A VIDEO TIME ENCODING MACHINE

Aurel A. Lazar and Eftychios A. Pnevmatikakis

Columbia University, Department of Electrical Engineering, New York, NY 10027, USA.
E-mail: aurel@ee.columbia.edu, eap2111@columbia.edu

ABSTRACT

Time encoding is a real-time asynchronus mechanism of mapping analog amplitude information into multidimensional time sequences. We investigate the exact representation of analog video streams with a Time Encoding Machine realized with a population of spiking neurons. We also provide an algorithm that perfectly recovers streaming video from the spike trains of the neural population. Finally, we analyze the quality of recovery of a space-time separable video stream encoded with a population of integrate-and-fire neurons and demonstrate that the quality of recovery increases as a function of the population size.

***Index Terms*—** time encoding, video coding, integrate-and-fire neurons, frames, Gabor wavelets

1. INTRODUCTION

Time Encoding Machines (TEMs) encode analog information in the *time domain* using only *asynchronous* circuits [1]. Representation in the time domain is an alternative to the classical sampling representation in the *amplitude domain*. Applications arise in low power nano-sensors for analog-to-discrete (A/D) conversion as well as in modeling olfactory systems, vision and hearing in neuroscience.

Asynchronous Sigma/Delta modulators as well as FM modulators have been shown to encode information in the time domain [1]. Multichannel TEMs realized with invertible filterbanks and invertible single integrate-and-fire neurons have been investigated in [2]. These TEMs were generalized to population models for single input (SIMO) [3] and for multiple input (MIMO) [4] systems.

In this paper we investigate whether the information contained in the components of an analog video stream can be encoded in the spike trains at the output of an ensemble of integrate-and-fire neurons. In order to do so, we provide a signal recovery scheme based on the spike times of the neural ensemble and derive conditions for *perfect recovery*. The key condition for recovery calls for the spike density of the neural ensemble to be above the Nyquist rate. Our results are based on the theory of frames [5].

Recovery theorems in signal processing with applications to A/D conversion are usually couched in the language of spike/sample density. In neuroscience, however, the natural abstraction is the neuron. We shall formulate here recovery results with conditions on the size of the neural population as opposed to spike density. We demonstrate that, the information contained in the video sensory input can be recovered from the output of the population of integrate-and-fire neurons provided that the number of neurons is beyond a threshold value. Therefore, while information about the signals can not be perfectly represented with a small number of neurons, this limitation can be overcome provided that the number of neurons is beyond a certain critical value. Increasing the number of neurons to achieve a faithful representation of the sensory world is consistent with basic neurobiological thought.

2. TIME ENCODING AND THE *T*-TRANSFORM

Let $\mathcal{H}$ denote the space of (real) analog video streams $I(x,y,t)$ which are bandlimited in time, continuous in space and have finite energy. We assume that the video streams are defined on a bounded spatial set $\mathbb{D}$ which is a compact subset of $\mathbb{R}^2$. By saying bandlimited in time, we mean that for every $(x_0,y_0)\in\mathbb{D}$, we have $I(x_0,y_0,t)\in\Xi$, where Ξ is the space of bandlimited functions of finite energy.

It is clear that the space $\mathcal{H}$, endowed with the inner product $\langle\cdot,\cdot\rangle:\mathcal{H}\times\mathcal{H}\mapsto\mathbb{R}$ defined by

$$\langle I_1, I_2\rangle = \iiint_{\mathbb{R}\otimes\mathbb{D}} I_1(x,y,t)I_2(x,y,t)\,dxdydt$$

is a Hilbert space.

In full generality we assume that each neuron $j, j=1,2,\ldots,N$, has a spatiotemporal receptive field described by the function $D^j(x,y,t)$. Filtering the video stream with the receptive field of neuron j gives the dendritic output $v^j(t)$

$$v^j(t) = \int_{\mathbb{R}}\left(\iint_{\mathbb{D}} D^j(x,y,s)I(x,y,t-s)\,dxdy\right)ds. \quad (1)$$

Subsequently, a constant bias b^j is added to the dendritic output and the sum $v^j(t)+b^j$ is passed through the axon

This work was supported by NIH grant R01 DC008701-01 and NSF grant CCF-06-35252. E. A. Pnevmatikakis was also supported by Onassis Public Benefit Foundation.

