Massively Parallel Neural Circuits for Stereoscopic Color Vision: Encoding, Decoding and Identification

Aurel A. Lazar[a,1,*], Yevgeniy B. Slutskiy[a,1], Yiyin Zhou[a,1]

[a] *Department of Electrical Engineering, Columbia University, New York, NY, USA*

Abstract

Past work demonstrated how monochromatic visual stimuli could be faithfully encoded and decoded under Nyquist-type rate conditions. Color visual stimuli were then traditionally encoded and decoded in multiple separate monochromatic channels. The brain, however, appears to mix information about color channels at the earliest stages of the visual system, including the retina itself. If information about color is mixed and encoded by a common pool of neurons, how can colors be demixed and perceived?

We present Color Video Time Encoding Machines (Color Video TEMs) for encoding color visual stimuli that take into account a variety of color representations within a single neural circuit. We then derive a Color Video Time Decoding Machine (Color Video TDM) algorithm for color demixing and reconstruction of color visual scenes from spikes produced by a population of visual neurons. In addition, we formulate Color Video Channel Identification Machines (Color Video CIMs) for functionally identifying color visual processing performed by a spiking neural circuit.

Furthermore, we derive a duality between TDMs and CIMs that unifies the two and leads to a general theory of neural information representation for

*Corresponding author. Department of Electrical Engineering, Columbia University, 500 W. 120th Street, New York, NY 10027, USA. Tel: +1 212-854-1747.

Email addresses: aurel@ee.columbia.edu (Aurel A. Lazar), yevgeniy@ee.columbia.edu (Yevgeniy B. Slutskiy), yiyin@ee.columbia.edu (Yiyin Zhou)

[1] The author's names are listed in alphabetical order.

stereoscopic color vision. We provide examples demonstrating that a massively parallel color visual neural circuit can be first identified with arbitrary precision and its spike trains can be subsequently used to reconstruct the encoded stimuli. We argue that evaluation of the functional identification methodology can be effectively and intuitively performed in the stimulus space. In this space, a signal reconstructed from spike trains generated by the identified neural circuit can be compared to the original stimulus.

Keywords: stereoscopic color vision, massively parallel neural circuits, time encoding machines, time decoding machines, channel identification machines.

---

1. Introduction

The sensation of light for many animals is primarily due to two of its properties. Light intensity provides information about the brightness of the scene and the shape of objects, while its wavelength provides information about the color. Although it has long been known that the separation of color space in humans and some of the primates is mediated by three types of cone photoreceptors having peak sensitivity at different wavelengths (roughly corresponding to red, green and blue light), the exact nature of color processing and representation downstream of photoreceptors remains to be elucidated (Dacey, 2000; Solomon & Lennie, 2007).

The study of color vision is complicated by the fact that the early visual system employs a large number of neurons to process and represent visual stimuli. This processing involves a variety of cell types, complex wiring of neurons into canonical circuits and a large number of outputs (Masland, 2012; Gollisch & Meister, 2010; da Silveira & Roska, 2011; Lazar et al., 2013). Furthermore, neurons in the early visual system appear to combine signals from different types of cones, thereby mixing colors on the level of individual cells so that a spike train of a single cell carries information about multiple colors present in the visual scene (Dacey, 2000; Solomon & Lennie, 2007). For example, retinal ganglion cells (RGCs) exhibit opponent channel processing, whereby differences between the responses of cones such as red vs. green or blue vs. yellow or black vs. white is encoded.

How is color information represented in the activity of different types of visual

neurons? Can we identify the color mixing performed by any given neuron? Can the color information be demixed once it is encoded by spiking neurons into an unlabeled set of spike trains?

In this paper, we approach these questions by proposing and studying a novel class of massively parallel neural circuits with spiking neurons that model the encoding of color visual scenes in early visual systems. For this novel class of circuits, we investigate three different but interrelated problems: encoding, decoding and functional identification.

Encoding describes the transformation of sensory inputs into neural codes, *e.g., spike trains*, by the sensory system. In a closely related setting of analog to digital conversion, the encoding procedure maps, or transforms, analog input signals into discrete amplitude sequences.

Decoding is the process of inverting the transformation established by the encoder. In other words, it is the process of reconstructing the sensory input from the spikes generated by the encoder. While decoding may not take place at every stage in the brain, it provides a quantification of the amount of information encoded by the neural circuit under investigation.

Functional identification aims at inferring the transformation performed by a neural circuit or system. In the setting of functional identification, the inputs are known and the outputs are observable, while the functionality of the neural circuit is unknown. Functional identification is critical in understanding the information processing taking place in biological systems and is operationally significant to decipher what a neural circuit does.

In this paper, we apply formal methods seeking the unification of encoding, decoding and functional identification algorithms operating on massively parallel neural circuits that share a common underlying representation of visual information in the spike domain. Our formal methods are based on and extend upon the rigorous theory of Time Encoding Machines (TEMs), Time Decoding Machines (TDMs) and Channel Identification Machines (CIMs).

TEMs are asynchronous nonlinear circuits that encode analog signals into multi-dimensional spike trains (Lazar & Tóth, 2004; Lazar, 2004). They naturally arise as models of early sensory systems. For a general TEM of interest, the inputs are first filtered by a (linear) receptive field and then fed to a spike generation mechanism (Lazar & Pnevmatikakis, 2008, 2011). The spike generators can be viewed as asynchronous samplers, while the receptive

fields can be considered akin to communication channels. Receptive fields may be implemented by complex neural circuits in cascade with dendritic trees feeding into the spike generator. TEMs have been investigated for both single neuron and population configurations, with applications to multiple sensory modalities, *e.g.*, audition and vision (Lazar & Pnevmatikakis, 2008, 2011).

TDMs reconstruct time encoded analog signals from spike trains (Lazar & Tóth, 2004; Lazar, 2004). Each TDM is a realization of an algorithm that recovers the analog signal from the output of its TEM counterpart. TDMs play important roles in understanding the encoding as well, since the decodability of spike trains quantifies how much information about the encoded stimuli is preserved in the spike sequence.

Functional identification of a neural circuit is formulated here in the setting of CIMs (Lazar & Slutskiy, 2012). CIMs are algorithms that identify neural circuit parameters (e.g., receptive fields and parameters of spike generators) directly from spike times generated in response to a collection of test stimuli. Some of the key advantages of parameter identification in the setting of CIMs are (i) clear lower bounds on the number of test stimuli and spikes required for identification can be specified and (ii) both synthetic and naturalistic stimuli can be used (Lazar & Slutskiy, 2014a,b).

TEM, TDM and CIM algorithms are all based on spike time sequences instead of average spike rates. They are versatile and have been applied in many contexts, including, neural circuits encoding time-varying stimuli, natural visual stimuli and multi-sensory circuits (Lazar & Pnevmatikakis, 2008; Lazar et al., 2010; Lazar & Slutskiy, 2013). However, they have typically been used separately, within their own context. Here we bridge the theoretical formalism of TEMs, TDMs and CIMs in the context of color vision.

Previously, Video TEMs have been considered for mono-chromatic videos, e.g., gray-scale visual stimuli (Lazar & Pnevmatikakis, 2011; Lazar et al., 2010). Encoding of color visual stimuli was modeled as mono-chromatic video encoding in separate color channels (Lazar & Zhou, 2012). We present in this paper a Color Video TEM for modeling the encoding of color visual stimuli that takes into account a variety of color representations in early visual systems within a single neural circuit. A decoding (color demixing) algorithm is devised that reconstructs color videos from a population of spike trains. In addition, we formulate CIMs for functionally identifying receptive

fields that carry visual information from multiple color channels. This allows us to identify the spatio-temporal structure of the receptive fields in each of the base color channels. We also demonstrate that demixing and identification results hold in full generality for Stereoscopic Color Video TEMs and CIMs, respectively.

Moreover, we extend the duality between TDMs and CIMs (Lazar & Slutskiy, 2012) and thereby obtain a general theory of neural representation of color visual information. We provide examples to demonstrate that a massively parallel color video encoding neural circuit of unknown parameters can be identified with arbitrary precision and its spike trains can be used to reconstruct the encoded stimuli.

Given the extended duality between TEMs and CIMs, we argue that evaluation of the functional identification methodology can be effectively and intuitively performed in the stimulus space. In this space the reconstructed signals using spike trains generated by an identified neural circuit can be compared to the original stimulus.

This paper is organized as follows. In section 2, we model the space of color visual stimuli as a vector-valued Reproducing Kernel Hilbert Space (RKHS). We formulate the encoding/decoding problem of color visual stimuli and provide an algorithm for stimulus recovery. We also provide rigorous methods for identifying circuits in the early visual system. In section 3 we present an unified framework for quantitatively evaluating identification algorithms of massively parallel neural circuits in the stimulus space and demonstrate the use of decoding algorithms as a means of model verification. Extensions to encoding, decoding and identification of stereoscopic color vision is dealt with in section 4. Finally, section 5 provides a discussion of the results and their implication, and future work.

2. Massively Parallel Neural Circuits for Color Vision

The majority of existing neural computation models employ grayscale or monochrome stimuli to test and describe the computation performed by neural circuits of the visual system. An exception is Lazar & Zhou (2012) where it is demonstrated how to encode an RGB (Red, Green and Blue) video stream for each color using separate, dedicated circuits for each color component.

In this section we present a novel approach that formally models the mixing of color information within a single neuron level, e.g., as performed by the retina. We begin by describing how both synthetic and natural grayscale stimuli can be effectively modeled as elements of a scalar-valued Reproducing Kernel Hilbert Space (RKHS) (Berlinet & Thomas-Agnan, 2004). We then extend the space of stimuli to a vector-valued RKHS to handle both color and stereoscopic visual stimuli. Due to space constraints, we discuss video stimuli only. However, images can be handled in a similar fashion.

2.1. Modeling Color Visual Stimuli

We model monochromatic visual stimuli $u = u(x, y, t), (x, y, t) \in \mathbb{D}$, as elements of an RKHS $\mathcal{H}$. The elements of the space $\mathcal{H}$ are scalar valued functions defined over the space-time domain $\mathbb{D} = \mathbb{D}_x \times \mathbb{D}_y \times \mathbb{D}_t$, where $\mathbb{D}_x = [0, T_x]$, $\mathbb{D}_y = [0, T_y]$ and $\mathbb{D}_t = [0, T_t]$, with $T_x, T_y, T_t \in \mathbb{R}_+$. The scalar functions represent the intensity of light at a particular point in a two-dimensional space (x, y) at time t.

For practical and computational reasons we choose to work with spaces of trigonometric polynomials. However, the theory developed below is quite general and applies to many other RKHSs (examples include Sobolev spaces and Paley-Wiener spaces; see Berlinet & Thomas-Agnan (2004) for an extensive list of alternatives). Each element $u \in \mathcal{H}$ is of the form

$$u(x, y, t) = \sum_{l_x=-L_x}^{L_x} \sum_{l_y=-L_y}^{L_y} \sum_{l_t=-L_t}^{L_t} c_{l_x l_y l_t} e_{l_x l_y l_t}(x, y, t), \tag{1}$$

with

$$e_{l_x l_y l_t}(x, y, t) = \frac{1}{\sqrt{T_x T_y T_t}} \exp\left[j \left(\frac{l_x \Omega_x x}{L_x} + \frac{l_y \Omega_y y}{L_y} + \frac{l_t \Omega_t t}{L_t} \right) \right], \tag{2}$$

where j denotes the imaginary number and L_x, L_y, L_t represent the order of the space $\mathcal{H}$ in each corresponding variable. The elements of $\mathcal{H}$ are periodic bandlimited functions with bandwidths Ω_x, Ω_y and Ω_t in space and in time, respectively. The period in each variable is associated with the space-time domain and is defined as

$$T_x = \frac{2\pi L_x}{\Omega_x}, T_y = \frac{2\pi L_y}{\Omega_y}, T_t = \frac{2\pi L_t}{\Omega_t}. \tag{3}$$

As already mentioned, the choice of the space of trigonometric polynomials was motivated by the ease of representation and the associated computational efficiency. In addition, the concept of bandwidth naturally arises from this representation and will become important in devising bounds for decoding and identification.

$\mathcal{H}$ is endowed with the inner product defined by

$$\langle u, w\rangle_{\mathcal{H}} = \int_{\mathbb{D}} u(x,y,t)\overline{w(x,y,t)}dxdydt, \tag{4}$$

where $\overline{w}$ denotes the complex conjugate of w. It is easy to see that the set of functions

$$\{e_{l_x l_y l_t} \,|\, l_x = -L_x, ..., L_x;\; l_y = -L_y, ..., L_y;\; l_t = -L_t, ..., L_t\}$$

forms an orthonormal basis in $\mathcal{H}$. The reproducing kernel of $\mathcal{H}$ is a function given by $K : \mathbb{D} \times \mathbb{D} \to \mathbb{R}$, where

$$K(x,y,t;x',y',t') = \sum_{l_x=-L_x}^{L_x} \sum_{l_y=-L_y}^{L_y} \sum_{l_t=-L_t}^{L_t} e_{l_x l_y l_t}(x-x', y-y', t-t') \tag{5}$$

satisfies the reproducing property

$$\langle u, K_{xyt}\rangle_{\mathcal{H}} = u(x,y,t), \text{for all } u \in \mathcal{H}, \tag{6}$$

and $K_{xyt}(x',y',t') = K(x,y,t;x',y',t')$. The RKHS $\mathcal{H}$ is very effective in modeling both synthetic and natural stimuli (Lazar et al., 2010). Moreover, it allows for (i) easy interpretation of the encoding of color visual stimuli as generalized sampling, (ii) easy derivation of sampling functions, and (iii) derivation of a reconstruction algorithm that solves a spline interpolation problem.

In what follows we denote the total dimension of $\mathcal{H}$ by $dim(\mathcal{H}) = (2L_x + 1)(2L_y + 1)(2L_t + 1)$, the spatial dimension of $\mathcal{H}$ by $dim_{xy}(\mathcal{H}) = (2L_x + 1)(2L_y + 1)$ and the temporal dimension by $dim_t(\mathcal{H}) = 2L_t + 1$. Clearly, $dim(\mathcal{H}) = dim_{xy}(\mathcal{H}) \cdot dim_t(\mathcal{H})$.

Color is the perception of the wavelength of light. Here we consider a discrete representation of wavelength, which is naturally provided by multiple types of cone photoreceptors having different peak spectral sensitivities. For example,

it is well known that the trichromacy in human vision arises as a result of the visual space being sampled by three different kinds of photoreceptors at the very first stage of the visual system. Specifically, the L-, M-, and S-cones of the retina provide an initial representation of the visual space in terms of the red, green, and blue color channels, respectively. Subsequent processing within and across these color channels affords enhanced scene segmentation, visual memory, as well as perception and recognition of objects and faces (Russell & Sinha, 2007; Gegenfurtner, 2003; Gegenfurtner & Rieger, 2000).

