The power of connectivity: Identity preserving transformations on visual streams in the spike domain

Aurel A. Lazar[c,1,*], Eftychios A. Pnevmatikakis[d,1], Yiyin Zhou[c,1]

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

[b] *Department of Statistics and Center for Theoretical Neuroscience, Columbia University, New York, NY, USA*

[*] Corresponding author. Tel: +1 212-854-1747; fax: +1 212-932-9421

Email addresses: aurel@ee.columbia.edu (Aurel A. Lazar), eftychios@stat.columbia.edu (Eftychios A. Pnevmatikakis), yiyin@ee.columbia.edu (Yiyin Zhou)

[1] The author's names are alphabetically listed.

Preprint submitted to Neural Networks, accepted for publication. *February 26, 2013*

The power of connectivity: Identity preserving transformations on visual streams in the spike domain

Aurel A. Lazar[c,1,*], Eftychios A. Pnevmatikakis[d,1], Yiyin Zhou[c,1]

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

[d] *Department of Statistics and Center for Theoretical Neuroscience, Columbia University, New York, NY, USA*

Abstract

We investigate neural architectures for identity preserving transformations (IPTs) on visual stimuli in the spike domain. The stimuli are encoded with a population of spiking neurons; the resulting spikes are processed and finally decoded. A number of IPTs are demonstrated including faithful stimulus recovery, as well as simple transformations on the original visual stimulus such as translations, rotations and zooming. We show that if the set of receptive fields satisfies certain symmetry properties, then IPTs can easily be realized and additionally, the same basic stimulus decoding algorithm can be employed to recover the transformed input stimulus. Using group theoretic methods we advance two different neural encoding architectures and discuss the realization of exact and approximate IPTs. These are realized in the spike domain processing block by a "switching matrix" that regulates the input/output connectivity between the stimulus encoding and decoding blocks. For example, for a particular connectivity setting of the switching matrix, the original stimulus is faithfully recovered. For other settings, translations, rotations and dilations (or combinations of these operations) of the original video stream are obtained. We evaluate our theoretical derivations through extensive simulations on natural video scenes, and discuss implications of our results on the problem of invariant object recognition in the spike domain.

Keywords: Identity Preserving Transformations, Invariant Representations,

[*] Corresponding author. Tel: +1 212-854-1747; fax: +1 212-932-9421

Email addresses: aurel@ee.columbia.edu (Aurel A. Lazar), eftychios@stat.columbia.edu (Eftychios A. Pnevmatikakis), yiyin@ee.columbia.edu (Yiyin Zhou)

[1] The author's names are alphabetically listed.

Spiking Neurons, Time Encoding Machines, Group Theory, Connectivity.

1. Introduction

The brain must be capable of forming object representations that are invariant with respect to the large number of fluctuations occurring on the retina (DiCarlo & Cox, 2007). These include object position, scale, pose and illumination, and the presence of clutter. In a simple model of the visual system in primates, the incoming visual stimulus is first represented in the responses of the retinal ganglion cells (RGC). Subsequently, the stimulus is re-represented at each neural layer starting with the first relay center (LGN) and followed by the visual cortex (V1, V2, V4 and IT cortex). Each of these representations can be modeled as an Identity Preserving Transformation (IPT). At the final stage, the visual objects are represented in a way that is amenable to an efficient comparison with an internal (memory) representation of the object. Since spike trains are the language of the brain, the latter representation is in the form of a neural population activity. Consequently, the decision whether the object is present or absent takes place in the spike domain (Logothetis & Sheinberg, 1996).

What are some plausible computational or neural mechanisms by which invariance could be achieved? An early pioneering work (Olshausen et al., 1993) provides a model mechanism for shifting and rescaling the representation of an object from its retinal reference frame into an object-centered reference frame (see also Anderson & Essen (1987)). In one class of models used in the invariant recognition literature, transformations of the incoming visual signal are matched with an existing stored version of the image (Bülthoff & Edelman, 1992). More formally, let I be a visual sensory object (stimulus). An IPT acting on I is modeled as an invertible transformation $\mathcal{T}$ that, in turn, consists of a composition of a set of elementary operators (e.g., rotation, dilation, translation, etc.). The set of all spike trains produced by $\mathcal{T}(I)$ for all possible IPTs $\mathcal{T}$ defines the object-manifold. For identifying the instantiation of a stored object in the incoming object-manifold, the algorithm presented in Arathorn (2002) calls for the identification of the operator $\mathcal{T}$ (and its inverse). More recent research focuses on routing/connectivity operators in support of information delivery (e.g., sensory information) to higher brain centers (Wolfrum & von der Malsburg, 2007).

In this paper we focus on the *realization* of IPTs in the *spike domain*. The spike domain is a non-linear, stimulus-dependent representation space. The non-linear nature of the stimulus representation has proven to be a major challenge for spike

domain computation. Our goal here is to put forth the first efficient rigorous computational model that allows formal reasoning in the spike domain while at the same time it is biologically relevant. Our model of computation can briefly be summarized in block diagram form in Figure 1. The input visual stimulus is encoded in the time domain by an instantiation of a Video Time Encoding Machine (Video TEM) (Lazar & Pnevmatikakis, 2011b). Video TEMs are spatio-temporal models of neural encoding that are realized with receptive fields in cascade with a population of spiking neurons (Lazar et al., 2010). The encoded stimulus (in the form of spikes) is first processed in the Time Domain Processing (TDP) block and then decoded by a Time Decoding Machine (TDM). The output of the TDM is again an analog signal. Our time (spike) domain computation chain resembles the traditional digital signal processing (Oppenheim et al., 1999) chain where an analog signal is converted into a digital signal using an analog-to-digital converter, then processed with a digital signal processor and finally converted back to an analog signal with a digital-to-analog converter.

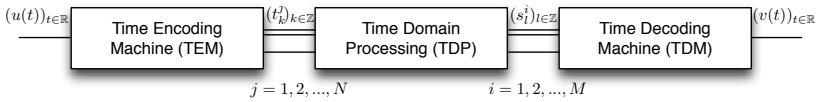


Figure 1: **General Signal Processing Chain with a Time Domain Core.**

An example of processing in the time domain appeared in Lazar (2006) where it was demonstrated how an arbitrary linear filter can be implemented in the time domain, using neural components. By building upon these results, any IPT acting on the input stimulus can be realized in the time domain. However, the setup of Lazar (2006) is rather complex as it requires a different TDM for any desired transformation of the sensory stimulus.

There are two types of operators that are used for encoding of stimuli with TEMs: linear operators (receptive fields) followed by non-linear operators (spiking neural circuits). These operators are cascaded. The efficient realizability of IPTs presented here is primarily due to the structure of the receptive fields of the Video TEM. These are required to form an overcomplete spatial (or spatiotemporal) filterbank. Furthermore the set of receptive fields has to exhibit certain symmetry properties (in group theoretic sense). If the receptive fields (linear filters) have a group structure transformations on the stimulus can be realized via transformations on the filters. However, these group operations cannot be, in general, "propagated" through the neural encoding circuits (formally non-linear operators). Surprisingly, however, under certain conditions described in the paper, rotations, scaling and translations can be efficiently executed in the spike domain.

We show that a large class of IPTs can be efficiently realized by making connectivity changes in the TDP block while the TDM block remains the same. The TDP block consists of a "switching matrix" that simply regulates the connectivity between the TEM and TDM blocks. We will show that different IPTs can be realized with different connectivity settings of the switching matrix. For example, for a particular setting of the switching matrix, the original stimulus is faithfully recovered. For other settings, translations, rotations and dilations (or combinations of these transformations) of the original video stream are obtained. We will also show that IPTs can be computed in parallel.