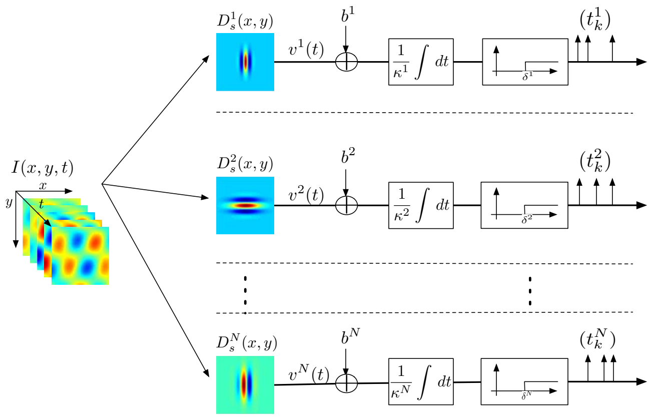

Fig. 1. Video Time Encoding Machine.

hillock. The later is modeled as an integrate-and-fire neuron with threshold δ^j and integration constant κ^j.

Here we are interested in the case where the receptive fields of the neurons have only spatial components, i.e., $D^j(x,y,t) = D_s^j(x,y)\delta(t)$, where $\delta(t)$ is the Dirac function. The neural population encoding model is depicted in Figure 1.

With $(t_k^j), k \in \mathbb{Z}$, the spike train of neuron $j, j = 1,2,\dots,N$, the t-transform can be written as

$$\int_{t_k^j}^{t_{k+1}^j} (v^j(s) + b^j)\, ds = \kappa^j \delta^j. \tag{2}$$

In inner product form the above equation can be written as

$$\langle I, \phi_k^j \rangle = q_k^j, \tag{3}$$

with $q_k^j = \kappa^j\delta^j - b^j(t_{k+1}^j - t_k^j)$. The sampling functions above are given by

$$\phi_k^j(x,y,t) = \tilde{D}^j(x,y,t) * g * 1_{[t_k^j, t_{k+1}^j]}(t), \tag{4}$$

where $\tilde{D}^j(x,y,t) = D^j(x,y,-t)$ and $g(t) = \sin(\Omega t)/\pi t$ is the impulse response of a low pass filter (LPF) with cutoff frequency Ω. For the case of spatial receptive fields, the sampling functions simply become $\phi_k^j(x,y,t) = D_s^j(x,y) \cdot \left(g * 1_{[t_k^j, t_{k+1}^j]}\right)(t)$.

3. TIME DECODING AND THE T-TRANSFORM INVERSE

For recovery we use the sequence of functions $(\psi_k^j), j = 1,2,\dots,N, k \in \mathbb{Z}$, with

$$\psi_k^j(x,y,t) = D^j(x,y,t) * g(t - s_k^j). \tag{5}$$

Lemma 1. *Assume that the filters modeling the receptive fields $D^j(x,y,t), j = 1,2,\dots,N$, are BIBO stable and their space-time frequency support is a superset of the space-time frequency range of interest. Then if the spike density and the number of neurons N are sufficiently large, the sequences (ϕ_k^j) and $(\psi_k^j), k \in \mathbb{Z}, j = 1,2,\dots,N$, are frames.*

Proof: For video streams discretized in space, the Nyquist rate can be bounded from above by $M\Omega/\pi$, where M is the number of pixels (spatial resolution) and Ω is the temporal bandwidth of the video. If the total spike density exceeds the Nyquist rate then, under mild conditions, the two sequences (ϕ_k^j) and $(\psi_k^j), k \in \mathbb{Z}, j = 1,2,\dots,N$, constitute frames [4]. For continuous video streams bandlimited in space, the Nyquist rate is again bounded and depends on the spatial and temporal bandwidths. □

Theorem 1. *Under the same assumptions as in Lemma 1, if $\sum_{j=1}^N b^j/\kappa^j\delta^j$ diverges in N, then there exists a number $\mathcal{N}$ such that if $N \geq \mathcal{N}$, the video stream $I = I(x,y,t)$, can be perfectly recovered as*

$$I(x,y,t) = \sum_{j=1}^{N} \sum_{k \in \mathbb{Z}} c_k^j \psi_k^j(x,y,t), \tag{6}$$

where the c_k^j, $k \in \mathbb{Z}$, $j = 1,2,\dots,N$, are suitable coefficients.