We now extend the problem setting presented above to the space of color stimuli. Without loss of generality, we assume the traditional red, green and blue, or RGB, color representation. We assume that color visual stimuli are elements of the space of trigonometric polynomials. Each visual stimulus $\mathbf{u}$ is a vector-valued function $\mathbf{u} : \mathbb{D} \to \mathbb{R}^3$ of the form

$$\mathbf{u}(x, y, t) = [u_1(x, y, t), u_2(x, y, t), u_3(x, y, t)]^T, \tag{7}$$

where each of the component functions u_1 (red channel), u_2 (green channel) and u_3 (blue channel) is a scalar-valued function in the RKHS $\mathcal{H}$. As the space we have in mind is a direct sum of three orthogonal spaces $\mathcal{H}$, we denote this color visual stimulus space as $\mathcal{H}^3$. For simplicity, we assume that the bandwidth and order of each of the considered subspaces are the same. By construction, the space $\mathcal{H}^3$ is endowed with the inner product

$$\langle \mathbf{u}, \mathbf{w} \rangle_{\mathcal{H}^3} = \sum_{m=1}^{3} \langle u_m, w_m \rangle_{\mathcal{H}}. \tag{8}$$

RKHSs with vector-valued function elements have been studied in depth (see Carmeli et al. (2006); Caponnetto et al. (2008) and reference within) and the reproducing kernel of $\mathcal{H}^3$ is given by $\mathbf{K} : \mathbb{D} \times \mathbb{D} \to \mathbf{M}(3, \mathbb{C})$, where $\mathbf{M}(3, \mathbb{C})$ is the space of 3×3 matrices (bounded linear operators on $\mathbb{R}^3$) given by

$$\mathbf{K} = \begin{bmatrix} K_1 & 0 & 0 \\ 0 & K_2 & 0 \\ 0 & 0 & K_3 \end{bmatrix}. \tag{9}$$

and K_m, $m = 1, 2, 3$, are reproducing kernels of $\mathcal{H}$ as in (5). The reproducing property of $\mathcal{H}^3$ is given by

$$\langle \mathbf{u}, \mathbf{K}_{x,y,t} \mathbf{v} \rangle_{\mathcal{H}^3} = \langle \mathbf{u}(x, y, t), \mathbf{v} \rangle_{\mathbb{R}^3}, \quad \text{for all } \mathbf{u} \in \mathcal{H}^3 \textit{ and } \mathbf{v} \in \mathbb{R}^3, \tag{10}$$

where $\mathbf{K}_{x,y,t}\mathbf{v} \in \mathcal{H}^3$ and is defined as

$$\mathbf{K}_{x,y,t}\mathbf{v} = \mathbf{K}(x, y, t; x', y', t')\mathbf{v}. \tag{11}$$

From the above, it is easy to see that for a unit vector $\mathbf{e}_m \in \mathbb{R}^3$, $m = 1, 2, 3$,

$$u_m(x, y, t) = \langle \mathbf{u}, \mathbf{K}_{x,y,t}\mathbf{e}_m \rangle_{\mathbb{R}^3}. \tag{12}$$

Note that the dimension of $\mathcal{H}^3$ is $dim(\mathcal{H}^3) = 3 \cdot dim(\mathcal{H})$.

2.2. Neural Encoding Circuits for Color Vision

We now describe how the color visual stimuli of the previous section can be faithfully encoded into a multidimensional sequence of spikes by a population of spiking neurons. We employ a massively parallel neural circuit consisting of thousands of neurons, in which each neuron is a fundamentally slow device capable of producing only a limited number of spikes per unit of time. Our goal is to devise a set of conditions on the population of neurons that guarantees a faithful, or loss-free, representation of color visual stimuli in the spike domain.

Consider the massively parallel neural circuit shown in Fig. 1. The color visual stimulus $\mathbf{u}$ consists of 3 components u_1, u_2, u_3, as in (7). These components (corresponding to the red, green, and blue channel, respectively) are assumed to be extracted by the photoreceptors and subsequently encoded by a population of N neurons. In the most general case, all neurons receive information from each photoreceptor type and multiplex (mix) and encode that information in the spike domain. Specifically, each neuron i is associated with a multi-component linear receptive field, or kernel, $\mathbf{h}^i(x, y, t)$, where

$$\mathbf{h}^i(x, y, t) = \left[h_1^i(x, y, t), h_2^i(x, y, t), h_3^i(x, y, t)\right]^T. \tag{13}$$

The components $h_m^i(x, y, t), m = 1, 2, 3, i = 1, ..., N$, are assumed to be causal in the time domain $\mathbb{D}_t$ and have a finite support in the spatial domains $\mathbb{D}_x$ and $\mathbb{D}_y$. In addition, we assume that all components of the kernel are bounded-input bounded-output (BIBO) stable. Therefore, the component filters belong to the filter kernel space $H = \mathbb{L}^1(\mathbb{D})$, where $\mathbb{L}^1(\mathbb{D})$ denotes the space of absolute integrable functions on $\mathbb{D}$.

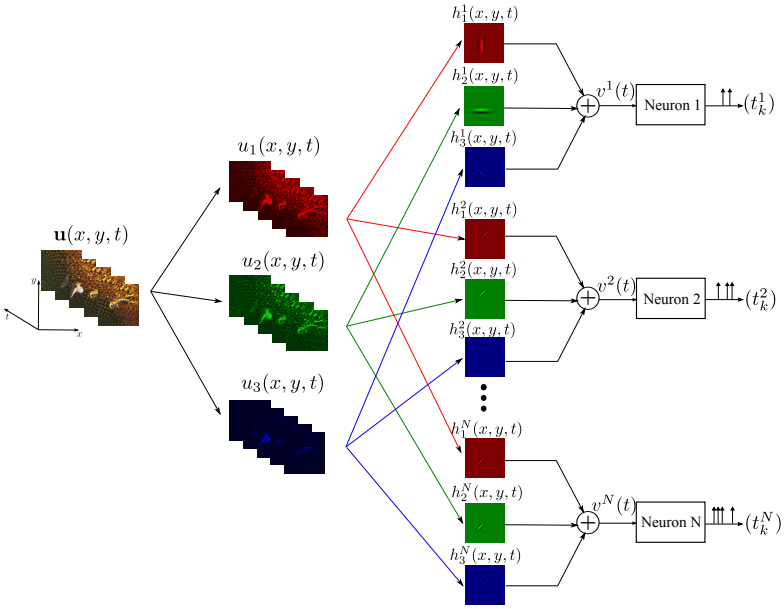


Figure 1: Diagram of the Encoding Neural Circuit for Color Visual Stimuli.

For every neuron i, each color channel $u_m(x,y,t)$, $m = 1, ..., 3$, of the input signal $\mathbf{u}$ is independently filtered by the corresponding component $h_m^i(x,y,t)$ of the receptive field $\mathbf{h}^i(x,y,t)$, yielding a temporal signal

$$v_m^i(t) = \int_{\mathbb{D}} h_m^i(x,y,t-s)u_m(x,y,s)dxdyds, \quad m = 1,2,3. \tag{14}$$

The outputs of the three receptive field components are then summed to provide an aggregate temporal input $v^i(t)$ to the ith neuron that amounts to

$$v^i(t) = \sum_{m=1}^{3} v_m^i(t) = \sum_{m=1}^{3} \left(\int_{\mathbb{D}} h_m^i(x,y,t-s)u_m(x,y,s)dxdyds \right). \tag{15}$$

The three-dimensional color visual stimulus $\mathbf{u}(x,y,t)$ is effectively transformed into a one-dimensional signal $v^i(t)$, in which colors, spatial and temporal attributes of $\mathbf{u}$ are multiplexed. v^i is then encoded by the ith neuron into a spike train, with the sequence of spike times denoted by $(t_k^i)_{k\in\mathbb{Z}}$, where k is the spike index. The summation here can be justified by experiments in retina and V1 that have shown that the response of many neurons can be captured by a linear combination of cone signals (Gegenfurtner & Kiper, 2003; Solomon & Lennie, 2007).

For simplicity, we assume here that each point neuron is an Integrate-and-Fire (IAF) neuron, as illustrated in Figure 2. However, many other point neuron models, including conductance-based models such as Hodgkin-Huxley, Morris-Lecar, Fitzhugh-Nagumo, Wang-Buzsáki, and Hindmarsh-Rose, threshold-and-fire neurons and oscillators with both multiplicative and additive coupling can be considered as well (Lazar & Slutskiy, 2014b, 2012).

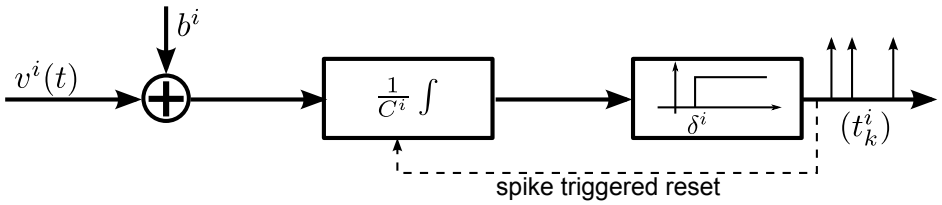


Figure 2: Block diagram of an (ideal) IAF neuron. The input v^i together with an additive bias b^i are passed through an integrator with integration constant C^i. A spike is generated whenever the integrator output reaches a threshold δ^i. The integrator is reset after every spike.

The IAF neuron i encodes its input v^i into the sequence of spike times t_k^i

$$\int_{t_k^i}^{t_{k+1}^i} v^i(s)ds = q_k^i, \quad k \in \mathbb{Z}, \tag{16}$$

where $q_k^i = C^i\delta^i - b^i(t_{k+1}^i - t_k^i)$. Here C^i, δ^i and b^i are the integration constant, threshold and bias, respectively, of the ith neuron. The encoding performed by the entire neural circuit can then be expressed by the following equations

$$\int_{t_k^i}^{t_{k+1}^i} \sum_{m=1}^{3} \left(\int_{\mathbb{D}} h_m^i(x,y,t-s)u_m(x,y,s)dxdyds \right) dt = q_k^i, \quad k \in \mathbb{Z}, \tag{17}$$

for all $i = 1, 2, ..., N$. By defining linear functionals $\mathcal{T}_k^i : \mathcal{H}^3 \to \mathbb{R}$, $i = 1, 2, ..., N, k \in \mathbb{Z}$, where

$$\mathcal{T}_k^i \mathbf{u} = \int_{t_k^i}^{t_{k+1}^i} \sum_{m=1}^{3} \left(\int_{\mathbb{D}} h_m^i(x,y,t-s)u_m(x,y,s)dxdyds \right) dt, \tag{18}$$

equation (17) can be compactly rewritten as

$$\mathcal{T}_k^i \mathbf{u} = q_k^i. \tag{19}$$

Called the t-transform (Lazar & Pnevmatikakis, 2011), equation (19) above describes the mapping of the analog signal $\mathbf{u}$ into a set of spikes $(t_k^i), i = 1, 2, ..., N, k \in \mathbb{Z}$.

By combining signals from different channels, each neuron may now carry different types of color information. For example, combining all three channels may provide luminance of the visual stimulus over a wide spectrum. Color opponency in the retina that typically takes the form of red versus green, blue versus yellow can be modeled as well.

The neural encoding circuit shown in Figure 1 is called the Color Video Time Encoding Machine (Color Video TEM). The Color Video TEM can be equally interpreted as a Multiple-Input Multiple-Output (MIMO) neural encoder, where $u_m, m = 1, 2, 3$, are seen as three separate inputs. By modeling the color video as a single element in $\mathcal{H}^3$, we highlight the fact that color is an intrinsic property of a natural visual stimulus.

Lemma 1 (The Geometry of Time Encoding)**.** *The Color Video TEM projects the stimulus* $\mathbf{u}$ *onto the set of sampling functions* $\boldsymbol{\phi}_k^i = [\phi_{1k}^i, \phi_{2k}^i, \phi_{3k}^i]^T$ *with* $\boldsymbol{\phi}_{mk}^i = \mathcal{T}_k^i \overline{\mathbf{K}_{xyt}\mathbf{e}_m}, m = 1, 2, 3$*, and*

$$\langle \mathbf{u}, \boldsymbol{\phi}_k^i \rangle_{\mathcal{H}^3} = q_k^i, \quad i = 1, 2, ..., N, \quad k \in \mathbb{Z}. \tag{20}$$

Proof: By the Riesz Representation Theorem (Berlinet & Thomas-Agnan, 2004), there exist functions $\boldsymbol{\phi}_k^i \in \mathcal{H}^3$ such that for all $\mathbf{u} \in \mathcal{H}^3$,

$$\mathcal{T}_k^i \mathbf{u} = \langle \mathbf{u}, \boldsymbol{\phi}_k^i \rangle_{\mathcal{H}^3}, \quad i = 1, 2, ..., N, \quad k \in \mathbb{Z}, \tag{21}$$

and therefore, the encoding of the color video $\mathbf{u}$ by the TEM can be expressed as

$$\langle \mathbf{u}, \boldsymbol{\phi}_k^i \rangle_{\mathcal{H}^3} = q_k^i, \quad i = 1, 2, ..., N, \quad k \in \mathbb{Z}.$$

The entries ϕ_{mk}^i of the sampling function $\boldsymbol{\phi}_k^i$ can be obtained by the reproducing property

$$\phi_{mk}^i(x, y, t) = \langle \boldsymbol{\phi}_k^i, \mathbf{K}_{xyt}\mathbf{e}_m \rangle = \overline{\langle \mathbf{K}_{xyt}\mathbf{e}_m, \boldsymbol{\phi}_k^i \rangle} = \mathcal{T}_k^i \overline{\mathbf{K}_{xyt}\mathbf{e}_m}, \quad m = 1, 2, 3. \tag{22}$$

□

In Appendix A, we demonstrate how ϕ_{mk}^i can be efficiently computed. Thus, similar to the monochrome video encoding (Lazar & Pnevmatikakis, 2011), the encoding of the color video has a simple geometrical interpretation as

sampling of $\mathbf{u}$ by a set of input dependent sampling functions (coordinates) $\boldsymbol{\phi}_k^i$, and the q_k^i, $k \in \mathbb{Z}$, are the corresponding measurements. The important difference is that the set of coordinates is not fixed. Rather, it changes at every spike time t_k^i.

2.3. Decoding Algorithms for Color Vision

2.3.1. Time Decoding Machines for Color Vision

Assuming that all receptive fields and parameters of the neurons are known, the decoding algorithm reconstructs the video $\mathbf{u}$ from the set of N spike trains $(t_k^i), i = 1, 2, ..., N$, $k = 1, 2, ..., n^i + 1$, produced by the encoding neural circuit, where $n^i + 1$ is the number of spikes generated by neuron i.