Our model can be viewed as a generalization of the shifting and rescaling mechanisms proposed in Olshausen et al. (1993). We extend these operations to include rotations and show how to efficiently implement them in the spike domain. We also discuss the constraints that the finite size of the neural population imposes on the set of achievable transformations. By starting from the continuous group on the plane characterizing all the possible IPTs, we advance two different encoding architectures whose receptive fields are defined on two different discrete grids. The first is a log-polar grid, similar to the ones used in models of foveated vision (Nattel & Yeshurun, 2000; Wohrer & Kornprobst, 2009; Weber & Triesch, 2009). On the log-polar grid the switching matrix can realize combinations of rotations and dilations in a lossless manner in the spike domain. The second is a Cartesian grid (Lee, 1996; Field & Chichilnisky, 2007). On the latter grid the switching matrix can realize combinations of dilations and translations in a lossless manner in the spike domain as well. Finally, we discuss how discrete approximations of the continuous symmetry group can be used to perform arbitrary but approximate IPTs in the spike domain. Examples are given that intuitively demonstrate our methodology.

2. Methods

2.1. The Architecture of the Model of Computation

An illustration of a general switching ("rewiring") architecture for encoding, processing and decoding video streams is shown in Figure 2. Our architecture follows the general one depicted in Figure 1.

The input signal is an analog video stream and is encoded by a canonical Video Time Encoding Machine (see Figure 2). A more formal overview of Video TEMs is available in Appendix A.1. Briefly, the Video TEM consists of a bank of linear filters/receptive fields $D^j(x, y, t), j = 1 \cdots, N$, in cascade with nonlinear spiking circuits (e.g., neural circuits realized with Integrate-and-Fire neurons). Hence, a

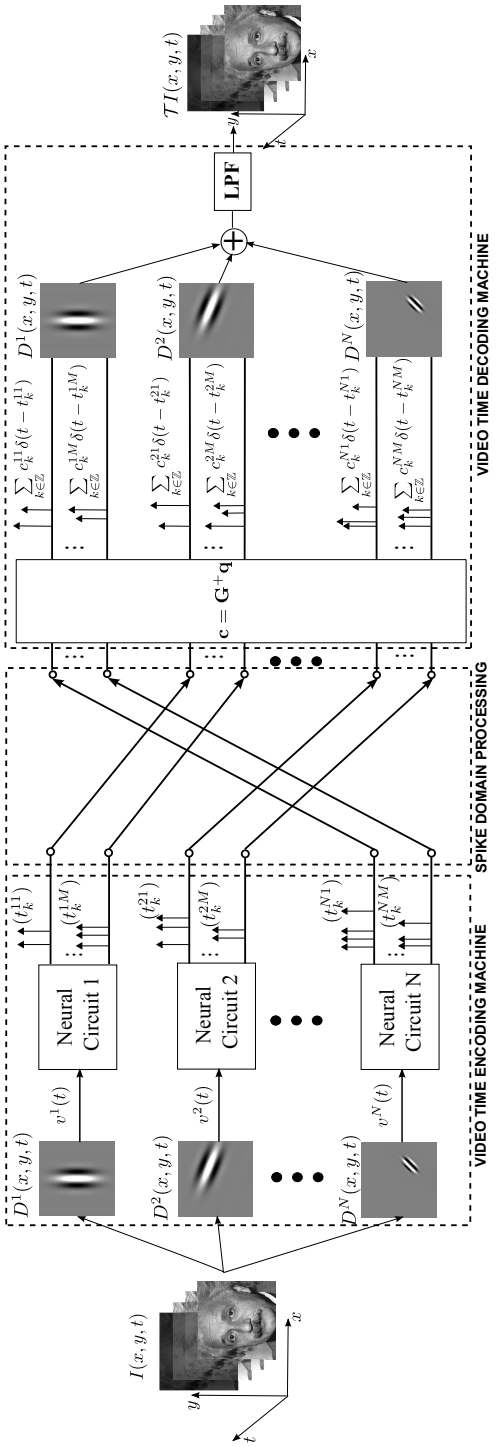


Figure 2: **General architecture of the video processing mechanism using spike domain switching techniques.**

Video TEM maps an input visual stimulus into a vector of spike trains. The spiking activity of the neural circuits can be interpreted as signal dependent sampling. This sampling operation, defined as the t-transform, is expressed by the following set of equations

$$\left\{ \langle I, \phi_k^j \rangle = q_k^j, k \in \mathbb{Z}, j = 1, \ldots, N \right\}, \tag{1}$$

where $I = I(x, y, t)$ is the input visual stimulus that belongs to a Hilbert space, ϕ_k^j is the sampling function associated with k-th spike of neuron j and q_k^j is its measurement (the projection of I onto the sampling function). The sampling functions are determined by the linear receptive fields, the spike times and the parameters of the neural circuits, whereas the outcomes of these projections depend on the spike times and the parameters of the neural circuits. Although the left-hand-side of equation (1) is an inner product, the sampling by the neural circuits is highly nonlinear, and the sampling functions are, through the spike times, stimulus dependent.

The TDM block implements decoding algorithms for the canonical Video TEM (see Figure 2). A more formal overview of Video TDMs is available in Appendix A.2. Under certain conditions the Video TEMs can faithfully encode the input video stream as a multidimensional sequence of spike trains. The TDM architecture implements a perfect decoding algorithm of the input video stream (see Appendix A.2). Briefly, the faithful representation condition ensures that (i) the set of linear receptive fields does not filter out any spatial information contained in the input stimulus (Lazar & Pnevmatikakis (2011a)) and (ii) the spiking frequency of the neurons is high enough so that it can represent the temporal information of the stimulus (Lazar & Pnevmatikakis (2011b)). (See also Appendix A.2). For the rest of this paper we assume that the number and the parameters of the neurons are such that the perfect stimulus recovery conditions are satisfied.

The architecture of the TDP block of Figure 2 is very simple. It consists of a *switching matrix* that regulates the connectivity between the TEM and TDM blocks. In other words, the switching matrix, directs the incoming spikes from the layer of neural circuits that represents the incoming video stimulus, to specific locations of the next layer, i.e., the TDM block or other layers of read-out neurons. For a finite number of N circuits, the switching matrix can have $N!$ different settings. Each setting corresponds to a permutation σ of the numbers $\{1, 2, \ldots, N\}$ mapping the spikes coming from the neural circuit j to the $\sigma(j)$-th entry of the next block or layer. Clearly such a transformation, although non-linear in general, is identity preserving in the time domain because it is invertible through the permutation σ^{-1}. However not all of these transformations have a clear physical interpretation. Moreover, as discussed in Section 1 the representations in the visual system have

to be invariant with respect to certain transformations. These include rotation, dilation (scaling) and translation among others. In what follows we show how the structure of the Video TEM together with the operation of the switching matrix can give rise to such invariant representations.

2.2. Identity-Preserving Transformations of Visual Streams in the Spike Domain

In this section we present the general architecture for the realization of Identity-Preserving Transformations (IPTs) in the spike domain by means of switching mechanisms that regulate connectivity. We argue that IPTs naturally arise in the time domain representation when using a Video TEM provided that the set of receptive fields has some special symmetry properties (sections 2.2.1 and 2.2.2). We present two different sets of receptive fields that can realize various IPTs in an exact form (sections 2.2.3 and 2.2.4) and also describe the realization of approximate arbitrary IPTs by our architecture (section 2.2.5). Finally, we assume that all spike generation model neurons have the same parameters (i.e., same threshold, bias and time constants for the case of Integrate-And-Fire (IAF) neurons and the same feedback loop if any).

2.2.1. The Structure of Receptive Fields

We consider receptive fields that are space-time separable, i.e., they satisfy

$$D^j(x, y, t) = D_s^j(x, y) D_\tau(t) \tag{2}$$

for all $j, j = 1, 2, \ldots, N$. Moreover, we assume that the temporal component D_τ is the same for all the receptive fields, and that its spectral support covers the frequency band of interest $[-\Omega, \Omega]$. Thus, there is no information loss due to temporal filtering.

