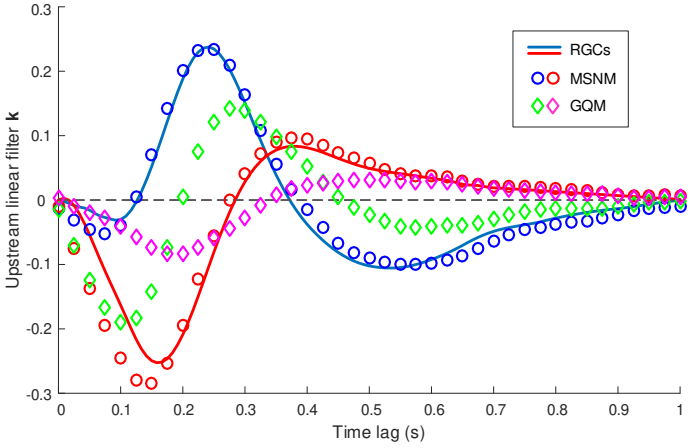


Fig. 6. Upstream linear filters $\boldsymbol{k}$ for different models. Solid lines show the filtering results of biological RGCs.

quadratic term, and when the downstream nonlinear function $F[\cdot]$ is discarded, the MSNM can be the LN model.

D. Performance Comparison

To compare the coding results and prediction accuracy of different models, we use k-fold cross validation. Fig. 7 evaluates the prediction performance in terms of the cross-correlation vs. log-likelihood LL. The horizontal axis gives the LL values, while the vertical axis means the cross-correlation values. The higher the cross-correlation and LL are, the better the prediction performance of RC models is. Intuitively, the more the markers in Figure 7 are distributed in the upper right, the better the prediction performance of the RC models is. According to the neuron populations, we can see that the proposed MSNM is remarkably accurate and powerful in visual coding compared to other models.

The spike raster and prediction firing rate of different models are shown in Fig. 8, in which 'RGCs' shows the biological firing records of retinal ganglion cells when a nature video is given [31]. Compared with the biological firing rate

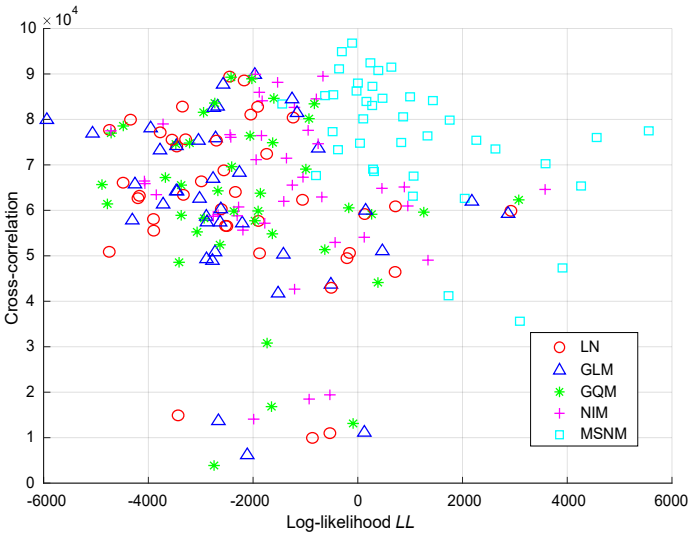


Fig. 7. The comparison results of the cross-correlation vs. log-likelihood LL for different RC models. Each marker shows the values of the cross-correlation and log-likelihood LL of each neuron. For example, the red circles show the prediction values of the 34 neurons in the LN model.

curve plotted by the solid red line in Fig. 8, the RC models successfully mimic the spike responses of neurons, as shown in the colored solid lines. However, retinal ganglion cells keep rest most of the frame times, which means the spike activation is sparse. When the frame time is bigger than 1300, the retinal ganglion cells hardly generate any firing spikes, i.e., there are few black dots in the RGCs region in Fig. 8. It can be seen that our MSNM exhibits such sparseness, leading to a good consistency with the biological firing records. This is because our MSNM designs the upstream linear filter with two parallel subunits which can capture the excitation and inhibitions of neurons. But other models still generate occasional spikes when RGCs keep rest, which causes a bias between actual and estimation results.