Given the assumption that $\mathbf{u} \in \mathcal{H}^3$, and the fact that encoding of the visual stimuli consists of projections of $\mathbf{u}$ onto a set of sampling functions, we formulate the reconstruction of the encoded video as the spline interpolation problem

$$\hat{\mathbf{u}} = \underset{\mathbf{u}\in\mathcal{H}^3, \{\mathcal{T}_k^i \mathbf{u} = q_k^i\}_{k=1,...,n^i}^{i=1,...,N}}{\operatorname{argmin}} \{\|\mathbf{u}\|_{\mathcal{H}^3}^2\}. \tag{23}$$

Theorem 1. *Let the color video $\mathbf{u} \in \mathcal{H}^3$ be encoded by the color Video TEM with N neurons and N linearly independent receptive fields. The color video can be reconstructed as*

$$\hat{\mathbf{u}} = \sum_{i=1}^{N} \sum_{k=1}^{n^i} c_k^i \boldsymbol{\phi}_k^i, \tag{24}$$

where the c_k^i's are the solution to the system of linear equations

$$\boldsymbol{\Phi}\mathbf{c} = \mathbf{q}, \tag{25}$$

with $\mathbf{c} = \left[c_1^1, c_2^1, ..., c_{n^1}^1, \cdots, c_1^N, c_2^N, ..., c_{n^N}^N\right]^T$, $\mathbf{q} = \left[q_1^1, q_2^1, ..., q_{n^1}^1, \cdots, q_1^N, q_2^N, ..., q_{n^N}^N\right]^T$, and $\boldsymbol{\Phi}$ is the block matrix

$$\boldsymbol{\Phi} = \left[\boldsymbol{\Phi}^{ij}\right], \tag{26}$$

where $i, j = 1, 2, ..., N$, and the block entries are given by (see also Appendix A)

$$\left[\boldsymbol{\Phi}^{ij}\right]_{kl} = \langle \boldsymbol{\phi}_k^i, \boldsymbol{\phi}_l^j \rangle_{\mathcal{H}^3}, \text{ for all } i, j = 1, 2, ..., N, \text{ and } k = 1, 2, ..., n^i, l = 1, 2 \cdots, n^j.$$

A necessary condition for perfect recovery of any arbitrary $\mathbf{u} \in \mathcal{H}^3$ *is that the set of sampling functions* $\boldsymbol{\phi} = (\boldsymbol{\phi}_k^i), k = 1, 2, ..., n^i, i = 1, 2, ..., N$, *span* $\mathcal{H}^3$. *This requires to have the number of neurons* $N \geq 3 \cdot dim_{xy}(\mathcal{H})$ *and a total of at least* $\dim(\mathcal{H}^3) + N$ *spikes.*

Proof: See Appendix B for a complete proof. □

Theorem 1 shows that color visual stimuli can be reconstructed by a linear combination of sampling functions. Moreover, color visual stimuli can be perfectly reconstructed only if the encoder generates enough spikes such that the sampling functions fully explore the input space.

Remark 1. *The solution of the optimization problem in Theorem 1 can easily be extended to an infinite dimensional space with* $\hat{\mathbf{u}}$ *reconstructed in the subspace generated by the span of the set of sampling functions* $(\boldsymbol{\phi}_k^i)$, $k = 1, 2, ..., n^i$, $i = 1, 2, ..., N$. *Moreover,* (25) *can be solved by a simple recursive algorithm instead of using the pseudoinverse. The recursive algorithm can efficiently be implemented on graphical processing units (GPUs) (Lazar & Zhou, 2012).*

The overall decoding procedure is summarized in the diagram shown in Fig. 3.

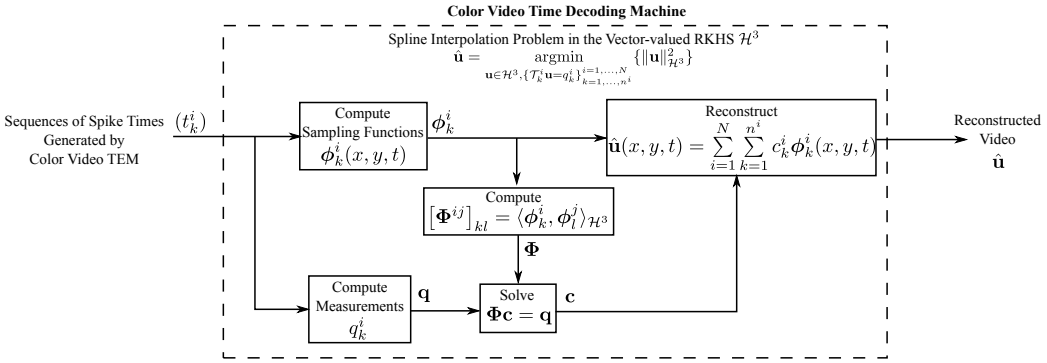


Figure 3: Color Video TDM diagram.

Remark 2. *By employing regularization our results can be readily extended to the case of noisy neurons (e.g., IAF neurons with random thresholds or Hodgkin-Huxley neurons with stochastic conductances) (Lazar et al., 2010).*

2.3.2. Evaluating Massively Parallel TDM Algorithms for Color Vision

In this section we describe an example of encoding and decoding of a color video sequence. The stimulus is a 10 [s], 160[px] $\times$ 90[px] natural color video. The video is encoded by a neural circuit consisting of $30,000$ color-component receptive fields in cascade with $30,000$ IAF neurons. The detailed construction of the encoding circuit is given in Appendix C.

In reconstruction, a spatio-temporal stitching algorithm similar to the one in Lazar & Zhou (2012) is deployed. The entire screen is divided into 4 overlapping regions of size 56[px] $\times$ 90[px] and time is cut into 150 [ms] slices. The stitching volume then becomes 56[px] $\times$ 90[px] $\times$ 0.15[s]. We picked the orders of the space to be $L_x = 24$, $L_y = 36$, $L_t = 8$, $\Omega_x = \Omega_y = 0.375 \cdot 2\pi$ and $\Omega_t = 10 \cdot 2\pi$ so that the overall period of the space is larger than each of the volumes. We did this in order to embed a typically non-periodic natural stimulus into a periodic space.

Remark 3. *Note that natural stimuli exhibit an "$1/f$ spectrum". Their Fourier coefficients are typically significant only in a certain spectral neighborhood around the origin. For simulation purposes, we restricted the set of spatial basis functions (l_x, l_y) to lie inside the elliptical set $\{(l_x, l_y) \,|\, l_x^2/L_x^2 + l_y^2/L_y^2 \leq 1\}$. Since the coefficients of all other basis functions in $\mathcal{H}$ are set to zero this restriction let us work instead in a subspace of $\mathcal{H}$ with an elliptical bandwidth profile. The spatial dimension $\dim_{xy}(\mathcal{H})$ is thereby approximated by the cardinality of the above subspace. Throughout this paper, if not stated otherwise, the cardinality of this subspace will provide the order of $\mathcal{H}$.*

The total number of spikes produced by encoding a 10-second long video was $9,001,700$. Each volume is typically reconstructed from about $90,000$ measurements. It took about 2 days to reconstruct the video on 9 GPUs. The reconstructed video is provided in Video S1 (see supplementary material). We show a snapshot of the original color video and the reconstructed video in Figure 4. The snapshot of the original video is shown on the left, which shows a bee on a sunflower. The reconstruction is shown in the middle and the error on the right. The error can be seen to be fairly small (zero error is shown in gray). The signal-to-noise ratio (SNR) of the reconstruction is 30.71 [dB]. The structural similarity (SSIM) index (Wang et al., 2004) of the reconstruction is 0.988. In addition, each color component can be individually accessed. Snapshots of all three channels corresponding to the time instant depicted in Fig. 4 are shown in Fig. 5. The corresponding video

is shown in Video S2 (see supplementary material). Since the original video is a natural video and is not strictly in the RKHS, the video is not perfectly reconstructed. However, it is still decoded with very high quality.

Figure 4: A snapshot of the original and the reconstructed color video. From left to right are respectively, the original video, reconstructed color video and the error.

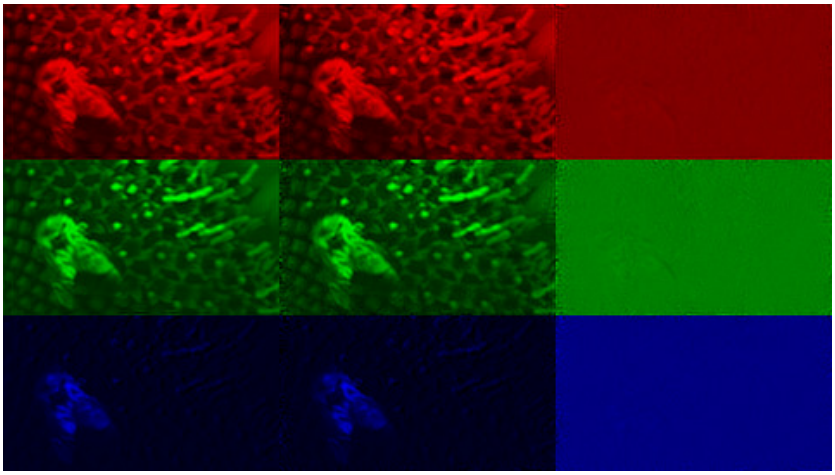


Figure 5: A snapshot of the original and the reconstructed videos of each color channel. From left to right are respectively, the original video, reconstructed color video and the error. Each color channel is a gray-scale image that is pseudo-colored to indicate the respective channels.

A limitation in modeling stimuli using the representation in (1) is that stimuli must be periodic. However, this is rarely true for natural visual stimuli. The imposed periodicity can be mitigated by using a larger period in reconstruction. We note that the periods T_x, T_y, T_t defined in the reconstruction space are larger than the size and duration of the stitching volume. By embedding the stitching volume in a larger periodic space, the reconstruction no longer has to be periodic within the stitching volume. This makes reconstruction of natural stimuli possible and the choice of space flexible. It may seem at first that the dimension of the space that the reconstruction is embedded into is

increased. However, this does not necessarily mean that the number of spikes required to reconstruct the stimulus has to increase. This is due to the fact that the sampling functions only need to span a subspace that the stitching volume is associated with. For example, in the reconstruction above, the space of choice is of dimension $137,751$ ($dim_{xy}(\mathcal{H}) = 2,701$, $dim_t(\mathcal{H}) = 17$). Already some $90,000$ measurements yield a high quality reconstruction of the visual stimuli. We note that the temporal period is $T_t = 0.8s$, while the duration of the stitching volume is only 0.15s. As long as the spikes are dense enough within the 0.15s time interval of interest, the stimulus can be reconstructed even if the number of spikes is lower than $137,751$. However, the sampling functions still have to span a minimal subspace that encloses the 0.15 time duration.

2.4. Identification of Neural Encoding Circuits for Color Vision

2.4.1. Channel Identification Machines for Color Vision

The color video encoded by the Color Video TEM can be reconstructed, given the spike times produced by a population of neurons and the parameters of each of the neurons. However, in many circumstances, the parameters of the neurons are not available apriori and need to be identified. In this scenario, the neurons are typically presented with one or more input test stimuli and their response, or output, is recorded so that neuron parameters can be identified using the input/output data. It can be shown that identification problems of this kind are mathematically dual to the decoding problem discussed in the previous section. Specifically, it can be shown that information about both the receptive fields and the spike generation mechanism can be faithfully encoded in the spike train of a neuron. Spike times are viewed as signatures of the entire system, and under appropriate conditions, these signatures can be used to identify both the receptive fields and the parameters of point neurons. The key experimental and theoretical insight is that the totality of spikes produced by a single neuron in N experimental trials can be treated as a single multidimensional spike train of a population of N neurons encoding fixed attributes of the neural circuit. Furthermore, it can be proven that only a projection of the neural circuit parameters onto the input stimulus space can be identified. The projection is determined by the particular choice of stimuli used during experiments and under natural

conditions it converges to the underlying parameters of the circuit (Lazar & Slutskiy, 2010, 2012, 2014b).

In this section we demonstrate that the ideas developed in Lazar & Slutskiy (2010), Lazar & Slutskiy (2012) and Lazar & Slutskiy (2014b) can be extended to massively parallel neural circuits that process (color) visual stimuli. For clarity, we consider identification of receptive fields only. Identification of spike generation parameters and/or connectivity between neurons can be handled similarly to what has been previously described (Lazar & Slutskiy, 2014b).

For presentation purposes, we consider the identification of a single receptive field associated with only one neuron, since identification of multiple receptive fields for a population of neurons can be performed in a serial fashion. We therefore drop the superscript i in h_m^i throughout this section and denote the m-th kernel component by h_m. Moreover, we introduce the natural notion of performing multiple experimental trials and use the same superscript i to index stimuli $\mathbf{u}^i$ on different trials $i = 1, ..., N$. In what follows, the neural circuit referred to as the Color Video TEM consists of a color receptive field $\mathbf{h} = (h_1, h_2, h_3)^T$ in cascade with a single IAF neuron.

Definition 1. *A signal $\mathbf{u}^i$, at the input to a Color Video TEM together with the resulting output $\mathbb{T}^i = (t_k^i)_{k\in\mathbb{Z}}$ is called an input/output (I/O) pair and is denoted by $(\mathbf{u}^i, \mathbb{T}^i)$.*

Definition 2. *The operator $\mathcal{P} : H^3 \to \mathcal{H}^3$ with elements $(\mathcal{P}\mathbf{h})_m$, $m = 1, 2, 3$, given by*

$$(\mathcal{P}\mathbf{h})_m(x, y, t) = \int_{\mathbb{D}} h_m(x', y', t')K_m(x, y, t; x', y', t')dx'dy'dt',$$

is called the projection operator.

Consider a single neuron receiving a stimulus $\mathbf{u}^i \in \mathcal{H}^3, i = 1, 2, ..., N$. The aggregate output $v^i = v^i(t)$, $t \in \mathbb{D}_t$, of the receptive field $\mathbf{h}$ produced in response to the stimulus $\mathbf{u}^i$ during the trial i is given by

$$v^i(t) = \sum_{m=1}^{3} \int_{\mathbb{D}} h_m(x, y, t - s)u_m^i(x, y, s)dxdyds, \tag{27}$$

where each signal u_m^i is an element of the space $\mathcal{H}$.

Let $\mathcal{L}_k^i : \mathcal{H}^3 \to \mathbb{R}$ be the linear functionals given by

$$\mathcal{L}_k^i\big[\mathcal{P}\mathbf{h}\big] = \int_{t_k^i}^{t_{k+1}^i} \sum_{m=1}^{3} \left(\int_{\mathbb{D}} u_m^i(x,y,t-s)(\mathcal{P}\mathbf{h})_m(x,y,s)dxdyds \right) dt, \tag{28}$$

for all $i = 1, ..., N$, and $k \in \mathbb{Z}$. Then, the Color Video TEM is described by the set of equations (see also Appendix D):

$$\mathcal{L}_k^i\big[\mathcal{P}\mathbf{h}\big] = q_k^i, \quad k \in \mathbb{Z}, \quad i = 1, ..., N. \tag{29}$$

We note that because $\mathcal{L}_k^i$ is linear and bounded, (29) can be expressed in inner product form as

$$\langle \mathcal{P}\mathbf{h}, \boldsymbol{\psi}_k^i \rangle = q_k^i, \tag{30}$$

where $\boldsymbol{\psi}_k^i(x,y,t) = \big[\psi_{1,k}^i(x,y,t), \psi_{2,k}^i(x,y,t), \psi_{3,k}^i(x,y,t)\big]^T$ and

$$\psi_{m,k}^i(x,y,t) = \mathcal{L}_k^i\left[\,\overline{\mathbf{K}\mathbf{e}_m}\,\right], \quad m = 1, 2, 3. \tag{31}$$

In contrast to equation (20) each inter-spike interval $[t_k^i, t_{k+1}^i)$ produced by the IAF neuron on experimental trial i is now treated as a quantal measurement q_k^i of the sum of the components of the receptive field $\mathbf{h}$, and not the stimulus $\mathbf{u}^i$. When considered together, equations (29) and (19) demonstrate that the identification problem can be converted into a neural encoding problem similar to the one discussed in the previous section. Note, however, that in (19) i denotes the neuron number whereas i in (29) denotes the trial number. This concept is further illustrated in Figure 6.