To generate a set of spatial receptive fields, we pick a *mother* function $\eta \in L^2(\mathbb{R}^2)$, similar to the mother wavelet in wavelet theory (Daubechies, 1992). Then, each individual receptive field is obtained by applying to η the unitary operator

$$\mathcal{T}([x_0, y_0], \alpha, \theta)\eta(x, y) = \tau_{x_0, y_0} D_\alpha R_\theta \eta(x, y), \tag{3}$$

where

1. $\tau_{x_0,y_0}, (x_0, y_0) \in \mathbb{R}^2$ with $\tau_{(x_0,y_0)}\eta(x, y) = \eta(x - x_0, y - y_0)$ is the translation operator;
2. $\mathcal{D}_\alpha, \alpha > 0$ with $\mathcal{D}_\alpha \eta(x, y) = \alpha^{-1} \eta\left(\frac{x}{\alpha}, \frac{y}{\alpha}\right)$ is the dilation operator;

3. $R_\theta, \theta \in [0, 2\pi)$, with $R_\theta \eta(x, y) = \eta(r_{-\theta}[x, y])$, where $r_\theta[x, y] = [x \cos\theta - y \sin\theta, x \sin\theta + y \cos\theta]$ is the rotation operator.

We *define* the operator of (3) to be equal to

$$\mathcal{T}([x_0, y_0], \alpha, \theta)\eta(x, y) = \alpha^{-1}\eta(\alpha^{-1} r_{-\theta}(x - x_0, y - y_0)). \tag{4}$$

In fact, the family of the operators $\mathcal{T}([x_0, y_0], \alpha, \theta)$ is the unique (up to unitary equivalence) unitary irreducible representation of a group called the *similitude group* SIM(2) (Antoine et al., 2004) (see group law in Appendix B).

By denoting the receptive field $\eta_g, g = ([x_0, y_0], \alpha, \theta)$, where

$$\eta_g(x, y) = \mathcal{T}(g)\eta(x, y), \tag{5}$$

we notice that the action of another operator $\mathcal{T}(g')$ on η_g will result in

$$\mathcal{T}(g')\eta_g = \eta_{g' \circ g}, \tag{6}$$

where the subscript of the resulting receptive field is given by the group law. With the above receptive field structure, we present next the generation of invariant transformations that are based on a switching mechanism.

2.2.2. Generation of Invariant Transformations

Let us denote by S a subset of the SIM(2) group, and assume that a signal of interest $A(x, y)$ (for simplicity we assume a constant image, although this approach is readily applicable to time-varying signals as well) can be represented as

$$A(x, y) = \sum_{s \in S} c_s \eta_s(x, y). \tag{7}$$

To apply a transformation $\mathcal{T}(g), g \in H \subseteq SIM(2)$, to A, we have

$$\begin{aligned} \mathcal{T}(g)A(x, y) &= \sum_{s \in S} c_s \mathcal{T}(g)\eta_s(x, y) \\ &= \sum_{s \in S} c_s \eta_{g \circ s}(x, y) \end{aligned} \tag{8}$$

If, in addition, $g \circ s \in S$ and $g^{-1} \circ s \in S$ for all $s \in S$, then the subset S is invariant under the action of $\mathcal{T}(g)$. Consequently, the transformation of A by $\mathcal{T}(g)$ can be rewritten as

$$\mathcal{T}(g)A(x, y) = \sum_{s \in S} c_{g^{-1} \circ s} \eta_s(x, y) \tag{9}$$

Therefore, we see that the transformed image $\mathcal{T}(g)A$ takes the same representation as A itself. The only changes are the coefficients $c_s, s \in S$. As indicated in (9), the coefficient of the receptive field η_s in $\mathcal{T}(g)A$, should be the coefficient of the receptive field $\eta_{g^{-1}\circ s}$ in A, for all $s \in S$. This is what the switching architecture of the spike domain process requires, *i.e.*, any IPT on the set of receptive fields maps to the same set of receptive fields. Once we know the representation of (7), the transformation can be easily realized with a switching circuit.

The invariance of the set S under the action of any $\mathcal{T}(g), g \in H$, is obvious if $H = S = SIM(2)$. In essence, the closure property of the group operation guarantees that the transformation of a receptive field by the unitary operator is another receptive field in the group. It is certainly true that under the continuous group structure, if there are uncountably many receptive fields $\eta_{([x_0,y_0],\alpha,\theta)}(x, y),$ for all $[x_0, y_0] \in \mathbb{R}^2, \alpha \in \mathbb{R}^*_+, \theta \in [0, 2\pi)$, then all rotations, translations and dilations can be implemented by a switching mechanism.

However, there is only a finite number of receptive fields since there is a finite number of neurons. Theoretically, we can make this number countably infinite, but it still requires the discretization of the SIM(2) group, and the restriction of all possible operations to a discrete set. Unfortunately, discretization breaks the nice properties of the continuous group and specific designs of the discretization and switching mechanisms are necessary. We present here a discretization based on a log-polar grid that is invariant under the action of a discrete set of rotations and dilations, and discuss the transformations that it can realize. In section 2.2.4 we present an alternative discretization that is based on a Cartesian grid and is invariant under the action of a discrete set of translations and dilations.

2.2.3. Exact IPTs Using a Log-Polar Grid

In the case of the log-polar grid, the centers of the receptive fields are placed in a rotation-invariant grid. More specifically, the discretization of the SIM(2) group is given by the subset (see also Figure S1 in the supplementary material)

$$S_p = \left\{ (\alpha_0^m r_{l\theta_0}[kb_0, 0], \alpha_0^m, n\omega_0 + l\theta_0) \middle| \begin{array}{l} b_0 > 0, k \in \mathbb{N}, m \in \mathbb{Z}, \alpha_0 > 1 \\ \omega_0 = 2\pi/N, N \in \mathbb{N}^*, n = 0, \cdots, N-1 \\ \theta_0 = 2\pi/L, L \in \mathbb{N}^*, l = 0, \cdots, L-1 \end{array} \right\}. \tag{10}$$

Having (4) in mind, we have that the general form of the receptive field constructed according to S_p is given by

$$\eta_{(\alpha_0^m r_{l\theta_0}[kb_0,0],\alpha_0^m,n\omega_0+l\theta_0)} = \alpha_0^{-m}\eta(\alpha_0^{-m} r_{-(n\omega_0+l\theta_0)}([x, y] - \alpha_0^m r_{l\theta_0}[kb_0, 0])) \tag{11}$$

We see that for each scale α_0^m, this set contains receptive fields that are centered in the points $\alpha_0^m r_{l\theta_0}[kb_0, 0]$, and have orientation $n\omega_0 + l\theta_0$. Note that the term $n\omega_0$ corresponds to the local orientation of the receptive fields around their center point. In the case where the receptive fields are isotropic, we can simply set $N = 1$. On the contrary, the term $l\theta_0$ corresponds to the global orientation, i.e., the angle between the line that connects the origin and the center of the receptive field and the x-axis. Since elements of S_p are uniquely determined by the parameters k, m, n and l, we use a more compact notation (k, m, n, l) to denote elements in S_p. We would like that the subset is invariant under some rotations and dilations. From the discretization of the SIM(2) group, it naturally arises that those rotations and dilations are derived from the following subset of SIM(2)

$$H_p = \left\{([0,0], \alpha_0^m, l\theta_0) \middle| m \in \mathbb{Z}, \alpha_0 > 1, \theta_0 = 2\pi/L, L \in \mathbb{N}^*, n = 0, \cdots, L-1\right\}. \tag{12}$$

Proposition 2.1. *For every choice of the parameters $(b_0, \alpha_0, \omega_0, \theta_0)$ with $b_0 > 0, \alpha_0 > 1, N, L \in \mathbb{N}^*$ of the log-polar grid, the set of all constructed receptive fields, denoted by S_p, is invariant under any transformations in the subset H_p. In addition, the action of each element of H_p induces a permutation of S_p.*