In order to further assess the performance of RC models in predicting the visual response characteristics of ganglion cells, we calculate the skewness values of the neuron firing rate curves of 102 neurons according to Eq. (17). Fig. 9 plots the skewness curves for different retinal coding models, in which the black line gives the skewness curve of the biological retinal ganglion cells. It can be seen that the proposed MSNM model is much closer to the RGCs result than its rivals, benefiting from the multi-modal stimulus matrix and excitation/inhibition filters in the MSNM.

In the retina, the visual coding of different ganglion cells has horizontal interaction. To find the correlation between neurons' sparse spiking, we use a Clayton copula function to construct the joint distribution of RGCs. As shown in Fig. 10, the Clayton copula function has a 'heavy tail' sensitive to covariates (u, v). The parameter ζ identifies how the (u, v) has a synchronous excitation/rest state. The higher the ζ is, the stronger the synchronicity of (u, v) is. Fig. 11 shows the Clayton copula function parameter ζ vs. the correlation coefficient. We can see that the results of other RC models cluster in the regions with ζ values less than 6. In contrast, our MSNM model is similar to biological ganglion cell responses and

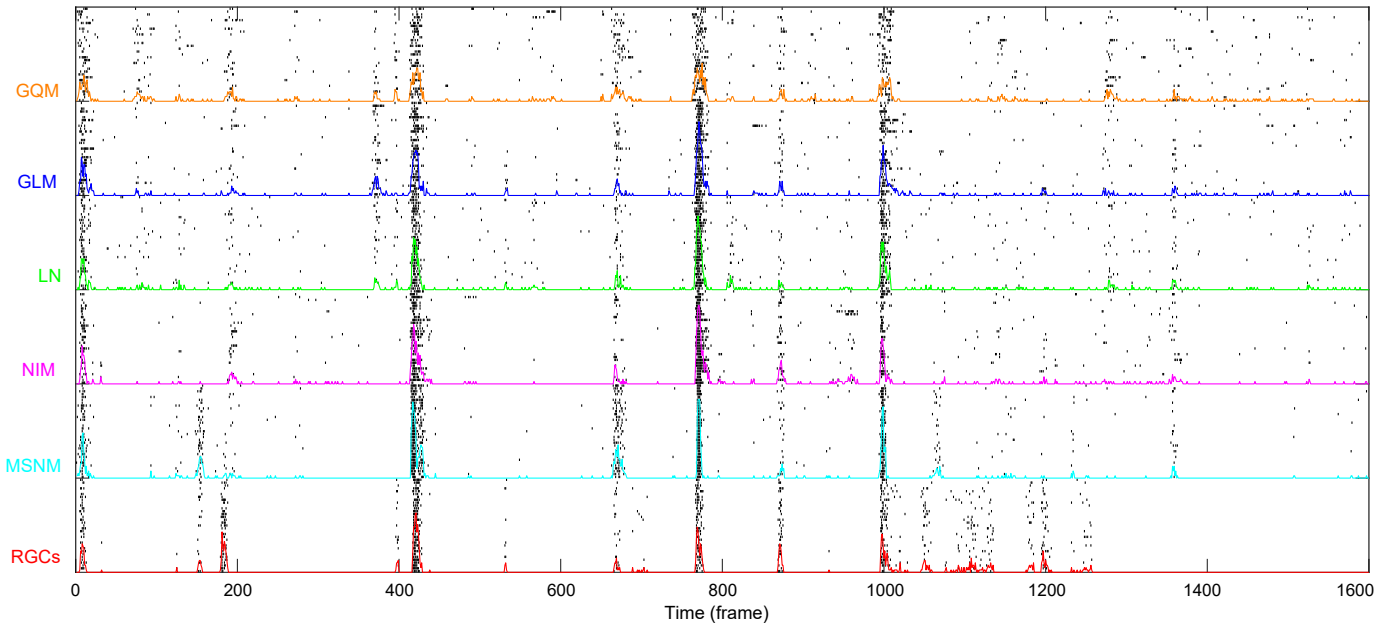


Fig. 8. The spike raster and prediction firing rate of different RC models. In this experiment, the natural input video is repeatedly played 33 times, and the 'RGCs' gives the biological firing records of retinal ganglion cells when a nature video is given [31]. The bottom 33 raster lines represent the impulse response of biological neurons to the video. The black dots show the firing spikes of the neurons, and the solid red line means the