We note that the spike trains produced by a Color Video TEM in response to test stimuli $\mathbf{u}^i$, $i = 1, ..., N$, carry only partial information about the underlying receptive field $\mathbf{h}$. Intuitively, the information content is determined by how well the test stimuli explore the system. More formally, given test stimuli $\mathbf{u}^i \in \mathcal{H}^3$, $i = 1, ..., N$, the original receptive field $\mathbf{h}$ is projected onto the space $\mathcal{H}^3$ and only that projection $\mathcal{P}\mathbf{h}$ is encoded in the neural circuit output. It follows from (30) that we should be able to identify the projection $\mathcal{P}\mathbf{h}$ from measurements q_k^i, $i = 1, ..., N$, $k \in \mathbb{Z}$.

We now provide an algorithm, called the Color Video Channel Identification Machine (Color Video CIM), to functionally identify a neural circuit processing color video stimuli. As discussed in the previous section, this algorithm can be considered to be the dual of the decoding algorithm of Theorem 1.

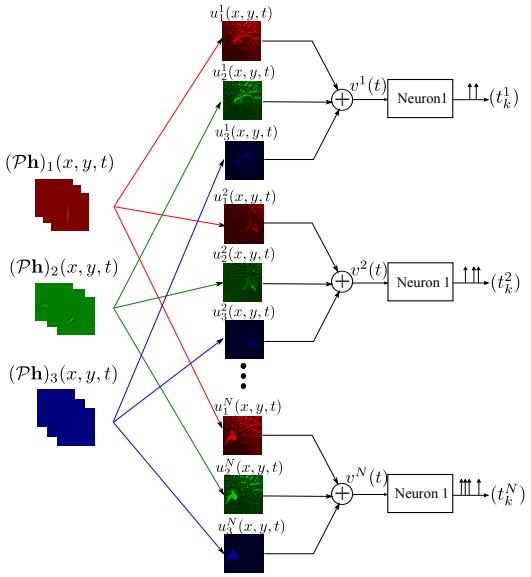


Figure 6: Block diagram of functional identification with multiple trials of controlled visual stimuli. Note that the same neuron is stimulated with N visual stimuli.

Theorem 2. *Let $\{\mathbf{u}^i | \mathbf{u}^i \in \mathcal{H}^3\}_{i=1}^N$ be a collection of N linearly independent stimuli at the input to a Color Video TEM with a receptive field $\mathbf{h}$. The projection $\mathcal{P}\mathbf{h}$ of the receptive field $\mathbf{h}$, can be identified from a collection of I/O pairs $\{\mathbf{u}^i, \mathbb{T}^i\}_{i=1}^N$ as a solution to the spline interpolation problem*

$$\widehat{\mathcal{P}\mathbf{h}} = \underset{\mathcal{P}\mathbf{h}\in\mathcal{H}^3, \{\mathcal{L}_k^i \mathbf{h} = q_k^i\}_{k=1,\cdots,n^i}^{i=1,\cdots,N}}{\operatorname{argmin}} \{\|\mathcal{P}\mathbf{h}\|_{\mathcal{H}^3}^2\}. \tag{32}$$

The solution is

$$\widehat{\mathcal{P}\mathbf{h}} = \sum_{i=1}^{N} \sum_{k=1}^{n^i} c_k^i \boldsymbol{\psi}_k^i,$$

where the c_k^i's satisfy the system of linear equations

$$\boldsymbol{\Psi}\mathbf{c} = \mathbf{q}, \tag{33}$$

with $\mathbf{c} = \left[c_1^1, c_2^1, ..., c_{n^1}^1, \cdots, c_1^N, c_2^N, ..., c_{n^N}^N\right]^T$, $\mathbf{q} = \left[q_1^1, q_2^1, ..., q_{n^1}^1, \cdots, q_1^N, q_2^N, ..., q_{n^N}^N\right]^T$, and $\boldsymbol{\Psi}$ is the block matrix

$$\boldsymbol{\Psi} = \left[\boldsymbol{\Psi}^{ij}\right], \tag{34}$$

where $i, j = 1, 2, ..., N$, and the block entries are given by (see also Appendix A)

$$\left[\mathbf{\Psi}^{ij}\right]_{kl} = \langle \boldsymbol{\psi}_k^i, \boldsymbol{\psi}_l^j \rangle_{\mathcal{H}^3}, \text{ for all } i, j = 1, 2, ..., N \text{ and } k = 1, 2, ..., n^i, l = 1, 2, ..., n^j.$$

A necessary condition for perfect identification of $\mathcal{P}\mathbf{h} \in \mathcal{H}^3$ is that the set of sampling functions $\boldsymbol{\psi} = (\boldsymbol{\psi}_k^i), k = 1, 2, ..., n^i, i = 1, 2, ..., N$, span $\mathcal{H}^3$. This requires to have the number of trials $N \geq 3 \cdot dim_{xy}(\mathcal{H})$ and a total of at least $\dim(\mathcal{H}^3) + N$ spikes. Equivalently, if the neuron produces ν spikes on each trial $i = 1, ..., N$, of duration T_t, then the number of trials

$$N \geq \begin{cases} \left\lceil \dfrac{3 \cdot dim(\mathcal{H})}{\nu - 1} \right\rceil, & \nu < 2L_t + 2 \\ 3 \cdot dim_{xy}(\mathcal{H}), & \nu \geq 2L_t + 2. \end{cases}$$

Proof: The proof is along the lines of the one for Theorem 1.
Remark 4. *Note that only the projection $\mathcal{P}\mathbf{h}$ of $\mathbf{h}$ onto $\mathcal{H}^3$ can be identified. In addition, notice the similarity of the identification and the decoding algorithms. This is a direct result of the duality of the functional identification and decoding problems.*

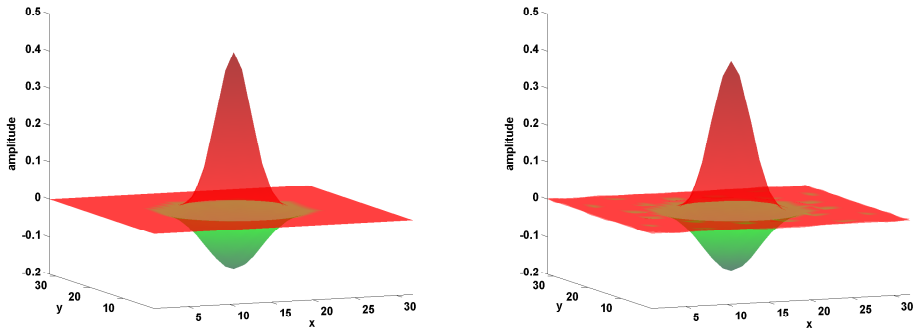
Theorem 2 provides, in addition to an algorithm, a lower bound on the number of video clips and the number of spikes required to perfectly identify the projection of receptive fields.

2.4.2. Evaluating Massively Parallel CIM Algorithms for Color Vision

In this section, we show an example of functional identification of a single non-separable spatio-temporal receptive field. We first demonstrate how one can use both natural video and artificially generated bandlimited noise to identify the receptive fields, and illustrate bounds for the number of video clips and number of spikes for perfect identification. Then, we perform a full-scale identification of the neural circuit described in Section 2.3.2 by using a long sequence of continuous natural stimuli instead of short video clip segments.

In the first example, the neuron to be identified has a receptive field that resembles that of a Red-On-Green-Off (R+G-) midget cell in the primate retina

(Benardete & Kaplan, 1997). The red and green components of the receptive field are modeled as space-time separable functions. They are Gaussian functions spatially, and resemble bandpass linear filters temporally. The blue component will also be identified, although it is set to zero. The temporal span of the filter is 150 [ms] and spatially it is confined to a 32 [px] $\times$ 32 [px] screen size. The receptive field is shown in Video S3 (see supplementary material). The snapshot of the receptive field at 40ms is depicted in Figure 7(a).



(a) (b)

Figure 7: An example of a spatio-temporal receptive field (150ms duration). (a) A snapshot of the receptive field to be identified, at 40ms. (b) A snapshot of the projection of the receptive field at 40ms. Parameters of the elements of the space $\mathcal{H}$ are: $T_t = 300$ms, $T_x = T_y = 32$, $L_x = L_y = 6$, $L_t = 12$, $\Omega_x = \Omega_y = 0.25\pi$ and $\Omega_t = 80\pi$. At this time instant, the red component of the receptive field provides the excitatory center (red surface) while the green component provides the inhibitory surround (green surface). The subtle difference between the original and the projected receptive field (23.12 [dB] SNR) indicates that the chosen input space is effective in exploring the receptive field. (see also Video S3 in supplementary material for full video)

To identify the receptive field, we consider $T_t = 300$ms, $T_x = T_y = 32$. We chose $L_x = L_y = 6$, $L_t = 12$, $\Omega_x = \Omega_y = 0.25\pi$ and $\Omega_t = 80\pi$. The total dimension of the space is $dim_{xy}(\mathcal{H}) \times (2L_t+1) \times 3 = 8,475$, where $\dim_{xy}(\mathcal{H}) = 113$, since we restricted the spatial basis functions to the set $\{(l_x, l_y) \,|\, l_x^2/L_x^2 + l_y^2/L_y^2 \leq 1\}$ (see also Remark 3). The projection of the receptive field $\mathcal{P}\mathbf{h}$, shown in Video S3 (see supplementary material), is close to the underlying receptive field $\mathbf{h}$ itself (see also Figure 7(b)). The Signal-to-Noise ratio (SNR) of the projection of the original filter onto the stimulus space is 23.12 [dB]

(the SNR is computed only for the red and green components). In what follows, we will only depict the SNR of the identified receptive fields and compare with the projection of the receptive field.

To identify the receptive field, we generate N video clips in the Hilbert space of interest by randomly picking coefficients for each basis. Note that by randomly picking coefficients we can easily end up with complex valued signals. We further ensure that the visual stimuli are real valued.

To illustrate the identification quality for different number of spikes and number of video clips, we modified the parameters of the IAF neuron while keeping the underlying receptive field the same. Note that the modification of parameters of the IAF neuron may not be biologically plausible; it is used here to better illustrate the bounds on the number of measurements.

First, we vary the number of video clips N while using the same number of spikes generated by each video clip. The SNR is shown is Figure 8(a). All three curves follow a general trend: the SNR increases as more video clips are used until it saturates at around 60 [dB], which indicates perfect identification (up to machine precision). Comparing the three curves, we see that if the neuron produces more than 25 measurements, or 26 spikes, per each video clip (blue and black curves), the SNR saturates at roughly $N = 3 \times \dim_{xy}(\mathcal{H}) = 3 \times 113 = 339$ video clips, as stated in Theorem 2. This corresponds to the lower bound on the number of stimuli needed in order to identify the receptive field. If the neuron generates fewer than 25 measurements, more stimuli are needed. For example, if 19 measurements, or $\nu = 20$ spikes, are produced in response to each stimulus, then a minimum of $\lceil 3 \times \dim_{xy}(\mathcal{H}) \times \dim_t(\mathcal{H})/(\nu - 1) \rceil = \lceil 3 \times 113 \times 25/19 \rceil = 447$ video clips is needed.

In Figure 8(b), we fix the number of video clips while varying the number of measurements of each video clip. Note that as the number of measurements per each video clip increases, so does the identification quality. However, the SNR cannot be further improved after the number of spikes for each video clip reaches $(2L_t + 1) + 1 = 26$ spikes, or 25 measurements. This is due to the fact that the freedom of the space in the temporal dimension is $(2L_t + 1)$. As shown in Theorem 2, once each neuron fires $(2L_t + 1) + 1 = 26$ spikes, it does not produce additionally informative measurements. Therefore, the identification quality cannot be improved further. Furthermore, the cyan and black curves demonstrate that even if the total number of measurements

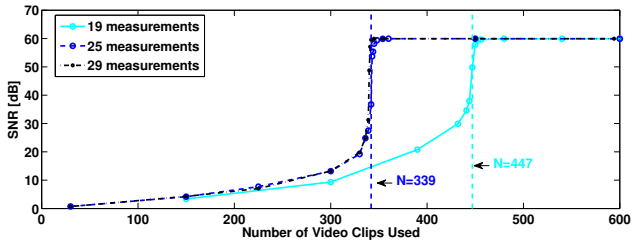
is larger than the dimension of the space, perfect identification may not be achieved if the lower bound $N = 3 \cdot \dim_{xy}(\mathcal{H}) = 3 \times 113 = 339$ is not met. For example, if the neuron generates 40 measurements per video clip, then both 180 and 336 video clips will result in the total number of spikes that is greater than the required number of $3 \times \dim(\mathcal{H}) = 8,475$ spikes.

In Figure 8(c), we vary the number of video clips while the total number of measurement remains fixed at around $9,000$. Identification quality is high when $N \geq 339$, illustrating the lower bound of videos/experiments needed in order to identify the receptive field with arbitrary precision. However, the identification quality saturates at $N = 342$, slightly shifted away from $N = 339$. This shift is mainly due to numerical errors and the random choice of video clips that does not guarantee linear independence among the sampling functions.

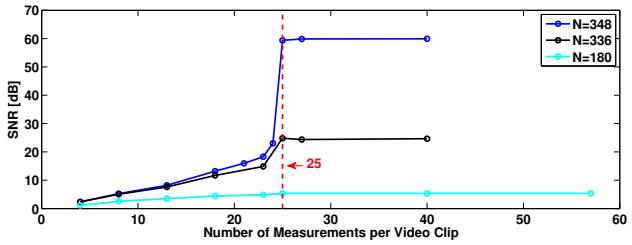
To sum up, for the first example, we have shown two useful bounds for perfectly identifying the projection $\mathcal{P}\mathbf{h}$ of a receptive field $\mathbf{h}$ onto a Hilbert space $\mathcal{H}^3$. The first lower bound is that the total number of measurements must be greater or equal to the dimension of the space $dim(\mathcal{H}^3)$. Equivalently, the totality of spikes produced in response to N experimental trials involving N different video clips must be greater than $dim(\mathcal{H}^3) + N$. The second lower bound is that the number of video clips N must be greater or equal to $3 \cdot dim_{xy}(\mathcal{H})$. Both conditions must be satisfied at the same time. In addition, we see that each video clip can provide a maximum of $2L_t + 1$ informative measurements towards the identification.

We now consider using a long sequence of continuous natural video in identifying the entire neural circuit. Colors in natural visual scenes are much more correlated than in randomly generated bandlimited signals using the above procedure. As neural systems are tuned to the natural statistics, it is likely that neurons will respond differently to natural stimuli. Thus, there is a need to be able to accommodate the use of natural stimuli during functional identification in real experiments. The machinery of RKHS, and spaces of trigonometric polynomials specifically, provide that capability. It is essential, however, to properly segment a long natural sequence into multiple segments with appropriate spike times associated with each. This is detailed in Appendix E.