Proof: The proposition is a direct consequence from the way the elements of S_p and H_p are constructed. Indeed, for each $h = ([0,0], \alpha_0^{m'}, l'\theta_0) \in H_p$ and $x = (\alpha_0^m r_{l\theta_0}[kb_0, 0], \alpha_0^m, n\omega_0 + l\theta_0) \in S_p$, we have

$$h \circ x = (\alpha_0^{m'+m} r_{(l'+l)\theta_0}[kb_0, 0], \alpha_0^{m'+m}, n\omega_0 + (l'+l)\theta_0) \in S_p. \tag{13}$$

Note that H_p is a subgroup of SIM(2) under the same action, since the identity element $([0,0], 1, 0) \in H_p$ and for each $h = ([0,0], \alpha_0^{m'}, l'\theta_0) \in H_p$ we have $h^{-1} = ([0,0], \alpha_0^{-m'}, (-l' \mod L)\theta_0) \in H_p$. Therefore, it is easy to see that for each $x_1, x_2 \in S_p$, if $hx_1 = hx_2$, then $h^{-1}hx_1 = h^{-1}hx_2$ and thus $x_1 = x_2$. Therefore, the mapping of S_p to itself by h is one-to-one and $h \in H$ induces a permutation of S_p.

Before we present the main result for this section we need the following definitions.

Definition 2.2. *A set $\boldsymbol{\tau} := \left\{(t_k^j), k \in \mathbb{Z}, j = 1, 2, \ldots, N\right\}$ of spike trains produced by the TEM of Figure 2 is said to represent the video stimulus I if the TDM of Figure 2 with given input the set of spike trains $(t_k^j), k \in \mathbb{Z}, j = 1, 2, \ldots, N$ recovers the video stream.*

Definition 2.3. *Let I be an arbitrary input video stream and $\boldsymbol{\tau}$ the set of spike trains produced by the TEM of Figure 2. An IPT $\mathcal{T}$ is said to be realizable in the time domain, if there is a connectivity setting σ of the switching matrix, such that the set of spike trains $\boldsymbol{\tau}$ represents the video stimulus $\mathcal{T}I$.*

Theorem 2.4. *For a Video TEM with receptive fields according to S_p the following IPTs are realizable in the time domain:*

- *The set of rotations $\mathcal{R}_\theta$ where $\theta = l\frac{2\pi}{L}, l = 0, 1, \ldots, L-1$.*
- *The set of dilations $\mathcal{D}_\alpha$ where $\alpha = \alpha_0^m, m \in \mathbb{Z}$.*
- *Any synthesis of the above operators.*

Proof: See Appendix C.

Definition 2.5. *Let I denote the video stream that represents a certain object and let $\boldsymbol{\tau}$ be the set of spike trains obtained from the Video TEM upon presentation of I. The set $\{\mathcal{T}I : \mathcal{T} \in H\}$ is called the object manifold under H for the object that is represented with I.*

Corollary 2.6. *The switching architecture of Figure 2 where the set of receptive fields is characterized by S_p generates the whole object manifold under H_p in real time.*

Proof: Follows directly from the above proof of Theorem 2.4.

Remark 2.7. *The theory presented above considers video streams with spatial support on $\mathbb{R}^2$. Moreover, an implicit requirement of Theorem 2.4 is that the number of elements in the set of receptive fields (which equals the number of spiking neural circuits) is infinite. In applications however, as well as in the visual system, the number of neurons is finite. This means that the number of possible scalings as well as the number of possible translations is finite. The finite number of possible translations implies that the spatial domain over which the input video stream is defined is of finite measure, that is clearly the case. The finite number of scalings implies that the input stimuli have also finite spatial bandwidth which also holds. These facts also restrict the set of possible IPTs that can be implemented in the time domain, to those that are supported by the characteristics of the set of receptive fields. Note however that the set of possible rotations remains unaffected.*

2.2.4. Exact IPTs using a Cartesian Grid

From the presentation of the log-polar grid in section 2.2.3, we see that rotation and dilation transformations can be realized in an exact fashion by the same switching circuit. This is because these transformations commute with each other, and thus it is possible to derive a discrete set that preserves the group structure. On the contrary, the translation transformation does not commute neither with rotations nor with dilations. As a result, it is impossible to derive a general discrete set that will be closed under *all* the combined transformations of translation and rotation or

dilation. However, we present here a discrete set that can realize a subset of these transformations on the Cartesian grid, in an exact fashion.

In this case the centers of the receptive fields for each scale are placed on a Cartesian grid. Spatial filters of multiple scales are constructed, with the scales placed logarithmically. The resulting discretization of the SIM(2) group is given by (see also Figure S2 in the supplementary material)

$$S_c = \left\{ (\alpha_0^{-m}[kb_x, lb_y], \alpha_0^{-m}, n\omega_0) \middle| \begin{array}{l} b_x, b_y > 0, k, l \in \mathbb{Z} \\ \alpha_0 \in \mathbb{N}^*, \alpha_0 > 1, m \in \mathbb{N} \\ \omega_0 = 2\pi/N, N \in \mathbb{N}^*, n = 0, \cdots, N-1 \end{array} \right\}. \tag{14}$$

Again from (4), the general receptive field obtained from the above discretization is given by

$$\eta_{(\alpha_0^{-m}[kb_x, lb_y], \alpha_0^{-m}, n\omega_0)} = \alpha_0^m \eta(\alpha_0^m r_{-n\omega_0}([x, y] - \alpha_0^{-m}[kb_x, lb_y])). \tag{15}$$

For each scale α_0^{-m} the set includes the receptive fields centered at $\alpha_0^{-m}[kb_x, lb_y]$, with orientation $n\omega_0$. Note that this discretization resembles closely the one of the discrete wavelet transform. We would like S_c to be invariant under some specific translations and dilations. Based on the above discretization, this set of transformations is given by

$$H_c = \left\{ (\alpha_0^{-m'}[k'b_x, l'b_y], \alpha_0^{-m'}, 0) | m' \in \mathbb{N}, k', l' \in \mathbb{Z} \right\}. \tag{16}$$

As in the case of the log-polar grid, consider an element $h = (\alpha_0^{-m'}[k'b_x, l'b_y], \alpha_0^{-m'}, 0) \in H_c$ and an element $x = (\alpha_0^{-m}[kb_x, lb_y], \alpha_0^{-m}, n\omega_0) \in S_c$. Then we have

$$\begin{aligned} h \circ x &= (\alpha_0^{-m'}[k'b_x, l'b_y] + \alpha_0^{-m'-m}[kb_x, lb_y], \alpha_0^{-(m+m')}, n\omega_0) \\ &= (\alpha_0^{-(m+m')}[(\alpha_0^m k' + k)b_x, (\alpha_0^m l' + l)b_y], \alpha_0^{-(m+m')}, n\omega_0), \end{aligned} \tag{17}$$

and since α_0 is a positive integer, we have that $\alpha_0^m k' + k, \alpha_0^m l' + l \in \mathbb{Z}$ and therefore $h \circ x \in S_p$.

Based on the above discussion, we have the following theorem for the IPTs that are implementable with a Cartesian grid.

Theorem 2.8. *For a Video TEM with receptive fields according to S_c the following IPTs are realizable in the time domain:*

- *The set of dilations $\mathcal{D}_\alpha$ where $\alpha = \alpha_0^{-m}, m \in \mathbb{N}$.*
- *The set of translations $\tau_{[x_0, y_0]}$ where $[x_0, y_0] = [kb_x, lb_y], k, l \in \mathbb{Z}$.*

- *Any synthesis of the above operators, i.e., simultaneous dilation $\mathcal{D}_{\alpha_0^{-m}}$ and translation*
  $\tau_{\alpha_0^{-m}[kb_x, lb_y]}, m \in \mathbb{N}, k, l \in \mathbb{Z}$.