average firing rate curve of ganglion cells. From the bottom to the top, the simulation output results and firing rate curves of different RC models are plotted, in which the colored solid lines show the average firing rate curves of different RC models, respectively.

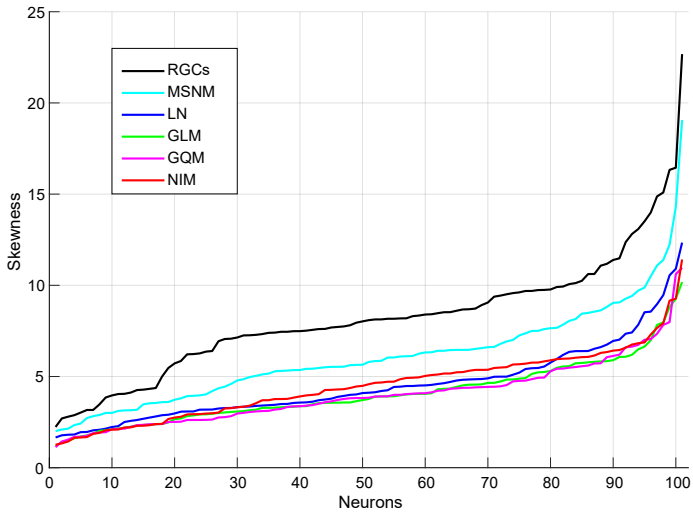


Fig. 9. The skewness values of 102 neurons for different RC models. The black line gives the skewness curve of the biological retinal ganglion cells.

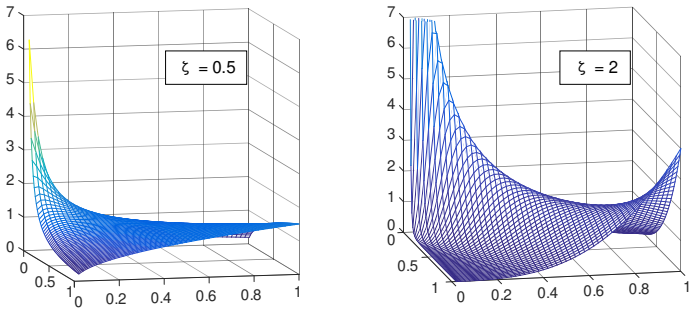


Fig. 10. Clayton distribution density. It can be divided into asynchronous and synchronous regions. The synchronous region is near the axis $y = x$, which means the spikes/rests of two neurons are relevant. The asynchronous spikes scatter in the region $(1, 0)$ or $(0, 1)$. A Clayton function with a high ζ has a lower value in the asynchronous region and a ridgeline in $y = x$. The higher the ζ is, the stronger the correlation is.

better reflects the strong and weak distribution characteristics of interactions between neurons.

IV. Conclusion

To improve the adaptability of retinal coding models in natural and complex scenes, this paper presents a nonlinear retinal coding model based on multi-modal stimulus input. It extracts the spatiotemporal features in the natural video and compiles the features by using basis functions to construct a multi-modal stimulus matrix. By employing the upstream linear subunits, the proposed model mimics the excitation-inhibition balance mechanism in biological neurons. And the nonlinear calculation effectively activates the generation function to generate firing spikes. Experimental results show our model is a bio-interpretable coding model and can accurately reproduce the internal biological computing properties of biological ganglion cells.