Proof: Clearly, the theorem holds if we can show that the sequence of functions $(\psi_k^j), k \in \mathbb{Z}, j = 1,2,\dots,N$, forms a frame for $\mathcal{H}$. By following the same procedure as in [4] the total spike density amounts to $\sum_{j=1}^N b^j/\kappa^j\delta^j$. Since this sum diverges, the result follows from Lemma 1. □

Corollary 1. *Let $[\mathbf{c}^j]_k = c_k^j$ and $\mathbf{c} = [\mathbf{c}^1, \mathbf{c}^2, \dots, \mathbf{c}^N]^T$. The coefficients $\mathbf{c}$ can be computed as*

$$\mathbf{c} = \mathbf{G}^+\mathbf{q}, \tag{7}$$

where T denotes the transpose, $\mathbf{q} = [\mathbf{q}^1, \mathbf{q}^2, \dots, \mathbf{q}^N]^T$, $[\mathbf{q}^j]_k = q_k^j$ and $\mathbf{G}^+$ denotes the pseudoinverse of $\mathbf{G}$. The entries of the matrix $\mathbf{G}$ are given by

$$\mathbf{G} = \begin{bmatrix} \mathbf{G^{11}} & \mathbf{G^{12}} & \dots & \mathbf{G^{1N}} \\ \mathbf{G^{21}} & \mathbf{G^{22}} & \dots & \mathbf{G^{2N}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G^{N1}} & \mathbf{G^{N2}} & \dots & \mathbf{G^{NN}} \end{bmatrix}, \tag{8}$$

$$[\mathbf{G}^{ij}]_{kl} = \int_{t_k^i}^{t_{k+1}^i} D^{ij}(s) * g(s - s_l^j)\, ds,$$

where

$$D^{ij}(s) = \iint_{\mathbb{D}} D^i(x,y,s) * D^j(x,y,s) dxdy.$$

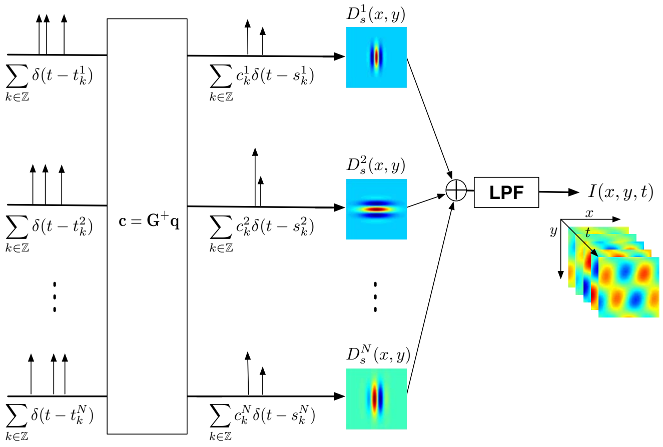
$D_s^1(x,y)$
$D_s^2(x,y)$
$D_s^N(x,y)$
$\sum_{k\in\mathbb{Z}} \delta(t - t_k^1)$
$\sum_{k\in\mathbb{Z}} \delta(t - t_k^2)$
$\sum_{k\in\mathbb{Z}} \delta(t - t_k^N)$
$\mathbf{c} = \mathbf{G}^{+}\mathbf{q}$
$\sum_{k\in\mathbb{Z}} c_k^1\delta(t - s_k^1)$
$\sum_{k\in\mathbb{Z}} c_k^2\delta(t - s_k^2)$
$\sum_{k\in\mathbb{Z}} c_k^N\delta(t - s_k^N)$
LPF
$I(x,y,t)$
x
y
t

Fig. 2. Video Time Decoding Machine.

Proof: Equation $\mathbf{c} = \mathbf{G}^{+}\mathbf{q}$ can be obtained by substituting the representation of $\mathbf{u}$ in equation (6) into the equation of the t-transform in (2). Since the sequences (ϕ_k^j) and $(\psi_k^j), k \in \mathbb{Z}, j = 1, 2, \ldots, N$, are frames for $\mathcal{H}$, the result follows [6]. The decoding scheme is depicted in Figure 2. □

Remark 1. *If the receptive field is only spatial*

$$\psi_k^j(x,y,t) = D_s^j(x,y)g(t - s_k^j)$$

$$[\mathbf{G}^{ij}]_{kl} = \iint_{\mathbb{D}} D_s^i(x,y)D_s^j(x,y)dxdy \int_{t_k^i}^{t_{k+1}^i} g(s - s_l^j)\, ds.$$

4. EXAMPLE

In what follows we provide an example of encoding of space-time separable video streams with a model of simple cells arising in the primary visual cortex. These neurons show a linear summation cross the spatial receptive field [7].