Proof: Similar to the one of Thm. 2.4. □

Remark 2.9. *Note that the only dilations that the Cartesian grid can implement are with a scale $\alpha_0^{m'}$, where m' is a strictly positive integer, and not a general integer as in the log-polar grid case. This corresponds only to the zoom-out transformation. The opposite zoom-in transformation cannot be performed in an exact way since for $m' < 0$, we also have $m + m' < 0$ for $m = 0, \ldots, -m' - 1$. However, note that even when this condition is not satisfied, we can perform approximate zoom-in transformations by keeping only the scales $-m', -m'+1, \ldots$ and disregarding the remaining ones. We present such an example in section 3.2.*

2.2.5. Approximate IPTs Using Nearest Neighbor Mapping

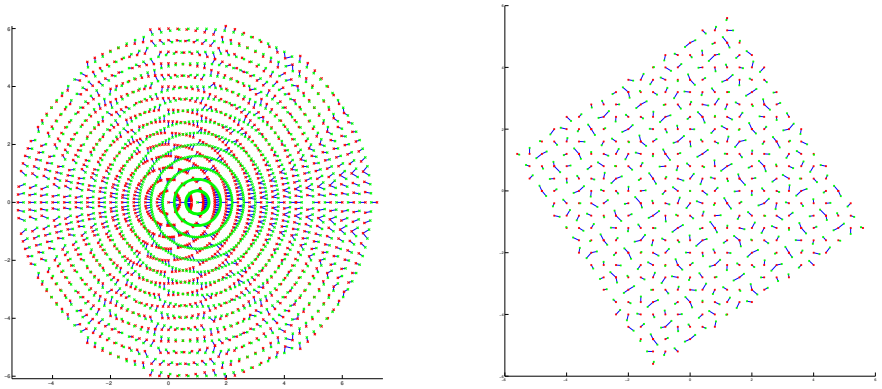
As we discussed, the two different grids, presented in sections 2.2.3 and 2.2.4 correspond to two different discretizations of the continuous SIM(2) group that consists of all possible rotations, translations and dilations. Conceptually, the SIM(2) group can generate all the possible IPTs and therefore has led in the past to the development of group theoretic approaches for the problem of visual perception (Hoffman, 1966; Dodwell, 1983). However, these ideas cannot lead to practical applications since they require an uncountable number of elements and the general group structure is not retained for arbitrary discretization schemes. Nevertheless, both of the grids can approximate the continuous one as they become denser, i.e., the distance between neighboring elements becomes smaller. The polar grid becomes dense in the continuous group as $N \to \infty$, $\alpha_0 \to 1^+$ and $b_0 \to 0^+$. Similarly, the Cartesian grid becomes dense in the continuous group as $b_x, b_y \to 0$, $\alpha_0 \to 1^+$ and $N \to \infty$. Therefore as the grids become denser, they also become more similar to the continuous one.

Using this argument we can also use a Video TEM with a set of receptive fields placed on a Cartesian grid to implement approximate rotations, of an arbitrary angle θ. To do so we set the switching matrix as follows: Upon the application of a rotation operator r_θ the spikes of the neural circuit with receptive field that corresponds to the point $(\alpha_0^{-m}[kb_x, lb_y], \alpha_0^{-m}, n\omega_0)$ is mapped, to its nearest neighbor with the same scale $(\alpha_0^{-m}[k'b_x, l'b_y], \alpha_0^{-m}, n'\omega_0)$ where k', l', n' are given by

$$
\begin{aligned}
[k', l'] &= \underset{i,j \in \mathbb{Z}}{\arg\min} \left\{ (kb_x \cos\theta + lb_y \sin\theta - ib_x)^2 + (-kb_x \sin\theta + lb_y \cos\theta - jb_y)^2 \right\} \\
&= [\text{nint}(k \cos\theta + lb_y \sin\theta / b_x), \text{nint}(-kb_x \sin\theta / b_y + l \cos\theta)] \,, \\
n' &= \underset{l \in \mathbb{Z}/N}{\arg\min} \, |n\omega_0 + \theta - l\omega_0|
\end{aligned}
\tag{18}
$$

where $\text{nint}(x)$ is the nearest integer to x. Note that the distance to the closest receptive field is always bounded by $\sqrt{b_x^2 + b_y^2}/2$ which ensures that as b_x, b_y becomes smaller, i.e., the grid becomes denser, the mapping becomes more accurate.

Using similar arguments, it is clear that using the log-polar grid we can also implement approximate translations. These will become more accurate as the grid becomes denser, i.e., b_0 and θ_0 become smaller.



(A) (B)

Figure 3: **(A)** Mapping of the translated receptive field of $\alpha = 0.5$ to their approximations in the existing polar grid. The green markers indicate the translated receptive fields, the red markers indicate the approximations, and the blue line shows the mapping from green to red. **(B)** Mapping of the rotated receptive field in dilation $\alpha = 0.5$ to their approximations in the existing Cartesian grid. The green markers indicate the translated receptive fields, the red markers indicate the approximations, and the blue line shows the mapping from green to red.

In Figure 3A, we show the nearest neighbor mapping of shifted receptive fields in one of the scales in the log-polar grid S_p. The green markers indicate the centers of

the translated receptive fields. The latter are connected by the blue lines, to the red markers that represent their nearest neighbor. As a comparison, we also show the nearest neighbor mapping of rotated receptive fields in one scale in the Cartesian grid S_c in Figure 3B.

3. Results

The theoretical tools developed in section 2 have been evaluated in a number of directions. This pertains to separable and nonseparable visual streams, and to the architecture of the video time encoders employed, including the choice of receptive fields and spiking neuron models. It also pertains to exploring the sampling of the IPT transformations, including dilations, rotations and translations. Finally, it pertains to the recovery of the encoded visual streams, including exact and approximate stimulus decoding algorithms.

We consider a space-time separable video stream of the form $I(x, y, t) = A(x, y)u(t)$, where $A(x, y)$ is an 256×256 pixel image defined on the spatial domain of $\mathbb{D} = [-8, 8] \times [-8, 8]$, and $u(t)$ is an one second long temporal signal of bandwidth 4 Hz. In the following examples, we show snapshots of the video stream I and the various transformed videos at the time instant $t|_{u(t)=1}$ for illustration purposes. The spatial component of the input stimulus is visualized in Figure 4A. The Structural Similarity (SSIM) index (Wang et al., 2004) of the shown snapshot is evaluated for each of the examples. Bilinear interpolation is used to perform transformations on the original signals to create references that the reconstructions are compared to.

In our evaluations we will also use a space-time nonseparable natural visual stream with the same characteristics as the separable video stream described above, that incidentally, exhibits a higher temporal bandwidth. The SSIM index is evaluated for the entire visual stimulus.

3.1. Rotations and Dilations on the Log-Polar Grid

We start by constructing the receptive fields on the log-polar grid and by providing the parameters of the IAF neurons defining the Video TEM. Subsequently, we describe how to achieve with the TDP block rotation and dilation transformations.