V. Acknowledge

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] Z. Yu, J. Liu, S. Jia, Y. Zhang, Y. Zheng, Y. Tian, and T. Huang, "Toward the next generation of retinal neuroprosthesis: Visual computation with spikes," *Engineering*, vol. 6, no. 4, pp. 449–461, Apr. 2020.

[2] S. Baccus and M. Meister, "Fast and slow contrast adaptation in retinal circuitry," *Neuron*, vol. 36, no. 5, pp. 909–919, Dec. 2002.

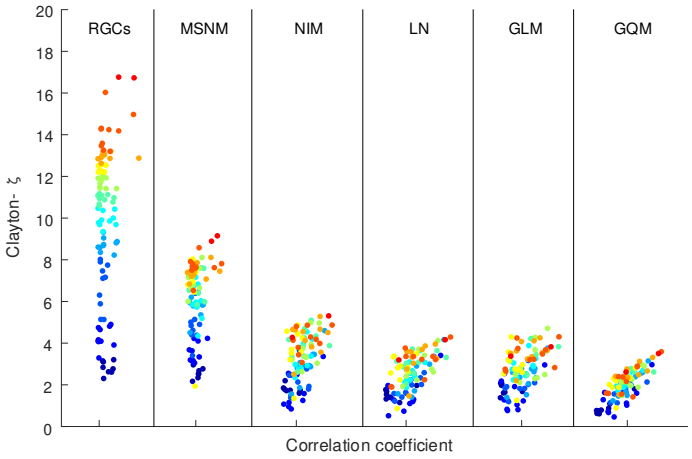


Fig. 11. Clayton copula function parameter ζ vs. the correlation coefficient. Each point is located by the correlation coefficient on the horizontal axis and the Clayton parameter ζ on the vertical axis.

[3] S. Kuo, G. Schwartz, and F. Rieke, "Nonlinear spatiotemporal integration by electrical and chemical synapses in the retina," *Neuron*, vol. 90, no. 2, pp. 320–332, Apr. 2016.

[4] J. Liu, H. Schreyer, A. Onken, F. Rozenblit, M. Khani, V. Krishnamoorthy, S. Panzeri, and T. Gollisch, "Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization," *Nature Communications*, vol. 8, no. 1, pp. 1–14, Jul. 2017.

[5] T. Gollisch, "Features and functions of nonlinear spatial integration by retinal ganglion cells," *Journal of Physiology - Paris*, vol. 107, no. 5, pp. 338–348, Nov. 2013.

[6] M. Carandini, D. Heeger, and J. Movshon, "Linearity and normalization in simple cells of the macaque primary visual cortex," *Journal of Neuroscience*, vol. 17, no. 21, pp. 8621–8644, Nov. 1997.

[7] J. Pillow, L. Paninski, V. Uzzell, E. Simoncelli, and E. Chichilnisky, "Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model," *Journal of Neuroscience*, vol. 25, no. 47, pp. 11 003–11 013, Nov. 2005.

[8] I. Park and J. Pillow, "Bayesian spike-triggered covariance analysis," *Advances in Neural Information Processing Systems*, vol. 24, pp. 1692–1700, Dec. 2011.

[9] T. Badea and J. Nathans, "Morphologies of mouse retinal ganglion cells expressing transcription factors brn3a, brn3b, and brn3c: Analysis of wild type and mutant cells using genetically-directed sparse labeling," *Vision Research*, vol. 51, no. 2, pp. 269–279, Jan. 2011.

[10] T. Baden, P. Berens, K. Franke, M. Román Rosón, M. Bethge, and T. Euler, "The functional diversity of retinal ganglion cells in the mouse," *Nature*, vol. 529, no. 7586, pp. 345–350, Jan. 2016.

[11] B. Völgyi, S. Chheda, and S. Bloomfield, "Tracer coupling patterns of the ganglion cell subtypes in the mouse retina," *Journal of Comparative Neurology*, vol. 512, no. 5, pp. 664–687, Feb. 2009.