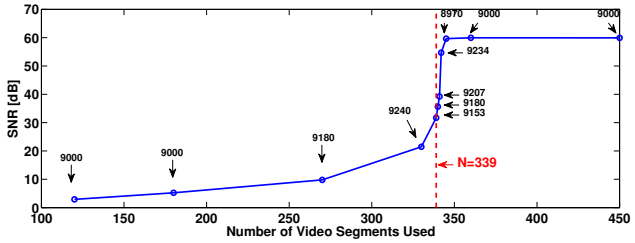
In examples below we use a custom natural video shot by a handheld device. The total length of the video is 200 seconds. We use this single video to



(a)



(b)



(c)

Figure 8: Effect of the number of video clips and the total number of measurements on the quality of identification. (a) A fixed number of measurements (19, 25 and 29) is used from each video clip. SNR increases as more video clips are used until the total number of measurements passes the dimension of the space (19, 25) and the number of video clips reaches $3 \cdot dim_{xy}(\mathcal{H})$ (25,29) (b) A fixed number of video clips are used in identification while the number of measurements from each video clip increases. The identification performance saturates after each video clip generates 25 measurements. This number corresponds to $dim_t(\mathcal{H})$. This suggests that a minimum of 26 spikes is needed to fully explore the space in the time dimension. (c) Varying the number of video clips in identification while keeping the number of total measurements above $dim(\mathcal{H}^3)$ at all times, the identification performance steeply increases after the number of video clips reaches $3 \cdot dim_{xy}(\mathcal{H})$. This suggests that in order to identify the color receptive field, not only the total number of spikes has to be larger than $dim(\mathcal{H}^3)$, but also the number of video clips used needs to be larger than $3 \cdot dim_{xy}(\mathcal{H})$.

identify the complete neural circuit, that is, the receptive fields of all $N = 30,000$ neurons. Due to computational constraints and in the interest of time, we identify each of the receptive field components $h_m^i(x, y, t)$ separately rather than the entire $\mathbf{h}^i$. This can easily be done by supplying a single color channel during the identification procedure.

For simplicity, we assume that the dilation parameter of the receptive field was known. For $\alpha = 2^{0.5}$, the chosen screen size is $24\,[\text{px}] \times 24\,[\text{px}]$, $\Omega_x = \Omega_y = 0.5$. For $\alpha = 2^{1.5}$, the chosen screen size is $48\,[\text{px}] \times 48\,[\text{px}]$ and $\Omega_x = \Omega_y = 0.25$. In both cases, $L_x = L_y = 12$, $L_t = 4$ and $\Omega_t = 2\pi \cdot 20$. The dimension of both spaces is $\dim_{xy}(\mathcal{H}) \times (2L_t + 1) = 3,969$, where $\dim_{xy}(\mathcal{H}) = 441$. In this example, each neuron in the population has fixed but different parameters and generates about 100 spikes per second, or about $10 = 2L_t + 2$ spikes per windowed video clip. Note here that since we only use spikes generated in the second half of the $T_t = 0.2$ window (see Figure 9(a)), the number of spikes used per windowed video clip is approximately $100 \times T_t/2 = 10$. This choice of stimulus and neuron parameters allows each neuron to provide the maximum number of informative spikes about each video clip in the simulation. We vary the number of spikes used in the identification. The number of video clips co-vary with the number of spikes as a result. Each receptive field is identified on a single GPU and it takes about 10 minutes when using 17,000 spikes to identify the receptive field. When using only 1,000 spikes, it takes 30 seconds to finish the identification process. Since a large number of neurons had to be identified, we performed simulation on a cluster that has 96 GPUs.

The SNR of the identified receptive fields over the original receptive fields are shown in Figure 9, where different colors are used to indicate a different number of total measurements used in identification. Each dot in the figure corresponds to the SNR of an identified receptive field for the corresponding neuron. Figure 9 shows a general trend that a larger number of measurements produces better identification results.

3. Jointly Evaluating Encoding, Decoding and Identification

Functional identification of a visual neural circuit provides a quantitative description of the relationship between the input video and the spiking activity

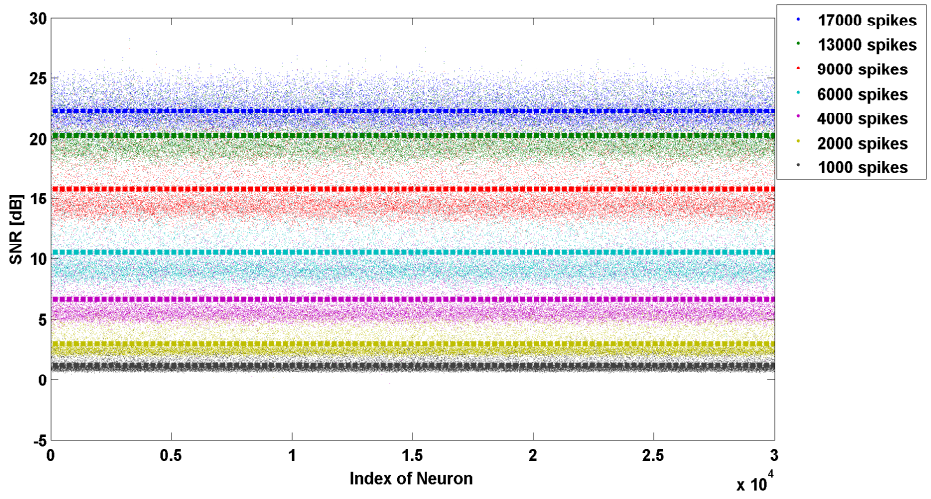


Figure 9: SNR of the 30,000 identified filters. Color indicates the number of total measurements used in the identification for each receptive field. The average SNR is shown by dashed line for corresponding colors.

of the neurons in the circuit. It is natural, then, to ask if this description can be used to reconstruct the actual stimulus that produces the spikes.

By connecting the CIM and TEM methodologies and exploiting their duality we provide an answer to this question in section 3.1. After functionally identifying the entire neural circuit, we apply a novel visual stimulus to the identified encoding circuit. From the resulting spike trains the visual stimulus is recovered using the identified circuit parameters (of the receptive fields). As shown in section 3.2 this provides a single measure for the quantitative evaluation of the identification process.

3.1. Stimulus Reconstruction Using the Identified Circuit

Decoding of a visual stimulus from spikes produced by a population of identified visual neurons was previously investigated under various contexts. Warland et al. (1997) attempted to decode the time course of stimuli with spatially uniform intensity by identifying an optimal linear decoder for retinal ganglion cells in salamanders. A similar decoding algorithm was used to decode natural visual scenes from neural activity in the cat LGN (Stanley et al., 1999). Both of these approaches are based on linear decoding using

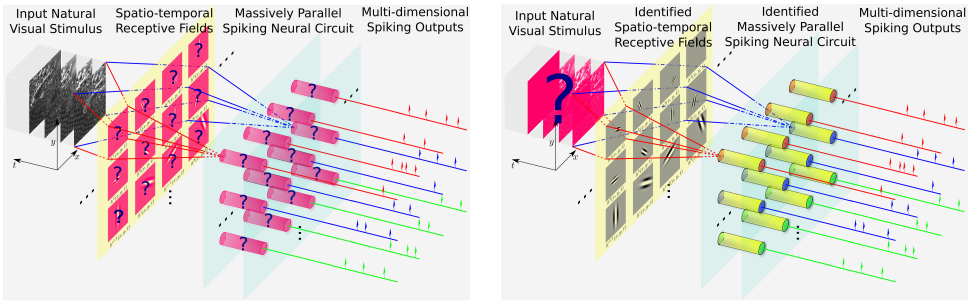
the firing rate of the neurons. In addition, only a small number of neurons was considered.

Our approach differs in three important ways. First, instead of decoding from firing rate of neurons with a linear decoder, we use a non-linear decoder applied to the exact spike times generated by the neurons. This is attractive from an experimental standpoint, since the stimulus can be recovered using a single experimental trial, eliminating the need to repeat the same experiment in order to compute the firing rate. The latter may not be possible due to the variability in experimental conditions and the state of the neural circuit. Second, the neural circuit we are investigating is a massively parallel neural circuit comprised of thousands of neurons. In recent years it has become apparent that information in the brain is in general represented by vast populations of neurons that process sensory stimuli in a parallel fashion. The required population size is both modality- and stimulus-specific and our findings provide an estimate for the lower bound on the number of neurons that are need to faithfully represent a color video stimulus in the spike domain. Finally, we have provided conditions under which the neural circuit can be identified and the input videos can be subsequently faithfully reconstructed. Note that, in identification, we can only identify the projection of the neural circuit parameters onto the RKHS. Therefore, using the identified neural circuit, it is only possible to reconstruct a stimulus up to its projection onto the same RKHS.

Furthermore, by bridging the identification and reconstruction problems, it is possible to evaluate the amount of information encoded by neurons. One approach is summarized in Fig. 10. Given a massively parallel neural circuit, we first identify its parameters (Fig. 10(a)). This can be done by presenting an input video and using the resulting spike train in identification. Second, for a novel stimulus presented to the same neural circuit, the spike train generated by the neural circuit can be used along with the identified parameters to reconstruct the novel stimulus (Fig. 10(b)).

3.2. Evaluation of Functional Identification in the Stimulus Space

There are three ways to evaluate the quality of the identification process, as schematically illustrated in Fig. 11. In simulation, the original receptive fields are known, and we can compare the identified circuit to the original (indicated by (1) in Figure 11). Such evaluation is performed in the *parameter*



(a) (b)

Figure 10: Massively parallel neural circuit. (a) Neural circuit identification. Known stimuli are presented to a neural circuit whose parameters are unknown, the circuit needs to be identified from the stimuli and the spike times generated. (b) Diagram of the encoding circuit. An unknown stimulus is present to the circuit. The circuit has been identified and spikes are known, the visual stimulus can be reconstructed from the identified parameters and the spike times. (Modified with permission from: Lazar & Zhou (2014), ©2014 IEEE)

space. The drawback of this approach is that in biological neural systems the ground truth is not known. In other words, there is no true parameter space to compare with.

An alternative and widely adopted approach is to use the identified circuit to predict the response of the circuit to novel stimuli (indicated by (2) in Figure 11). A novel stimulus is presented to both the neural circuit and the identified circuit and their responses are compared. In other words, this type of verification is performed in the *spike train space*. In this approach the identified circuit can be viewed as an I/O equivalent or phenomenological circuit. Often, however, such I/O equivalence cannot be made on a spike by spike basis. Typically, only coarser measurements such as the Peri-Stimulus Time Histogram (PSTH) are available for making predictions (Carandini et al., 2005). Although these coarser metrics are often good useful indicators of the identification performance, they exhibit several shortcomings, especially in population encoding and natural stimuli settings. When evaluating the identification performance of a massively parallel neural circuit, checking the PSTH for every neuron alone typically amounts to thousands of measurements. Moreover, quantitative distance measures between PSTHs are often times not well defined and are hard to interpret. Most importantly, individual PSTH predictions may not speak for the overall functionality of the

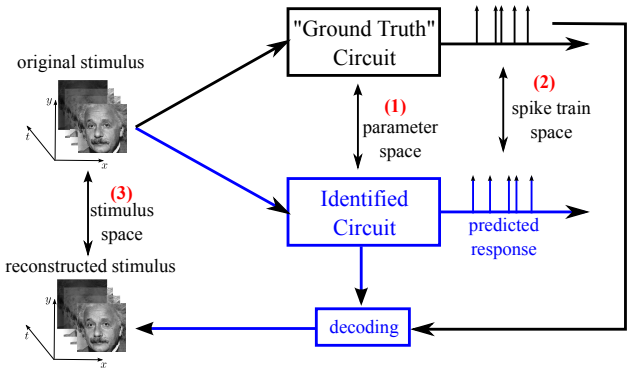


Figure 11: Three ways of evaluating the identification quality.

entire identified circuit.

Within the framework of time encoding and channel identification, we introduce a third way of directly assessing the quality of the identification using the information content retained in the spike train. This metric is the distance between the original visual stimulus and the stimulus decoded using the identified circuit parameters (indicated by (3) in Fig. 11). When using the identified filters in decoding instead of the underlying filters, the t-transform corresponding to (19) is no longer precise since the identified filters deviate from the projections of the underlying filters. However, it is expected that the error in the t-transform should decrease when the identification quality increases, thereby leading to improved reconstruction. Similarly, a poor quality of the decoded cross-validating stimulus suggests that the identified circuit parameters deviate from their true values. By using large scale simulations, the converse can be observed, *i.e.*, a high quality reconstruction indicates a high quality functional identification.

By evaluating the identified circuit in the *stimulus space*, we obtain a metric that quantifies the identification result. In other words, the evaluation of the entire identified neural circuit is reduced to intuitive comparisons in the stimulus space, thereby augmenting the usual neuron-by-neuron comparison in the spike-train space. For images and video, such an evaluation can be visually performed. In addition, many well defined distance measures are readily available in the signal processing literature (see Wang et al. (2004) and references therein).

This methodology of using identified parameters in decoding may allow us

to identify a relevant subset of neurons in a large scale neural circuit and, at the same time, evaluate the goodness of the identification algorithm for that neuronal subset. In the ideal case, the entire circuit can be identified. We use here faithful representation as an example. By bridging the identification problem (channel identification) with the encoding problem (time encoding), we are able to evaluate the function of the entire circuit by decoding the stimulus and quantifying the information content retained in the spikes.

3.3. Joint Performance Evaluation

To illustrate the results of the methodology presented in the previous section, we performed computer simulations of encoding, decoding and identification using the massively parallel neural circuit described in section 2.3.2 and identified in 2.4.2. We also evaluated the performance of the identification algorithm in the stimulus space. Here we present the results obtained and discuss practical issues arising in using the TEMs, TDMs and CIMs with natural videos.

In section 2.4.2, we identified a massively parallel neural circuit with $30,000$ neurons. The identified receptive fields were compared to the ground truth in the parameter space (see also Figure 11), and it was shown that the quality of the identified receptive fields improves with the length of the test stimulus. Using the identified receptive fields, we evaluate the quality of functional identification in the stimulus space.

The neural circuit is first identified with 7 different settings. In each setting, a stimulus of a given duration was used for identifying the receptive fields (see also Figure 9). As the length of the stimuli increases, more spikes are obtained for identification Using each set of identified parameters, we decode the spikes generated by the Color Video TEM when encoding novel stimuli. Since a perfect reconstruction of the video can be demonstrated for the case when the underlying receptive fields are known (see Section 2.3.2), we know that it is possible to reconstruct the stimulus with high quality if the circuit is identified well. The decoding procedure, including the parameter of the space and the stitching window, was taken to be exactly the same as that in Section 2.3.2. The decoding using identified neural circuit parameters takes the same time as the decoding in Section 2.3.2, that is, about 2 days for 10 seconds of video.