Center-surround receptive fields of RGCs and neurons in the LGN have been modeled with the Difference of Gaussians (DoG) wavelet (Kuffler, 1953; Rodieck, 1965). Here the mother wavelet of the receptive fields is the DoG

$$\eta(x, y) = \frac{1}{2\alpha_1^2} \exp\left(-\frac{x^2 + y^2}{2\alpha_1^2}\right) - \frac{1}{2\alpha_2^2} \exp\left(-\frac{x^2 + y^2}{2\alpha_2^2}\right), \qquad (19)$$

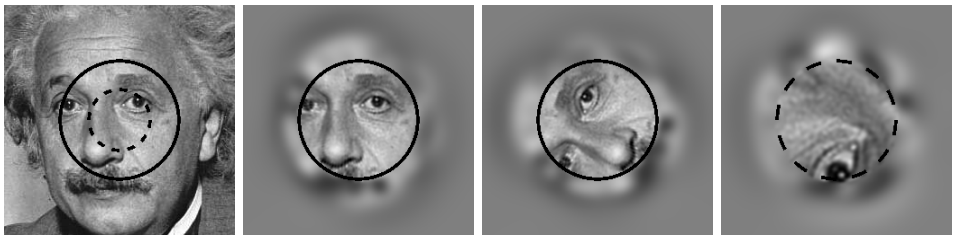


Figure 4: **Rotations and dilations (zoom-in) on the log-polar grid. (A)** Spatial component of the original stimulus. The reconstruction is performed only for the region inside the solid black circle. The dashed circle indicates the region that was zoomed into. **(B)** Recovery with a null connectivity setting switching matrix. **(C)** Recovery with a switching matrix connectivity setting realizing a rotation of 69 degrees counter-clockwise. **(D)** Recovery with a switching matrix connectivity setting realizing a rotation of 135 degree clockwise and a dilation factor of 2. The region corresponds to the one inside the dashed circle in the original stimulus in (A). The recovery was multiplied by 2 before being shown. The SSIM index for (B-D) is, respectively, $0.90, 0.93, 0.94$.

with parameters $\alpha_1 = 0.5, \alpha_2 = 1.6\alpha_1$.

The parameters for generating the filter bank are defined on the log-polar grid (see section 2.2.3):

- $\alpha_0 = 2, m \in \{-3, -2, -1, 0, 1\}$,
- $\theta_0 = 2\pi/L, L = 120, l \in \{0, 1, \cdots, 119\}, \omega_0 = 2\pi, N = 1$,
- $b_0 = 0.8$.

Each receptive field output was fed into an IAF neuron, with bias $b = 0.4$, threshold $\delta = 0.03$ and integration constant $\kappa = 1$. The total number of receptive fields, as well as the total number of neurons was $18,605$; a total of $245,690$ spikes were fired. We focussed on the reconstruction of the circular region $x^2 + y^2 \leq 4^2$ as indicated by the solid circle in Figure 4A. All the receptive fields centered in the domain were taken into account, together with a small number of receptive fields whose centers are immediate neighbors of the circular domain.

We present several examples showing the recovery of stimuli encoded with the Video TEM described above. Two experiments are described. In the first, the encoded stimulus is space-time separable. In the second, a natural video stream is

used that is non-separable.

We first demonstrate the ability of the TDP block with various connectivity settings to recover, rotate and dilate a separable visual stream. Perfect recovery was realized by the TDP block with the null setting of the switching matrix performing an identity transformation. Figure 4B shows the recovery of the visual stimulus within the framed region. Since $\theta_0 = 2\pi/120$ in the log-polar grid, any rotations of multiples of 3 degrees can be realized. In this example, a rotation of 69 degrees in the counter-clockwise direction is performed. The spike train coming from the neuron whose spatial receptive field is the element $(\alpha_0^m r_{l\theta_0}[kb_0, 0], \alpha_0^m, l\theta_0)$ of the filter bank is mapped into the reconstruction filter with parameter $(\alpha_0^m r_{(l+23)\theta_0}[kb_0, 0], \alpha_0^m, (l+23)\theta_0)$. The resulting reconstruction is shown in Figure 4C.

We now present an example of a simultaneous rotation of 135 degrees clockwise and a dilation by a factor of 2. This transformation can be achieved by routing the spikes fired by the neuron whose spatial receptive field is the element $(\alpha_0^m r_{l\theta_0}[kb_0, 0], \alpha_0^m, l\theta_0)$ to the reconstruction filter $(\alpha_0^{m+1} r_{(l-45)\theta_0}[kb_0, 0], \alpha_0^{m+1}, (l-45)\theta_0)$, $m = -3, -2, -1, 0$. The spikes of the encoding neurons at scale $m = 1$ were ignored since the scale $m = 2$, at which they were to be shifted, did not exist. The result is shown in Figure 4D. Note that the reconstruction in the figure had been multiplied by 2 for illustration purposes since the dilated version is scaled by $\frac{1}{2}$ due to the unitary condition.

Second, we demonstrate rotation and dilation transformations executed by the TDM on a space-time non-separable natural visual stream, the size of which is the same as the separable video stream in the previous examples. Four snapshots of the original visual stimulus are shown in Figure 5A. We used the same set of receptive fields and neurons, except that the bias and threshold of each of the neurons were respectively set to $b = 1.2$ and $\delta = 0.04$. These parameters guarantee the high quality recovery of the encoded video stimuli. The ensemble of neurons fired $550,722$ spikes in the 1 second duration of the visual stream. We performed a rotation of 63 degrees counter-clockwise and zooming out by a factor of 2. These transformations are achieved by routing the spikes fired by the neuron whose spatial receptive field is the element $(\alpha_0^m r_{l\theta_0}[kb_0, 0], \alpha_0^m, l\theta_0)$ to the reconstruction filter $(\alpha_0^{m-1} r_{(l+21)\theta_0}[kb_0, 0], \alpha_0^{m-1}, (l+21)\theta_0)$, $m = -2, -1, 0, 1$. The spikes of the encoding neurons at scale $m = -3$ were ignored since the scale $m = -4$, at which they were to be shifted, did not exist. The result is shown in Figure 5B. Four snapshots of the transformed visual stimulus are shown, each corresponding to theones in Figure 5A. Again, the zoomed out reconstruction was multiplied by 0.5 before being shown. The entire video of this transformation, together with the videos of other rotation and dilation transformations can be found in the supple-

mentary material (video1, video2).

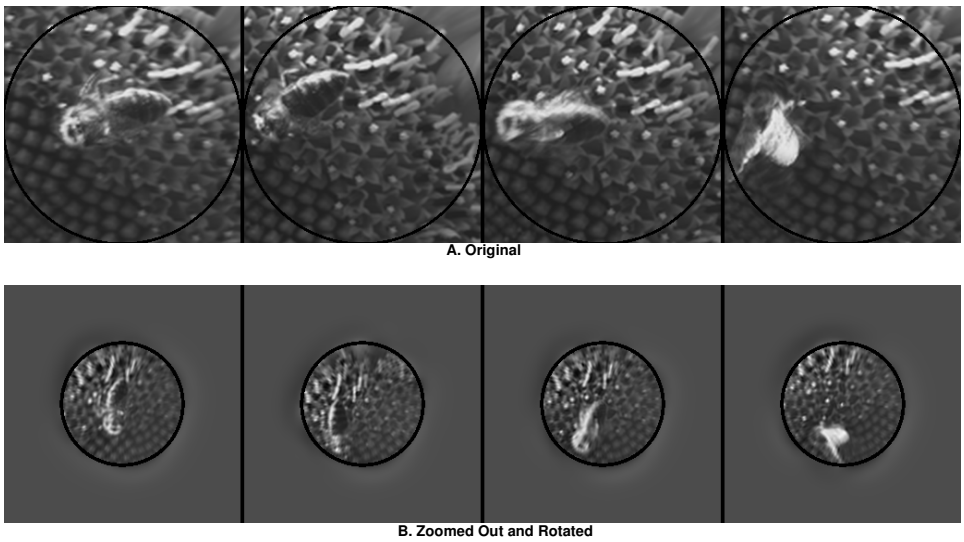


Figure 5: **Rotations and dilations (zoom-out) of non-separable natural scenes on the log-polar grid. (A)** 4 snapshots of the original visual stream. **(B)** The recovered video is dilated by a factor of 0.5 and rotated counter-clockwise by 63 degrees. The recovery is performed only for the region inside the black circle as indicated in the bottom panel. The recovery is multiplied by 0.5 before being shown.

From the results obtained in this section, it is clear that IPTs such as rotations and dilations can be performed in the spike domain with a simple switching mechanism as we have described in Section 2.2.3.

3.2. Translations and Dilations on the Cartesian Grid

We start by presenting the architecture of the Video TEM. This is followed by the TDP block connectivity settings to achieve translation and dilation transformations.