[12] G. Zeck and R. Masland, "Spike train signatures of retinal ganglion cell types," *European Journal of Neuroscience*, vol. 26, no. 2, pp. 367–380, May. 2007.

[13] I. Park, M. Meister, A. Huk, and J. Pillow, "Encoding and decoding in parietal cortex during sensorimotor decision-making," *Nature Neuroscience*, vol. 17, no. 10, pp. 1395–1403, Aug. 2014.

[14] A. Heitman, N. Brackbill, M. Greschner, A. Sher, and E. Chichilnisky, "Testing pseudo-linear models of responses to natural scenes in primate retina," *BioRxiv*, p. 045336, Mar. 2016.

[15] N. Maheswaranathan, D. Kastner, S. Baccus, and S. Ganguli, "Inferring hidden structure in multilayered neural circuits," *PLoS Computational Biology*, vol. 14, no. 8, p. e1006291, Aug. 2018.

[16] Y. Zhang, S. Jia, Y. Zheng, Z. Yu, Y. Tian, S. Ma, T. Huang, and J. Liu, "Reconstruction of natural visual scenes from neural spikes with deep neural networks," *Neural Networks*, vol. 125, pp. 19–30, May. 2020.

[17] L. Mcintosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. Baccus, "Deep learning models of the retinal response to natural scenes," *Advances in Neural Information Processing Systems*, vol. 29, pp. 1369–1377, Jul. 2017.

[18] E. Batty, J. Merel, N. Brackbill, A. Heitman, A. Sher, A. Litke, E. Chichilnisky, and L. Paninski, "Multilayer recurrent network models of primate retinal ganglion cell responses," Mar. 2016.

[19] F. Sinz, A. Ecker, P. Fahey, E. Walker, E. Cobos, E. Froudarakis, D. Yatsenko, Z. Pitkow, J. Reimer, and A. Tolias, "Stimulus domain transfer in recurrent models for large scale cortical population prediction on video," *Advances in Neural Information Processing Systems*, vol. 31, pp. 7199–7210, Dec. 2018.

[20] Y. Zheng, S. Jia, Z. Yu, J. Liu, and T. Huang, "Unraveling neural coding of dynamic natural visual scenes via convolutional recurrent neural networks," *Patterns*, vol. 2, no. 10, p. 100350, Sep. 2021.

[21] J. Liu, D. Karamanlis, and T. Gollisch, "Simple model for encoding natural images by retinal ganglion cells with nonlinear spatial integration," *PLoS Computational Biology*, vol. 18, no. 3, p. e1009925, Mar. 2022.

[22] R. Whitney, "Most efficient quantum thermoelectric at finite power output," *Physical review letters*, vol. 112, no. 13, p. 130601, Apr. 2014.

[23] H. Kekre, S. Sahasrabudhe, and N. Goyal, "Raised cosine function for image data interpolation," *Computers & Electrical Engineering*, vol. 9, no. 3-4, pp. 131–152, May. 1982.

[24] J. jcbyts, "Mtlipglm," https://github.com/jcbyts/mtlipglm, 2020.

[25] J. McFarland, Y. Cui, and D. Butts, "Inferring nonlinear neuronal computation based on physiologically plausible inputs," *PLoS Computational Biology*, vol. 9, no. 7, p. e1003143, Jul. 2013.

[26] L. Paninski, "Maximum likelihood estimation of cascade point-process neural encoding models," *Network: Computation in Neural Systems*, vol. 15, no. 4, pp. 243–262, Sep. 2004.

[27] D. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, Jan. 1997.

[28] C. Knox, "Cross-correlation functions for a neuronal model," *Biophysical Journal*, vol. 14, no. 8, pp. 567–582, Aug. 1974.

[29] R. Sharma and R. Bhandari, "Skewness, kurtosis and newton's inequality," *The Rocky Mountain Journal of Mathematics*, vol. 45, no. 5, pp. 1639–1643, Dec. 2015.