The reconstruction quality is shown in Figure 12. The entire video sequence can be found in Video S4a (see supplementary material). The SNR of the decoded video increases as more measurements are used for identifying each receptive field in the neural circuit. This indicates a better overall quality in the identification step. Such an evaluation in the stimulus space is also consistent with the parameter space evaluation in Section 2.4.2 that is based on the ground truth neural circuit. Moreover, the reconstructed videos are visually self-explanatory. The reconstruction artifacts are clearly visible in the videos to the left of the vertical line that marks the theoretical lower bound on the number of measurements required for identification. Those videos to the right of the vertical line have a superior quality and are visually close to what can be reconstructed when using the underlying receptive fields. In Video S4b, we show the reconstruction evaluated using the SSIM index as the metric (see supplementary material).

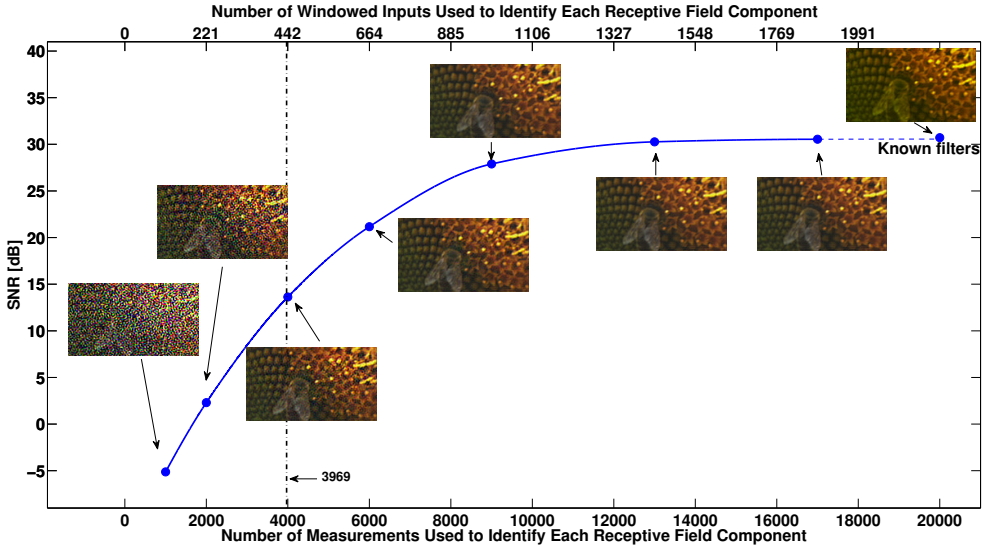


Figure 12: Evaluation of identification in the stimulus space. Neural circuit identified with different number of measurements is used to reconstruct a novel stimulus. The increase of SNR of the reconstructed video shows the improved identification quality as the number of measurements used in identifying each receptive field in the neural circuit increases. Snapshots of the corresponding reconstruction are shown with their SNR.

Comparing to the evaluation in Section 2.4.2, the quality assessment in the

stimulus space appears to be much more straightforward and more intuitive. It also provides a measure of the encoding capability of the identified circuit. Specifically, one can readily evaluate what information about the stimulus is preserved by the neural circuit by simply comparing the original video and the video decoded using the identified circuit parameters.

4. Massively Parallel Neural Circuits for Stereoscopic Color Vision

The encoding, decoding, identification and its evaluation in the stimulus space for color videos discussed in the previous sections provides a basis for encoding, decoding and identification of multi-dimensional videos. We discuss here these extensions with particular emphasis on stereoscopic color vision.

In this section we present examples for encoding/decoding of stereoscopic monochrome video and stereoscopic color video. In addition, examples are provided, demonstrating that mixing binocular information as well as color information can be treated within our theoretical framework. Identification results hold similarly. However, due to space constraints, parameter identification examples in the stereoscopic video setting are omitted.

The current formulation of encoding in a vector-valued RKHS also provides the flexibility to model videos that have a total of p components. Examples include color videos defined with a different color scheme, and multiview videos that correspond to the same visual scene sampled by more than one visual sensor. The extension to a $\mathbb{R}^p$-valued RKHS is straightforward, since the space of signals can be modeled as $\mathcal{H}^p$. In what follows we discuss two applications based on different values of p.

4.1. Massively Parallel Neural Circuits for Stereoscopic Video

Stereoscopic videos are two different streams of video that are projected onto the left and right eyes. Typically, the two video streams represent views of the same visual scene taken from slightly different angles. They arise naturally in the early visual system of vertebrates where binocular vision dominates. By combining multiple views of the visual scene, binocular vision makes it possible to extract the depth information about the visual scene.

A massively parallel neural circuit for encoding monochrome (grayscale) stereoscopic video is shown in Fig. 13. The input videos, denoted by (abuse

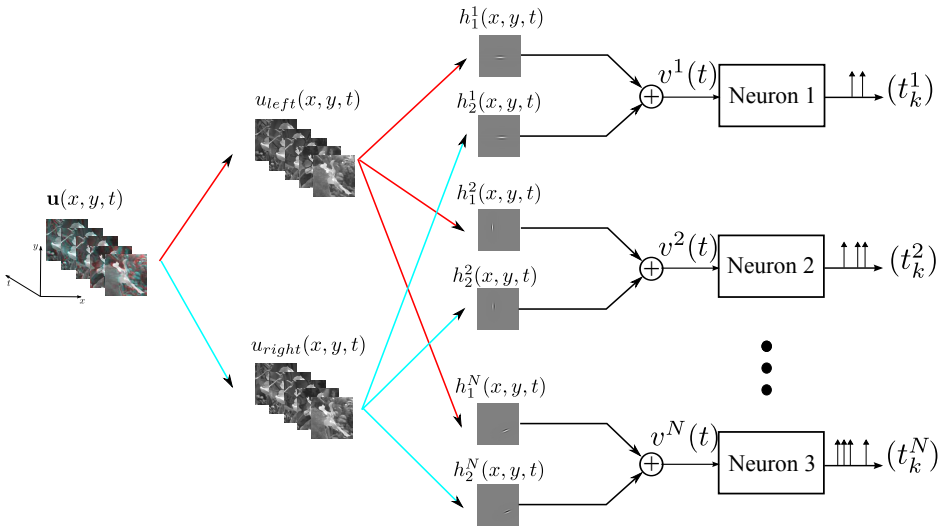


Figure 13: Diagram of massively parallel neural circuit for encoding stereoscopic video.

of notation),

$$\mathbf{u}(x,y,t) = [u_1(x,y,t), u_2(x,y,t)]^T,$$

may come from a single visual scene but are sensed by two eyes, where u_1 denotes the monochrome video sensed by the left eye and u_2 denotes that sensed by the right eye. In the visual cortex, the information from both eyes is combined in some of the neurons (Qian, 1997). This is modeled by the multi-component receptive fields $\mathbf{h}^i(x,y,t)$, where, by abuse of notation,

$$\mathbf{h}^i(x,y,t) = \left[h_1^i(x,y,t), h_2^i(x,y,t)\right]^T. \tag{35}$$

Again, each component $h_m^i(x,y,t), m = 1,2, i = 1,...,N$, is assumed to be causal with finite support, and BIBO stable. Each component receptive field performs a linear filtering operation on its corresponding input video before the outcomes are summed and fed into an IAF neuron (Freeman & Ohzawa, 1990; Zhu & Qian, 1996). The above neural encoding circuit forms a Stereoscopic Video TEM.

We provide an example here demonstrating the encoding of stereoscopic videos and their reconstruction. We omit the example of identification and

the performance evaluation, since they will be similar to the case of color video.

The stereoscopic video has a view of 192 [px] $\times$ 108 [px] in each component and was shot by two cameras calibrated to match binocular vision and provide a 3D visual perception (Inoue, 2009). Parameters of the space are $L_x = 72$, $L_y = 40$, $\Omega_x = 0.75\pi$, $\Omega_y = 0.74\pi$, $L_t = 8$ and $\Omega_t = 20\pi$. The decoded stereoscopic video is shown in Video S5 (see supplementary material). A snapshot of the reconstruction is shown in Fig. 14. SNR of the reconstruction is 35.77dB, SSIM index is 0.980. The reconstructions of separate eye channels are shown in Fig. 15. The corresponding video is shown in Video S6 (see supplementary material)

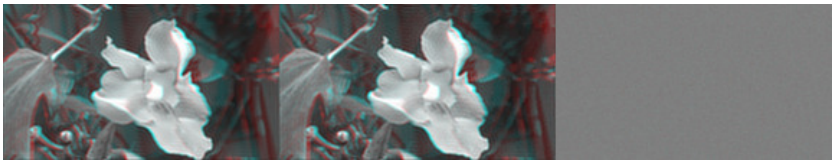


Figure 14: A snapshot of the original stereo video and the reconstructed stereo video. From left to right are respectively, the original video, reconstructed stereo video and the error. The 3D effects can be visualized by wearing red-cyan 3D glasses.

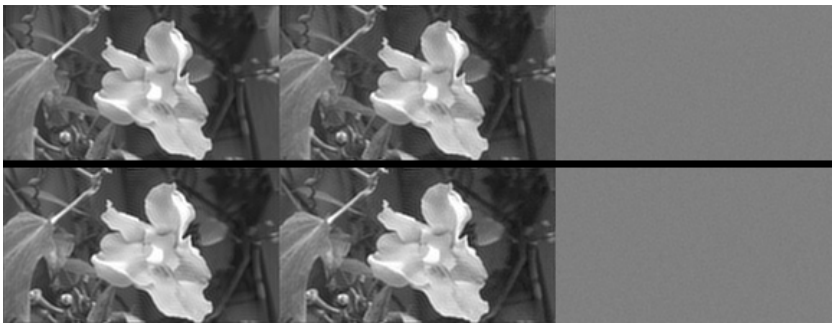


Figure 15: A snapshot of the original stereo video and the reconstructed stereo video in separate channels. The left eye channel is shown in the top row and the right eye channel in the bottom row. From left to right are respectively, the original video, reconstructed video and the error.

4.2. Massively Parallel Neural Circuits for Stereoscopic Color Video

The massively parallel neural circuits for color video and stereoscopic video can be combined to work with stereoscopic color video. The RKHS of interest

then becomes $\mathcal{H}^6$. Figure 16 depicts a block diagram of the massively parallel neural circut for encoding stereoscopic color video. Neurons in the circuit can now encode information in all the color channels of both eyes.

The encoding, decoding and functional identification based on this circuit can be formulated similarly as described in Section 2 and 4.1.

We show here a video in Video S7 (see supplementary material) for demonstration of the decoding of stereo color video. A snapshot of the original 3D color video and reconstruction is shown in Fig. 17. SNR of the reconstruction is 27.37dB, SSIM index is 0.960. The reconstruction of individual channels are shown in Fig. 18. The corresponding video is shown in Video S8 (see supplementary material).

5. Discussion

We presented TEMs and derived TDMs for color and stereoscopic visual stimuli. A common feature of encoding of all these stimuli is the use of multiple sensors to extract and to subsequently combine information from these sensors. Color visual scenes were decomposed into three color channels. Neurons then sampled, compared or composed information contained in the output of various color channels and multiplexed that information in the time domain using spikes. For stereoscopic vision, the visual scene was separately sensed by two horizontally displaced observers (eyes) and fed into a population of neurons. The receptive fields of each neuron can individually process and compose the information received from both eyes.

Natural scenes are highly complex with variations in intensity, wavelength and geometry. It is interesting to note that in order to perceive the complexity of the visual world, the visual system seems to mix different types of information. The TEMs we formulated here for stereoscopic color vision are instances of such mixing.

Mixed signals encoding may be important in a number of ways. First, each of the color channels represents an aspect of a visual scene. Information can be highly redundant across multiple channels. For example, all RGB channels carry information about the same objects in a visual scene. The shapes of these objects are shared in all channels. A change in color intensity is more likely to happen at the boundary between two objects and this change

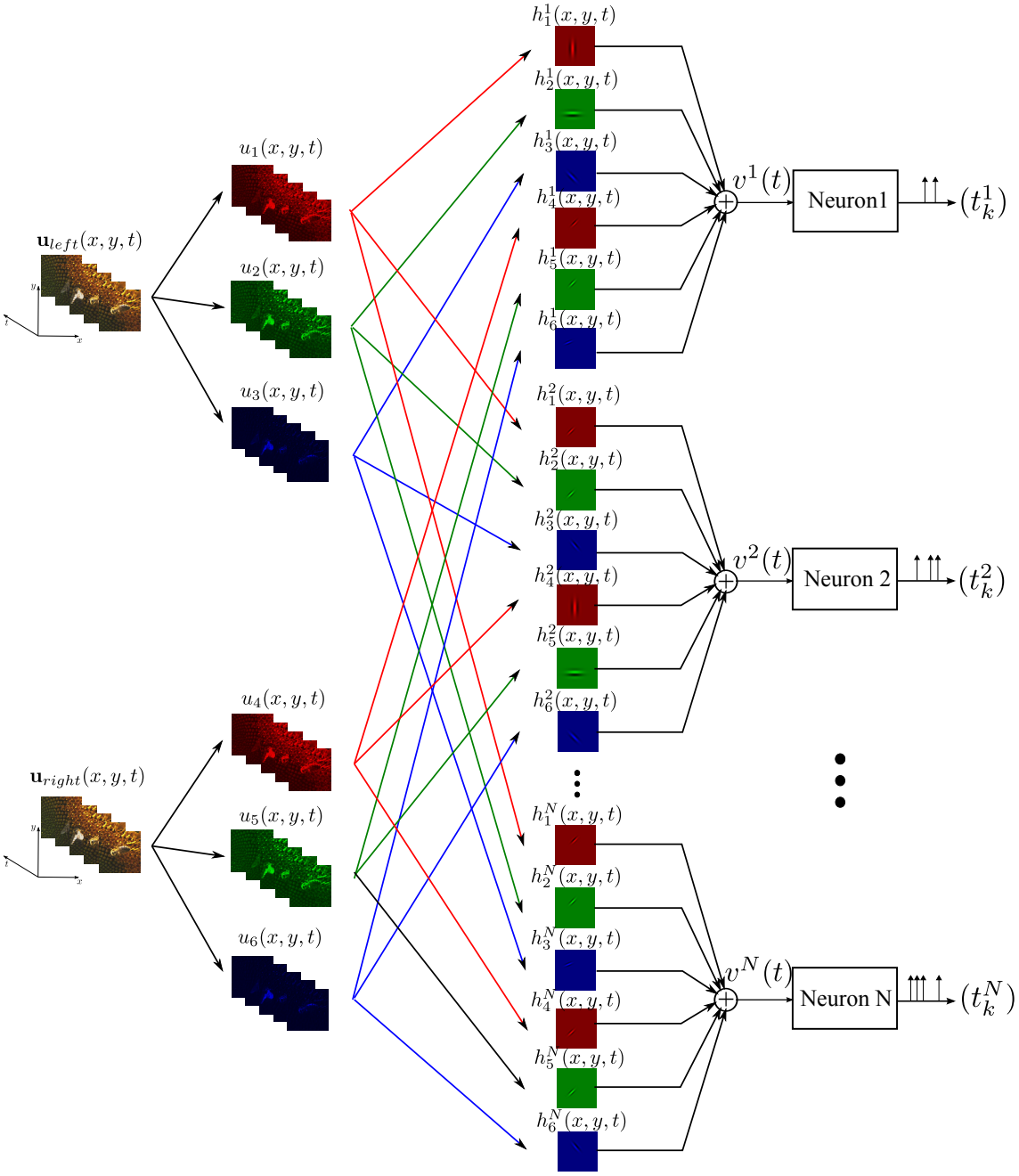


Figure 16: Diagram of the massively parallel neural circuit for encoding stereoscopic color video.