The mother wavelet that was chosen in this case was a Gabor wavelet of the form

$$\eta(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} - \frac{y^2}{8}\right)\left(e^{i\kappa x} - e^{-\kappa^2/2}\right), \tag{20}$$

where real and imaginary parts are separated into two real receptive fields. The Gabor wavelet models well the linear transformations observed in the receptive fields of simple cells in V1(Jones & Palmer, 1987).

The parameters for generating the filterbank are located on the Cartesian grid (see also section 2.2.4):

- $\alpha_0 = 2, m \in \{0, 1, 2, 3\}$
- $b_0 = 2$,
- $\omega_0 = 2\pi/N, N = 8, n \in \{0, 1, 2, 3, 4, 5, 6, 7\}$,

leading to a total of 99,136 filters. Note that we also added 8 orientations for the Gabor filter, which as discussed, does not affect the translation and dilation transformations. Since the simulation can only be performed for a finite grid instead of countably infinite one, we again only consider the reconstruction inside the $[-4, 4] \times [-4, 4]$ region in the middle of the image, as highlighted by the rectangle in Figure 6A. Receptive fields centered in this region and in its immediate neighborhood were taken into account. All the IAF neurons had parameters $\kappa = 1, \delta = 0.02$ and $b = 0.3$.

Perfect recovery can be implemented by the null setting switching matrix performing an identity transformation in the TDP building block. Figure 6B shows the reconstruction of the visual stimulus within the framed region.

Translation by 2 units both to the right and upwards corresponds to the connectivity setting of the switching matrix that maps the receptive field representing the element $(\alpha_0^{-m}[kb_x, lb_y], \alpha_0^{-m}, n\omega_0)$ to $(\alpha_0^{-m}[(k+\alpha_0^m \cdot 1)b_x, (l+\alpha_0^m \cdot 1)b_y], \alpha_0^{-m}, n\omega_0)$. The translation result is shown in Figure 6C, and the shift is indicated by the arrow.

We now demonstrate simultaneous dilation by a factor of $\frac{1}{2}$ and translation upwards by 1 unit. This was be achieved by the switching matrix setting that wires receptive fields representing the elements $(\alpha_0^{-m}[kb_x, lb_y], \alpha_0^{-m}, n\omega_0)$ to the elements $(\alpha^{-m-1}[kb_x, (\alpha_0^m + l)b_y], \alpha^{-m-1}, n\omega_0)$, for $m = 0, 1, 2$. The spikes of the neurons belonging to the dilation level $m = 3$ were ignored. The result of the simultaneous translation and dilation transformations is shown in Figure 6D.

Several videos demonstrating the reconstruction of non-separable natural scenes can be found in the supplementary material (video3, video4). Both recovery and translations as well as, dilations and translations are, respectively, shown on the Cartesian grid.

Concluding, if the spatial receptive fields are defined on a Cartesian grid as described in Section 2.2.4, then admissible translation and dilation IPTs can be achieved

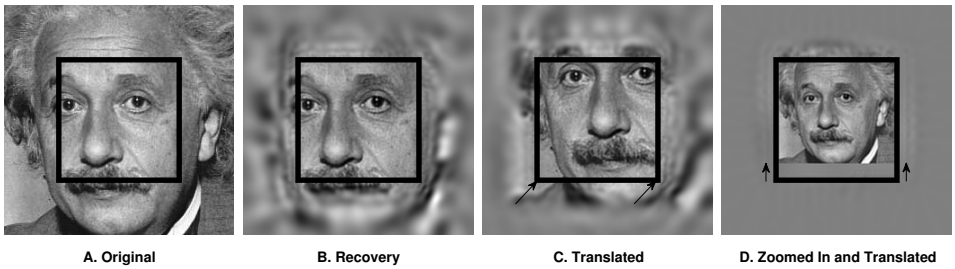


Figure 6: **Translations and dilations on the Cartesian grid. (A)** Original stimulus. The recovery is performed only for the region inside the black frame. **(B)** Recovery with a switching matrix with null connectivity setting. **(C)** Recovery with a switching matrix connectivity setting realizing a translation by 2 units to the right and 2 units upwards, as indicated by the arrow. **(D)** Recovery with a switching matrix connectivity setting realizing a dilation of factor of 0.5 and a translation by 1 unit upwards, as indicated by the arrow. The recovery was multiplied by 0.5 before being shown. The SSIM index for (B-D) is, respectively, $0.95, 0.95, 0.92$.

using a simple switching mechanism.

3.3. Approximate Transformations

In the previous sections we have shown that the "natural" grid for constructing the Video TEMs for rotations and dilations is the log-polar grid. We have also demonstrated that the natural grid for translations and dilations is the Cartesian grid.

In this section we explore the performance in recovery when the receptive fields of the Video TEMs are built on the log-polar grid and translation and dilation transformations are performed in the spike domain. Conversely, for receptive fields of the Video TEMs defined on the Cartesian grid, rotations and translations are considered.

Throughout, we constructed the receptive fields of the Video TEM with the DoG mother wavelet given in (19). The parameters of the IAF neurons are as described in section 3.1.

In order to explore approximate rotation transformations, the receptive fields were placed on a Cartesian grid with the following parameters:

- $\alpha_0 = 2, m \in \{0, 1, 2, 3\}$,
- $\omega_0 = 2\pi, N = 1$,
- $b_0 = 0.8$.

We performed several clockwise rotations with successive multiples of 23°. The results are shown in Figure 7. Only the recovery shown in Figure 7A is exact. The other rotations in Figure 7B-H were approximate. Nevertheless, the original image can easily be identified with the desired rotations. Transformations of space-time non-separable natural scenes have also been investigated. The reconstruction for approximate rotation transformations can be found in the supplementary materials (video5).

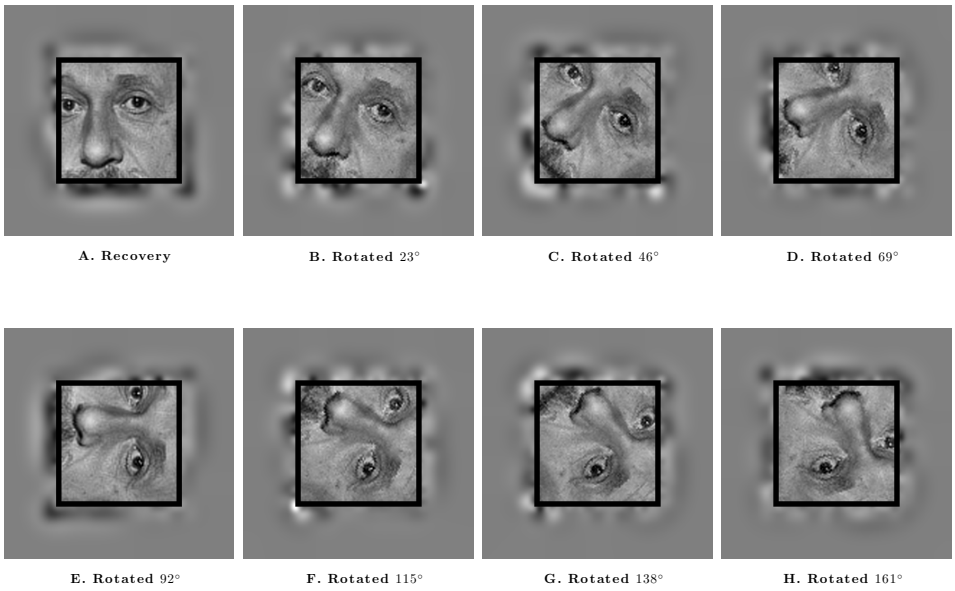


Figure 7: **Approximate arbitrary rotations on the Cartesian grid.** Recovery of a successive multiple of 23° clockwise rotations. Only the focused area of the video is being shown. **(A)** Recovery. **(B)** Rotation by 23°. **(C)** Rotation by 46°. **(D)** Rotation by 69°. **(E)** Rotation by 92°. **(F)** Rotation by 115°. **(G)** Rotation by 138°. **(H)** Rotation by 161°. The SSIM index for (A-H) is, respectively, $0.94, 0.87, 0.88, 0.87, 0.88, 0.87, 0.88, 0.86$.