4.1. Constructing Receptive Fields for Simple Cells

Gabor functions have been extensively used to model the spatial receptive fields of simple cells [7]. The general form of a Gabor function is

$$S(x,y) \propto \exp\left(-\frac{(x-x_0)^2}{\sigma_x^2} - \frac{(y-y_0)^2}{\sigma_y^2}\right) e^{i(kx+\nu y+\phi)},$$

where (x_0, y_0) is the center of the receptive field in the spatial domain and (k, ν) is the optimal frequency of the filter in the frequency domain. σ_x, σ_y are the standard deviations along x and y, and ϕ is the preferred spatial phase.

The Gabor wavelet filterbank that matches neural data can be generated from the following mother Gabor wavelet [8]

$$\psi(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{8}(4x^2 + y^2)\right)\left(e^{ikx} - e^{-k^2/2}\right). \tag{9}$$

For real signals the Gabor mother wavelet decomposes into two mother wavelets consisting of its real and imaginary parts. The admissible operations that can be performed on the mother wavelet to obtain the other functions of the wavelet family are

1. Dilation $\mathcal{D}_\alpha, \alpha \in \mathbb{R}\backslash\{0\}$, with
   $\mathcal{D}_\alpha\psi(x,y) = |\alpha|^{-1}\psi\left(\frac{x}{\alpha}, \frac{y}{\alpha}\right)$.
2. Translation $\mathcal{T}_{(x_0,y_0)}, (x_0, y_0) \in \mathbb{R}^2$, with
   $\mathcal{T}_{(x_0,y_0)}\psi(x,y) = \psi(x - x_0, y - y_0)$.
3. Rotation $\mathcal{R}_\theta, \theta \in [0, 2\pi)$, with
   $\mathcal{R}_\theta\psi(x,y) = \psi(x\cos\theta + y\sin\theta, -x\sin\theta + y\cos\theta)$.

In order to model spatial receptive fields with Gabor filters, we need to ensure that the set of the Gabor filters spans the space of interest, i.e., it forms a frame for the space of interest (usually $L^2(\mathbb{D})$, where $\mathbb{D} \subset \mathbb{R}^2$ compact). A set of discrete values for α, θ, x_0, y_0, can be obtained with the following scheme:

- $\alpha = \alpha_0^m, \alpha_0 > 1, m \in \mathbb{Z}$,
- $\theta = l\theta_0, l \in [0, 1, \ldots, L-1], \theta_0 = 2\pi/L$,
- $(x_0, y_0) = (nb_0\alpha_0^m, kb_0\alpha_0^m), (n,k) \in \mathbb{Z}^2$.

For the above set of discrete values Lee [8] found conditions that guarantee that the set of Gabor functions represent a frame for $L^2(\mathbb{R}^2)$. Moreover, he approximated the frame bounds and showed that as the density of the filterbank increases, i.e., the parameters α_0, θ_0, b_0 become smaller, the frame becomes tighter and eventually practically tight.

4.2. Video Time Encoding with a Gabor Filterbank

We consider a video stream $I = I(x,y,t)$ defined over the spatial domain $\mathbb{D} = [-3,3] \times [-3,3]$ and bandlimited in time to $\Omega = 2\pi\cdot 20$ Hz. The video stream has separable spatial and temporal components, i.e., $I(x,y,t) = S(x,y)u(t)$ with a spatial resolution of 51×51 pixels. We used respectively 36 and 108 neurons to encode the video with spatial receptive fields constructed as explained above. As an example, for the case of 108 neurons the parameters of the filterbank were chosen to be

- $\alpha = \alpha_0^m, \alpha_0 = \sqrt{2}, m \in [0, 1]$,
- $\theta = l\theta_0, \theta_0 = 2\pi/3, l \in [0, 1, 2]$,
- $(x_0, y_0) = (nb_0\alpha_0^m, kb_0\alpha_0^m), b_0 = 1$,
  $(n,k) \in [-1, 0, 1]^2$.