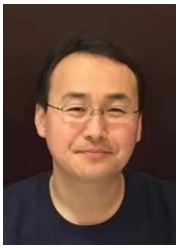
[30] H. Bekrizadeh, G. Parham, and M. Zadkarmi, "Weighted clayton copulas and their characterizations: application to probable modeling of the hydrology data," *Journal of Data Science*, vol. 11, no. 2, pp. 293–303, Apr. 2013.

[31] A. Onken, J. Liu, P. Karunasekara, I. Delis, T. Gollisch, and S. Panzeri, "Using matrix and tensor factorizations for the single-trial analysis of population spike trains," *PLoS Computational Biology*, vol. 12, no. 11, p. e1005189, Nov. 2016.

Lingling An received the B.S. and M.S. degrees in computer science and technology and the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2002, 2005, and 2011, respectively. She is currently a Professor with the School of Computer Science and Technology of Xidian University. Her research interests include neural computation and multimedia processing.

Zhen Yan received the B.S. degree in geophysics from Nanjing University of China, Nanjing, China, in 2017, and the M.S. degree in computer science and technology from Xidian University, Xi'an, China, in 2021. His research interests include machine learning and computational neuroscience.

Weizheng Wang received the B.S. degree in software engineering from Yangzhou University, Yangzhou, China, in 2019, the M.S. degree in computer science and engineering from the University of Aizu, Aizu-Wakamatsu, Japan, in 2021. He is currently working toward the Ph.D. degree in computer science with the Department of Computer Science, City University of Hong Kong, Hong Kong. He is currently a Research Associate with the University of Aizu. His research interests include applied cryptography, IoT system, and machine learning.



Jian K. Liu received the Ph. D. degree in mathematics from University of California t Los Angeles, USA, in 2009. He is currently a Lecturer with School of Computing, University of Leeds, UK. His research interests include computational neuroscience and brain-inspired computing.

Keping Yu (M'17) received the M.E. and Ph.D. degrees from the Graduate School of Global Information and Telecommunication Studies, Waseda University, Tokyo, Japan, in 2012 and 2016, respectively. He was a Research Associate, Junior Researcher, Researcher with the Global Information and Telecommunication Institute, Waseda University, from 2015 to 2019, 2019 to 2020, 2020 to 2022, respectively. He is currently an associate professor at Hosei University and a visiting scientist at the RIKEN Center for Advanced Intelligence Project, Japan. Dr. Yu has hosted and participated in more than ten projects, is involved in many standardization activities organized by ITU-T and ICNRG of IRTF, and has contributed to ITU-T Standards Y.3071 and Supplement 35. He received the IEEE Outstanding Leadership Award from IEEE BigDataSE 2021, the Best Paper Award from IEEE Consumer Electronics Magazine Award 2022 (1st Place Winner), IEEE ICFTIC 2021, ITU Kaleidoscope 2020, the Student Presentation Award from JSST 2014. He has authored 100+ publications including papers in prestigious journal/conferences such as the IEEE WCM, CM, NetMag, IoTJ, TFS, TII, T-ITS, TVT, TMC, JBHI, TR, TCOM, TNSM, TIM, TNSE, TGCN, TCSS, CEM, IOTM, ICC, GLOBECOM etc. He is an Associate Editor of IEEE Open Journal of Vehicular Technology, Journal of Intelligent Manufacturing, Journal of Circuits, Systems and Computers. He has been a Guest Editor for more than 20 journals such as IEEE TCSS. He served as general co-chair and publicity co-chair of the IEEE VTC2020-Spring 1st EBTSRA workshop, general co-chair of IEEE ICCC2020 2nd EBTSRA workshop, general co-chair of IEEE TrustCom2021 3nd EBTSRA workshop, session chair of IEEE ICCC2020, ITU Kaleidoscope 2016. His research interests include smart grids, information-centric networking, the Internet of Things, artificial intelligence, blockchain, and information security.