Figure 17: A snapshot of the original stereoscopic color video and the reconstructed stereoscopic color video. From left to right are respectively, the original video, reconstructed color video and the error. The 3D effects can be visualized by wearing red-cyan 3D glasses.

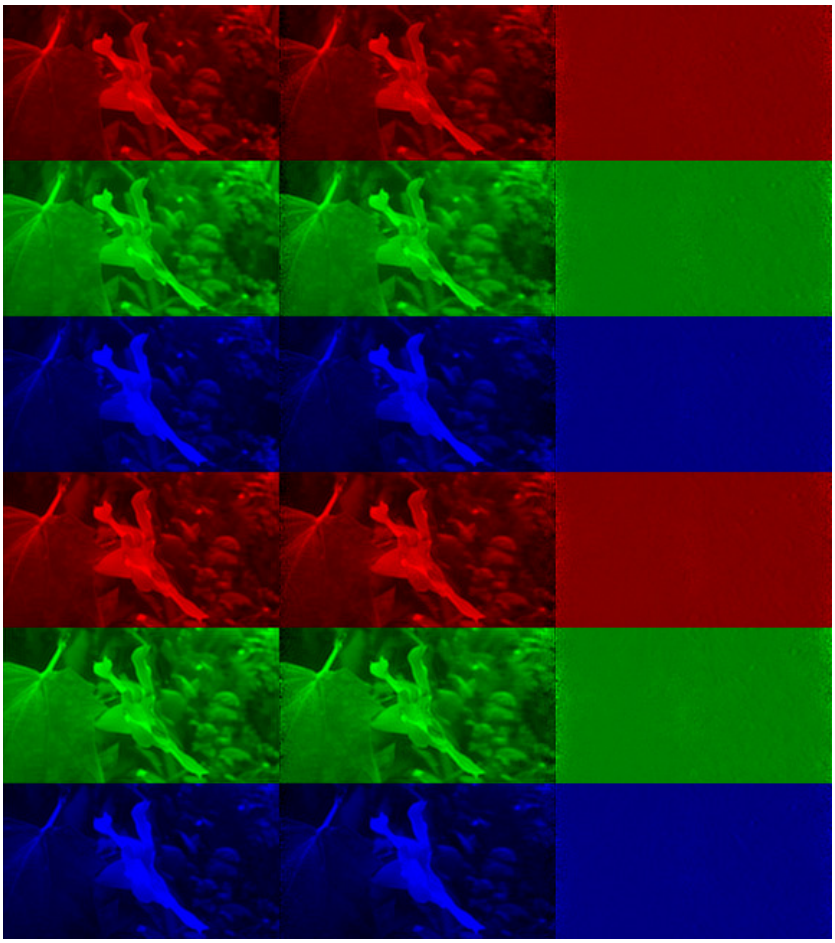


Figure 18: A snapshot of the original stereo color video and the reconstructed in separate channels. The first three rows are the color channels in the left eye and the last three rows are the color channels in the right eye. From left to right are respectively, the original video, reconstructed video and the error.

is propagated across color channels. Combining information from multiple channels may enhance the encoding efficiency and provide a simpler decoding algorithm. The YUV or YCbCr video formats, for example, have long been used in digital video technology where some of the channels can be subsampled while, at the same time, keeping a similar perceptual quality level. We did not explore here such a method of redundancy reduction. Rather, we provided a framework for representing multiple channels of information, for recovering the scene and for identifying channel parameters, such that these facilitate redundancy reduction.

Second, the mixing of cone signals can be utilized as coordinate transformations in the space/time color space. Such transformations may be useful in object recognition or in the separation of color and contrast information.

Third, mixing multiple channel output signals allows multiple information patterns to be represented together and therefore enables readout of different aspects of signals anywhere in the system. In other words, it provides broadcast information to multiple "receivers". This makes it possible for higher level neural systems to extract information from a common pool of spikes.

We have presented a comprehensive, yet highly intuitive method for evaluating the functional identification of massively parallel neural circuits. The problem was formulated in the space of color visual scenes, but can generally be applied to other stimulus spaces, *e.g.*, monochrome videos, or other sensory modalities.

The key result that led to the evaluation of identification algorithms in the stimulus space is the duality between

- the decoding problem of a single stimulus encoded by a bank of filters in cascade with a population of neurons, and,
- the identification problem of a single receptive field using different stimuli in multiple experimental trials.

We have implicitly assumed that the parameters of the spike generator are always known, even in the functional identification setting. This may not hold true in practice and additional experiments may be required to estimate these parameters before the identification of the receptive field (Lazar &

Slutskiy, 2014b). However, evaluation in the stimulus space is still applicable to assess the identification quality of both filters and neurons.

Our results were formulated using an encoding architecture that preserves the information contained in the input stimuli and therefore perfect reconstruction is possible. We shall expect that this takes place in the early stages of sensory processing, for example, in the retina.

In practice, it may be difficult to access the entire population of neurons of a sensory system. However, as multi-electrode devices become more accessible and powerful, it is possible to sample a subset of neurons that are restricted to certain spatial domains. In addition, a sensory circuit may be further divided into subcircuits based on output cell types (da Silveira & Roska, 2011) and their functionality. The encoding, decoding and identification problem is readily extensible to such problem settings.

Finally, the decoding formalism for evaluating functional identification quality can also be used to estimate the overall bandwidth support of the ensemble of receptive fields of the neural circuit. As we have pointed out, only the projection of the filters can be identified, so the decoding quality is also dependent on the stimulus space one chooses in the identification process. We should expect that for an input space with lower bandwidth, decoding using the identified projection of the receptive field should lead to high reconstruction quality. As the bandwidth of the stimulus space increases, the projection of filters converges to the actual filters. When the bandwidth of the stimulus space exceeds the overall bandwidth of the circuit, the filters can no longer support the entire input (signal) space. The decoding quality will degrade even with a set of known filters. Therefore, the degradation of the decoding quality with identified filters indicates the value of the bandwidth of the receptive field of the encoding circuit.

The scalar-valued RKHS we have focused on in this paper is the space of trigonometric polynomials. The finite dimensionality of this space allowed us to derive bounds on the number of spikes and the number of neurons/trials for perfect reconstruction/identification. The structure of the space also enabled us to use faster algorithms to perform decoding and identification. However, the choice of the base RKHS is flexible and does not exclude infinite dimensional spaces, and the formulation of decoding and functional identification by a variational approach is readily applicable to infinite dimensional spaces. While bounds on number of spikes may no longer be appropriate, the inter-

pretation of the interpolation spline algorithm still holds: the reconstruction is still generated by the subspace spanned by the finite number of sampling functions. That is, based on the observations in the sampling stage.

Acknowledgements

The research reported here was supported by AFOSR under grant #FA9550-12-1-0232 and, in part, by a grant of computer time from the City University of New York HPCC under NSF Grants CNS-0855217 and CNS-0958379. We thank Katsuhiko Inoue for kindly granting us permission to use the stereoscopic video.

Appendix A. Computation of the Sampling Functions and Φ matrix

To compute the entries for matrix $\mathbf{\Phi}$ in (25), we note from (22) that $\phi^j_{ik}(x_1, x_2, t)$ amounts to

$$
\begin{aligned}
&= \int_{t^i_k}^{t^i_{k+1}} \left(\int_{\mathbb{D}} h^i_m(x', y', s - t') K_m(x, y, t; x', y', t') dx' dy' dt' \right) ds \\
&= \int_{t^i_k}^{t^i_{k+1}} \left(\int_{\mathbb{D}} h^i_m(x', y', t') K_m(x, y, t; x', y', s - t') dx' dy' dt' \right) ds \\
&= \sum_{l_x=-L_x}^{L_x} \sum_{l_y=-L_y}^{L_x} \sum_{l_t=-L_t}^{L_t} \int_{t^i_k}^{t^i_{k+1}} \left(\int_{\mathbb{D}} h^i_m(x', y', t') e_{l_x l_y l_t}(x - x', x - y', t + t' - s) dx' dy' dt' \right) ds \\
&= \sum_{l_x=-L_x}^{L_x} \sum_{l_y=-L_y}^{L_y} \sum_{l_t=-L_t}^{L_t} e_{l_1 l_2 l_t}(x, y, t) \int_{t^i_k}^{t^i_{k+1}} \left(\int_{\mathbb{D}} h^i_m(x', y', t') e_{-l_x, -l_y, -l_t}(x', y', s - t') dx' dy' dt' \right) ds \\
&= \sum_{l_x=-L_x}^{L_x} \sum_{l_y=-L_y}^{L_y} \sum_{l_t=-L_t}^{L_t} e_{l_x l_y l_t}(x, y, t) \int_{t^i_k}^{t^i_{k+1}} e_{-l_t}(s) ds \int_{\mathbb{D}} h^i_m(x', y', t') e_{-l_x, -l_y, l_t}(x', y', t') dx' dy' dt'.
\end{aligned}
\tag{A.1}
$$

Since the $e_{l_x l_x l_t}(x, y, t)$'s form the orthonormal base in $\mathcal{H}$, we see that

$$
\phi^i_{mk}(x, y, t) = \sum_{l_x=-L_x}^{L_x} \sum_{l_y=-L_y}^{L_y} \sum_{l_t=-L_t}^{L_t} a^i_{mkl_x l_y l_t} e_{l_x, l_y, l_t}(x, y, t), \tag{A.2}
$$

where $a^i_{mkl_xl_yl_t}$ are the coefficients of the linear combination of bases and

$$a^i_{mkl_xl_yl_t} = \int_{t^j_k}^{t^j_{k+1}} e_{-l_t}(s)ds \cdot \int_{\mathbb{D}} h^i_m(x',y',t')e_{-l_x,-l_y,l_t}(x',y',t')dx'dy'dt', \quad \text{(A.3)}$$

or

$$a^i_{mkl_xl_y,-l_t} = \left(\int_{t^i_k}^{t^i_{k+1}} e_{l_t}(s)ds\right)\left(\int_{\mathbb{D}} h^i_m(x',y',t')e_{-l_x,-l_y,-l_t}(x',y',t')dx'dy'dt'\right). \quad \text{(A.4)}$$

Let

$$h^i_{ml_xl_yl_t} = \int_{\mathbb{D}} h^i_m(x',y',t')e_{-l_x,-l_y,-l_t}(x',y',t')dx'dy'dt', \quad \text{(A.5)}$$

we have

$$a^i_{mkl_xl_yl_t} = \begin{cases} \left(t^i_{k+1} - t^i_k\right) h^i_{ml_xl_y,-l_t}, & l_t = 0 \\ \dfrac{jL_t}{\Omega_t l_t}\left(e_{-l_t}(t_{k+1}) - e_{-l_t}(t_k)\right) h^i_{ml_xl_y,-l_t}, & l_t \neq 0 \end{cases} \quad \text{(A.6)}$$

The computation of the coefficients in (A.5) can be simplified by considering the space-time domain $\mathbb{D}$ to be exactly one period of the function in $\mathcal{H}$, and by numerically evaluating the integral in the second half of (A.4) using the rectangular rule with uniform grid. Since the result is closely related to the 3D-DFT coefficients of $h^i_m(x,y,t)$, these coefficients can be very efficiently obtained. Note also that the $a^i_{mkl_xl_yl_t}$ clearly depends on the particular neuron model and the spatio-temporal receptive field used in the encoding. Equation (A.4) shows, however, that this dependency can easily be separated into two terms. The term in the first parenthesis depends only on the IAF neuron and the term in the second parenthesis depends only on the receptive field.

Therefore,

$$\begin{aligned} \left[\mathbf{\Phi}^{ij}\right]_{kl} &= \langle \boldsymbol{\phi}^i_k, \boldsymbol{\phi}^j_l \rangle_{\mathcal{H}^3} \\ &= \sum_{m=1}^{3} \langle \phi^i_{mk}, \phi^j_{ml} \rangle_{\mathcal{H}} \\ &= \sum_{m=1}^{3} \sum_{l_x=-L_x}^{L_x} \sum_{l_y=-L_y}^{L_y} \sum_{l_t=-l_t}^{l_t} a^i_{mkl_xl_yl_t} \overline{a^j_{mll_xl_yl_t}}. \end{aligned} \quad \text{(A.7)}$$

Appendix B. Proof of Theorem 1

We present the proof of Theorem 1 in this section.

The form of the solution (24) is given by the Representer Theorem and the solution is unique. Substituting (24) into equation (23), the coefficients c_k^i can be obtained by solving the constraint optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\mathbf{c}^T\mathbf{\Phi}\mathbf{c} \\ \text{subject to} & \mathbf{\Phi}\mathbf{c} = \mathbf{q} \end{array}. \tag{B.1}$$

We note that all solutions to $\mathbf{\Phi}\mathbf{c} = \mathbf{q}$ lead to the same value in $\frac{1}{2}\mathbf{c}^T\mathbf{\Phi}\mathbf{c}$. Therefore, the solution to $\mathbf{\Phi}\mathbf{c} = \mathbf{q}$ verifies (B.1). Since $\hat{\mathbf{u}}$ is unique, we shall expect any solution to this system of equations leads to the same $\hat{\mathbf{u}}$.

The necessary condition for perfect recovery can be more readily observed when we consider using the basis representation of $\mathbf{u} = [u_1(x,y,t), u_2(x,y,t), u_3(x,y,t)]^T$ with

$$u_i(x,y,t) = \sum_{l_x=-L_x}^{L_x} \sum_{l_y=-L_y}^{L_y} \sum_{l_t=-L_t}^{L_t} d_{il_xl_yl_t} e_{l_xl_yl_t}(x,y,t), \tag{B.2}$$

Substituting (B.2) into equation (23), the coefficients $d_{il_xl_yl_t}$ in (B.2) have to verify the system of equations

$$\mathbf{\Xi}\mathbf{d} = \mathbf{q}, \tag{B.3}$$

where

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \mathbf{d}_3 \end{bmatrix}$$

with $[\mathbf{d}_i]_l = d_{il}, i = 1,2,3$ and the column index l traverses all possible subscript combinations of l_x, l_y, l_t. $\mathbf{\Xi}$ is a block matrix

$$\mathbf{\Xi} = \begin{bmatrix} \mathbf{\Xi}^1 \\ \vdots \\ \mathbf{\Xi}^N \end{bmatrix},$$

and

$$\mathbf{\Xi}^i = \left[\mathbf{\Xi}_1^i, \mathbf{\Xi}_2^i, \mathbf{\Xi}_3^i\right],$$

with

$$\left[\boldsymbol{\Xi}_m^i\right]_{kl} = \langle \phi_{mk}^i, e_l \rangle \overset{(a)}{=} a_{mkl}^i,$$

for all $i = 1, \cdots, N, m = 1, 2, 3$, and the column index l traverses all possible subscript combinations of l_x, l_y, l_t, and (a) is given by (A.2). Since $\mathcal{H}^3$ is finite dimensional, the two approaches are equivalent in the absence of noise. This can be observed by noticing that $\boldsymbol{\Phi} = \boldsymbol{\Xi}\boldsymbol{\Xi}^H$ (see also (A.7)).