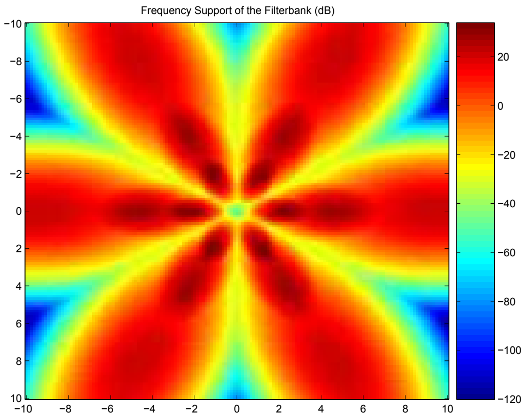

Fig. 3. Frequency support for the spatial Gabor filterbank.

Figure 3 shows the frequency support of the constructed filterbank. Figure 4 depicts the recovery results for the spatial component of the video stream. Increasing the number of neurons in the architecture results in a recovery with an improved spatial resolution. In general, the minimum number of neurons that is required for perfect recovery depends in a linear fashion on the spatial bandwidth of the video stream and on the area of the spatial domain $\mathbb{D}$, upon which the video stream is defined.

5. CONCLUSIONS

We presented a model of encoding an analog video stream into a multidimensional time sequence using an asynchronous real-time circuit. In its quantized form, the time sequence $(t_k^j), k \in \mathbb{Z}, j = 1, 2, \ldots, N,$ can be used for transmission and for further processing in any digital communications and/or signal processing system. Thus, Video TEMs with quantized output play the same role as A/D converters in conventional signal processing systems but with the added benefit of being asynchronous.

For SIMO TEMs [3] the quality of stimulus reconstruction gracefully degrades when additive white noise is present at the input. The structurally closely related recovery algorithm of the Video Time Decoding Machine is robust with respect to additive noise as well (data not shown).

Video TEMs may also be used as a template architecture for the design of purely nonlinear brain-machine interfaces with high performance characteristics. Current models of brain-machine-interfaces are based on a linear architecture and exhibit limited performance [9].

From a neuroscience standpoint our model provides theoretical support for modeling arbitrary linear operators associated with dendritic trees. The analysis described here demonstrates that the visual sensory world can be faithfully represented by a population of neurons, provided that the number of neurons is above a critical value.

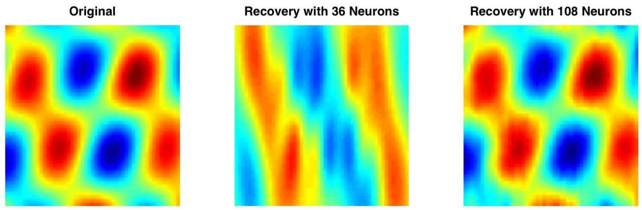

Fig. 4. Recovery of the spatial component for a space-time separable video stream. The quality of recovery increases with the number of neurons.

6. REFERENCES

[1] A.A. Lazar and L.T. Tóth, "Perfect Recovery and Sensitivity Analysis of Time Encoded Bandlimited Signals," *IEEE Transactions on Circuits and Systems-I: Regular Papers*, vol. 51, no. 10, pp. 2060–2073, October 2004.

[2] A.A. Lazar, "Multichannel Time Encoding with Integrate-and-Fire Neurons," *Neurocomputing*, vol. 65-66, pp. 401–407, 2005.

[3] A.A. Lazar and E.A. Pnevmatikakis, "Faithful Representation of Stimuli with a Population of Integrate-and-Fire Neurons," *Neural Computation*, 2008, to appear.

[4] A.A. Lazar and E.A. Pnevmatikakis, "A MIMO Time Encoding Machine," Submitted, 2008.

[5] O. Christensen, *An Introduction to Frames and Riesz Bases*, Applied and Numerical Harmonic Analysis. Birkhäuser, 2003.

[6] Y.C. Eldar and T. Werther, "General Framework for Consistent Sampling in Hilbert Spaces," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 3, no. 3, pp. 347–359, 2005.

[7] P. Dayan and L.F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, MIT Press, 2001.

[8] T.S. Lee, "Image Representation Using 2D Gabor Wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[9] S.-P. Kim, J.C. Sanchez, Y.N. Rao, D. Erdogmus, J.M. Carmena, M.A. Lebedev, M.A.L. Nicolelis, and J.C. Principe, "A Comparison of optimal MIMO Linear and Nonlinear Models for Brain-Machine Interfaces," *Journal of Neural Engineering*, vol. 3, pp. 145–161, 2006.