The columns of $\boldsymbol{\Xi}$ are associated with the basis functions in $\mathcal{H}^3$ and the number of variables to be solved in this case is $dim(\mathcal{H}^3)$. To achieve perfect reconstruction of any arbitrary $\mathbf{u}$, $\mathbf{d}$ must be uniquely determined and it is necessary to have $rank(\boldsymbol{\Xi}) = dim(\mathcal{H}^3)$. As each row of $\boldsymbol{\Xi}$ is essentially the sampling function $\boldsymbol{\phi}_k^i$, it is thereby necessary to have the set of sampling functions $\boldsymbol{\phi}$ span $\mathcal{H}^3$. Consequently, the number of rows must be greater than or equal to the number of columns, *i.e.*, the number of basis functions. Therefore, a necessary condition for perfect recovery is that the number of measurements/sampling functions must be at least $dim(\mathcal{H}^3)$. This implies at least $dim(\mathcal{H}^3) + N$ spikes are needed with a neural circuit consists of N neurons. In addition, since each individual neuron encodes a temporal signal, at most $dim_t(\mathcal{H})$ measurements per neuron are informative. Therefore, the number of neurons should at least be $dim(\mathcal{H}^3)/dim_t(\mathcal{H}) = 3 \cdot dim_{xy}(\mathcal{H})$. □

Appendix C. Constructing Gabor Receptive Fields and IAF Neurons

Each receptive field component has a profile modeled as a spatial Gabor filter derived from the mother function

$$D(x, y, \eta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} - \frac{y^2}{8}\right) \cos(-2.5x + \eta),$$

with translations

$$\mathcal{T}_{(x_0,y_0)} D(x, y, \eta) = D(x - x_0, y - y_0, \eta),$$

dilations

$$\mathcal{D}_\alpha D(x, y, \eta) = \frac{1}{\alpha} D\left(\frac{1}{\alpha}x, \frac{1}{\alpha}y, \eta\right),$$

and rotations

$$\mathcal{R}_\theta D(x, y, \eta) = D\left(\cos(\theta)x + \sin(\theta)y, -\sin(\theta)x + \cos(\theta)y, \eta\right).$$

The phase η provides additional flexibility in modeling phase selectivity of Gabor receptive fields (Reid et al., 1991).

We consider an initial orientation θ_m^i and phase η_m^i picked from a uniform distribution $[0, 2\pi)$, as well as two levels of dilation $\alpha_m^i \in \{2^{0.5}, 2^{1.5}\}$, with probability 0.8 and 0.2, respectively. The center coordinates of the red-component receptive fields (x_0^i, y_0^i) are picked randomly from a uniform distribution. The center coordinates of green- and blue-component receptive fields are picked around the red-component center with Gaussian distributions $\mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is 2×2 identity matrix. Note that while the parameters above are randomly chosen, once picked they are assumed to be known (or identifiable).

To create a non-separable spatio-temporal receptive field, we add a temporal component to the Gabor functions such that the receptive fields rotate at an angular speed $v = 2.5\pi(rad/s)$ around their respective centers (x_m^i, y_m^i). Furthermore, the temporal dynamics is modulated by a raised cosine function

$$f(t) = \begin{cases} 1 - \cos(2\pi \cdot 10 \cdot t), & 0 \leq t \leq 0.1[s] \\ 0, & otherwise \end{cases}$$

to ensure that the spatiotemporal receptive field is causal in the time variable and has finite memory.

The overall receptive field can be expressed as

$$h_m^i(x, y, t) = f(t)\mathcal{T}_{x_m^i, y_m^i}\mathcal{D}_{\alpha_m^i}\mathcal{R}_{\theta_m^i + 2.5\pi t} D(x, y, \eta_m^i).$$

The bias, threshold and integration constant of all IAF neurons are picked to be the same, and they are $b^i = 3$, $\delta^i = 0.1$, and $\kappa^i = 1$, respectively. In simulations, since the input video had a frame rate of 100 frames per second, the inputs to the IAF neurons had a time step of 0.01 second. The time occurrences of spikes generated by the IAF neurons are analytically computed using linear interpolation between two consecutive time steps. The encoding process took 5 minutes on a single Nvidia M2050 GPU.

Appendix D. Proof of Equation (29)

Since $\mathcal{H}$ is an RKHS, by the reproducing property we have $u_m^i(x,y,t) = \langle u_m^i, \mathbf{K}_{xyt}\mathbf{e}_m\rangle_{\mathcal{H}}$. It follows that the mth term of the sum in Eq. (27) can be written as

$$\int_{\mathbb{D}} h_m(x,y,t-s)u_m^i(x,y,s)dsdxdy = \int_{\mathbb{D}} h_m(x,y,s)u_m^i(x,y,t-s)dxdyds =$$
$$\overset{(a)}{=} \int_{\mathbb{D}} h_m(x,y,s)u_m^i(x',y',t')K_m(x,y,t-s;x',y',t')dx'dy'dt'dxdyds$$
$$\overset{(b)}{=} \int_{\mathbb{D}} u_m^i(x',y',t')\int_{\mathbb{D}} h_m(x,y,s)K_m(x',y',t-t';x,y,s)dxdydsdx'dy'dt'$$
$$\overset{(c)}{=} \int_{\mathbb{D}} u_m^i(x',y',t')(\mathcal{P}\mathbf{h})_m(x',y',t-t')dx'dy'dt'$$
$$= \int_{\mathbb{D}} u_m^i(x',y',t-t')(\mathcal{P}\mathbf{h})_m(x',y',t')dx'dy'dt',$$

where $^{(a)}$ follows from the reproducing property of the kernel K_m, $^{(b)}$ from the symmetry and the structure of the reproducing kernel $K_m(x,y,t;x',y',t')$ given in (5), and $^{(c)}$ from Definition 2.

Appendix E. Segmenting Continuous Visual Stimuli for Identification

We use a sliding temporal window to create multiple video clips from a single continuous natural video. This is needed to fix one of the complications arising in using natural video with the introduced methodology, namely how to properly segment a long natural sequence into multiple segments of videos. Since the spatio-temporal receptive field has temporal memory of length $S \triangleq \text{supp}(\mathbf{h})$, i.e., it extends S seconds into the past, the timing of a spike at a time t_k is affected by the stimulus on the time interval of length S preceding the spike, i.e., by values of the stimulus $\mathbf{u}(t)$ on $t \in (t_k - S, t_k]$. Therefore, when recording spikes in response to a stimulus $\mathbf{u}(t)$, care should be taken so that the recording is longer than the temporal support of the receptive field and only those spikes occurring S seconds after the start of the recording are used.

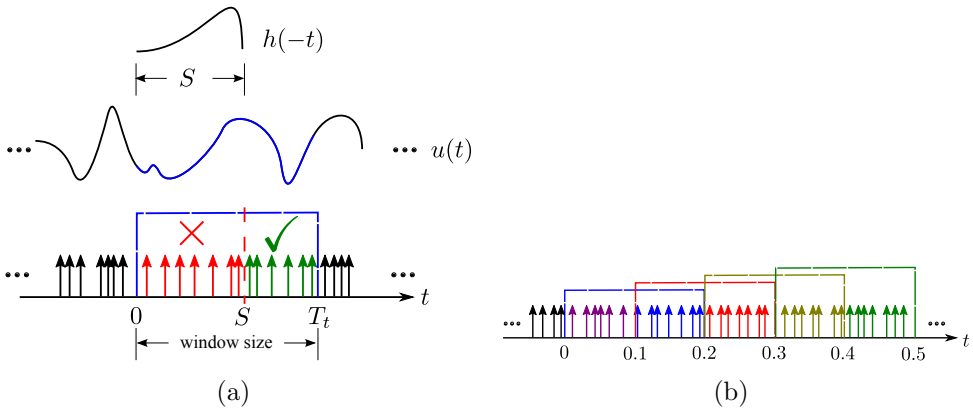


Figure E.19: Schematic illustration of segmenting a continuous visual stimulus used in identification into overlapping video clips and choosing the valid spikes for each video clip. (a) If the temporal support of the receptive field is S (top), one can choose a video clip of duration $T_t > S$. The spikes that are valid for this video clip are shown in green at the bottom. Although the spikes indicated in red are generated during the presentation of this video clip, they are not valid since they contain information outside the duration of this video clip. (b) In identification, a continuous stream of visual stimuli can be presented to the neural circuit. To perform identification, the continuous visual stimulus needs to be segmented into overlapping video clips with appropriate spikes chosen for each of the video clips. Here, an example with window size 0.2 [s] and $S = 0.1$ [s] is given. The cut-off times of each video clip are shown with rectangular boxes of various colors. The valid spikes for each video clip are shown in their respective color. As a result, only a small number of measurements are discarded.

In Figure 19(a), we illustrate the valid spikes given a window size T_t and filter size S: The sliding window is of the same length as the temporal period of the RKHS and is shown in blue. The corresponding clip of the signal $u(t)$ is highlighted by blue. The filter $h(t)$ (the shape of $h(t)$ in Figure 19(a) is shown for illustration purposes only) has temporal support of S. From all spikes generated in $[0, T_t]$, only the spikes generated in the interval $(S, T_t]$ (green spikes) can be used in identification. Phrased differently, if the window size is T_t and one uses spikes generated in the interval $(R, T_t]$ in identification, then the identified receptive field is only valid if the temporal support of it is within $[0, R]$.

The sliding window size we choose is $0.2s$ and the step between windows is $0.1s$, as schematically shown in Figure 19(b), where the color of the spikes indicates its use in the corresponding window. Note that practically no spikes are discarded as the windows overlap.

References

References

Benardete, E. A., & Kaplan, E. (1997). The receptive field of the primate P retinal ganglion cell, i: Linear dynamics. *Visual Neuroscience*, *14*, 169–185.

Berlinet, A., & Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers.

Caponnetto, A., Micchelli, C. A., Pontil, M., & Ying, Y. (2008). Universal multi-task kernels. *Journal of Machine Learning Research*, *9*, 1615–1646.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., & Rust, N. C. (2005). Do we know what the early visual system does? *The Journal of Neuroscience*, *25*, 10577–10597.

Carmeli, C., Vito, E. D., & Toigo, A. (2006). Vector valued reproducing kernel Hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, *4*, 377–408.

Dacey, D. M. (2000). Parallel pathways for spectral coding in primate retina. *Annual Review Neuroscience*, *23*, 743–775.

Freeman, R. D., & Ohzawa, I. (1990). On the neurophysiological organization of binocular vision. *Vision Research*, *30*, 1661–1676.

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, *4*.

Gegenfurtner, K. R., & Kiper, D. C. (2003). Color vision. *Annual Review Neuroscience*, *26*, 181–206.

Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, *10*.

Gollisch, T., & Meister, M. (2010). Eye smarter than scientists believed, neural computations in circuits of the retina. *Neuron*, *65*, 150–164.

Inoue, K. (2009). 3D waltz of the flowers. URL: `http://www.youtube.com/watch?v=0GcLW0g_c1s`.

Lazar, A. A. (2004). Time Encoding with an Integrate-and-Fire Neuron with a Refractory Period. *Neurocomputing*, *58-60*, 53–58. URL: `http://www.sciencedirect.com/science/journal/09252312`.

Lazar, A. A., Li, W., Ukani, N. H., Yeh, C.-H., & Zhou, Y. (2013). Neural circuit abstractions in the fruit fly brain. In *Society for Neuroscience Abstracts*. San Diego, CA.

Lazar, A. A., & Pnevmatikakis, E. A. (2008). Faithful representation of stimuli with a population of integrate-and-fire neurons. *Neural Computation*, *20*, 2715–2744.

Lazar, A. A., & Pnevmatikakis, E. A. (2011). Video time encoding machines. *IEEE Transactions on Neural Networks*, *22*, 461–473.

Lazar, A. A., Pnevmatikakis, E. A., & Zhou, Y. (2010). Encoding natural scenes with neural circuits with random thresholds. *Vision Research*, *50*, 2200–2212.

Lazar, A. A., & Slutskiy, Y. B. (2010). Identifying dendritic processing. *Advances in Neural Information Processing Systems*, *23*, 1261–1269.

Lazar, A. A., & Slutskiy, Y. B. (2012). Channel identification machines. *Journal of Computational Intelligence and Neuroscience*, *2012*, 1–20.

Lazar, A. A., & Slutskiy, Y. B. (2013). Multisensory encoding, decoding, and identification. In *Advances in Neural Information Processing Systems 26*. L. Bottou, C.J.C. Burges, M. Welling and Z. Ghahramani.

Lazar, A. A., & Slutskiy, Y. B. (2014a). Channel identification machines for multidimensional receptive fields. *Frontiers in Computational Neuroscience*, *8*.

Lazar, A. A., & Slutskiy, Y. B. (2014b). Functional identification of spike-processing neural circuits. *Neural Computation*, *26*, 264–305.

Lazar, A. A., & Tóth, L. T. (2004). Perfect recovery and sensitivity analysis of time encoded bandlimited signals. *IEEE Transactions on Circuits and Systems-I: Regular Papers*, *51*, 2060–2073.

Lazar, A. A., & Zhou, Y. (2012). Massively parallel neural encoding and decoding of visual stimuli. *Neural Networks*, *32*, 303–312. Special Issue: IJCNN 2011.

Lazar, A. A., & Zhou, Y. (2014). Reconstructing natural visual scenes from spike times. *Proceedings of the IEEE*, *102*, 1500–1519. doi:10.1109/JPROC.2014.2346465.

Masland, R. H. (2012). The neuronal organization of the retina. *Neuron*, *76*, 266–280.

Qian, N. (1997). Binocular disparity and the perception of depth. *Neuron*, *18*, 359–368.

Reid, R. C., Soodak, R. E., & Shapley, R. M. (1991). Directional selectivity and spatiotemporal structure of receptive fields of simple cells in cat striate cortex. *Journal of Neurophysiology*, *66*, 505–529.

Russell, R., & Sinha, P. (2007). Real-world face recognition: the importance of surface reflectance properties. *Perception*, *36*.

da Silveira, R. A., & Roska, B. (2011). Cell types, circuits, computation. *Current Opinion in Neurobiology*, *21*, 664–671.

Solomon, S. G., & Lennie, P. (2007). The machinery of colour vision. *Nature Review Neuroscience*, *8*, 276–286.

Stanley, G. B., Li, F. F., & Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *The Journal of Neuroscience*, *19*, 8036–8042.

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, *13*, 600–612.

Warland, D. K., Reinagel, P., & Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, *78*, 2336–2350.

Zhu, Y.-D., & Qian, N. (1996). Binocular receptive field models, disparity tuning, and characteristic disparity. *Neural Computation*, *8*, 1611–1641.