To illustrate the effect of the density of the Cartesian grid on arbitrary rotations, we performed three experiments with increasing grid density. By setting $b_0 =$

$1.0, 0.8, 0.6$ in each experiment, respectively, we rotated the video by $144°$ clockwise. The results are shown in Figure 8B-D. The rotations were performed correctly for all three b_0's. Moreover, as we decreased b_0 from 1.0 in Figure 8D, to 0.8 in Figure 8C and finally to 0.6 in Figure 8B, the quality of the approximate transformation increasingly improved. We then examined the quality of stimulus recovery using the grids mentioned above. The grid density has an effect on the quality of reconstruction. Since the identity transformation can be exactly implemented, the quality degradation of the approximate rotation of stimuli parametrized by different grid densities can be further evaluated. The evaluation is based on comparing for each stimulus the approximate rotation with the identity transform on the same grid. For $b_0 = 0.6$, the quality of recovery of the approximate rotation only differs from the quality of the identity transform by 0.01 SSIM. For $b_0 = 0.8$, the quality of recovery of the rotation is lower than that of the identity transformation by 0.07 SSIM, while for $b_0 = 1.0$, the difference is 0.13 SSIM. Therefore, as the grid becomes denser, the approximate transformation converges to a "faithful" one.

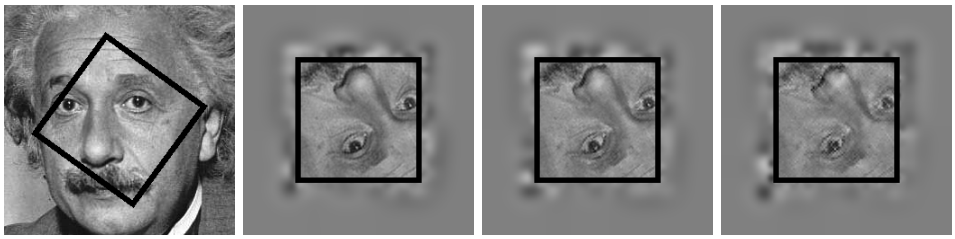


Figure 8: **Effect of grid density on rotations with nearest-neighbor mapping on the Cartesian grid. (A)** Original stimulus. **(B-D)** Recoveries of rotation by $144°$ with grid parameter (B) $b_0 = 0.6$, (C) $b_0 = 0.8$, (D) $b_0 = 1.0$. As the grid becomes denser (smaller b_0) the quality of the approximate rotation improves. The SSIM index for (B-D) is, respectively, 0.94, 0.88, 0.76. For comparison, the SSIM indexes of the recovery are $0.95, 0.94, 0.89$ for the grids in (B-D), respectively.

Remark 3.1. *In practice approximate rotations of arbitrary degree can be performed using the nearest neighbor mapping described in section 2.2.5. Note, however, that rotations of very similar degree, can potentially lead to the same mapping and therefore become indiscriminable, a phenomenon that has been reported in the psychophysics literature (Westheimer & Beard, 1998). The rotation angle resolution depends on the resolution of the filterbank, i.e., the parameters b_x, b_y and α_0.*

In order to evaluate approximate translations on the log-polar grid we used the same encoding architecture as in the examples in section 3.1. To perform a transla-

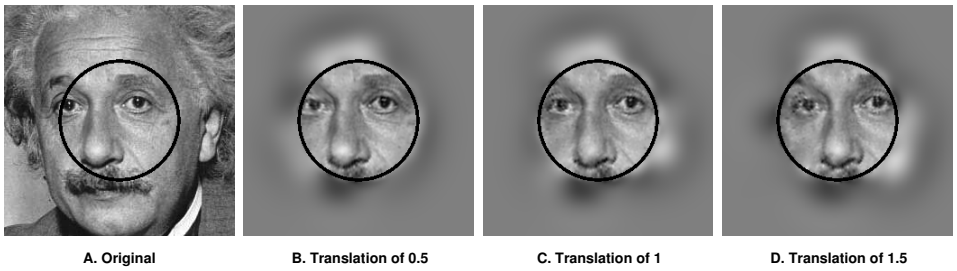


Figure 9: **Approximate translations on a log-polar grid. (A)** Original stimulus. **(B-D)** Recovery of translations by 0.5 units in (B), by 1 unit in (C) , and by 1.5 units in (D). Smaller translations exhibit improved quality of recovery because of the non-uniform tiling of the log-polar grid. The SSIM index for (B-D) is, respectively, $0.81, 0.80, 0.79$.

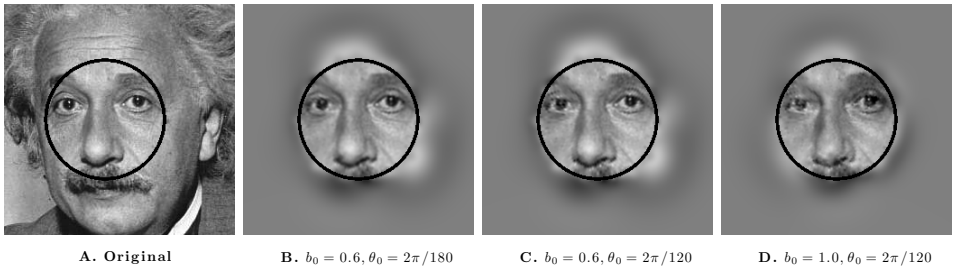


Figure 10: **Effect of polar grid density on translations. (A)** Original stimulus. **(B-D)** Recovery of translations by 1 unit, the grid densities are: (B) $(b_0 = 0.6, \theta_0 = 2\pi/180)$, (C) $(b_0 = 0.6, \theta_0 = 2\pi/120)$, (D) $(b_0 = 1.0, \theta_0 = 2\pi/120)$. As the grid becomes denser (smaller θ_0) the quality of the approximate translation improves. The SSIM index for (B-D) is, respectively, $0.85, 0.83, 0.75$. For comparison, the SSIM indexes are $0.90, 0.90, 0.89$ for the grids in (B-D), respectively. Compare also with Figure 9(C) where the grid density is $(b_0 = 0.8, \theta_0 = 2\pi/120)$.

tion, the receptive fields are mapped to the nearest existing receptive field after the translation. In Figure 9B-D we show, respectively, results obtained by translation to the right by $0.5, 1$ and 1.5 units. The same translations on the non-separable natural scenes was also performed. The complete video for these three translations can be found in supplementary material (video6).

We now investigate the effect of the density of the grid on the reconstruction. There are two ways to increase the density of the grid. One is to use a smaller b_0, the other is to decrease θ_0; we show their effects in Figure 10. Again, we can observe that the denser the grid is, the better the reconstruction quality of image (or natural scenes) translations. For $b_0 = 0.6, \theta_0 = 2\pi/180$, the recovery is less noisy. As before, since the reconstruction quality may also depend on the grid density itself, we tested the identity transformation on the same grids. The difference between the SSIM index of the identity transform and the approximate translations decreases as the grid becomes denser. However, it should be noted that the recovery quality of the approximate translation depends not only on the grid density, but also on the value of the translation performed, since the grid is denser in the center and coarser away from the center. These observations are consistent with several experimental studies (Kravitz et al., 2008).

It can be seen that, translations on the polar grid are not performing nearly as well for coarser grids, especially when compared with the reconstruction quality of rotations on a Cartesian grid. Such an effect is expected, as the polar grid is not uniformly tiling the space.

4. Discussion

In this paper we presented a general model for the realization of identity-preserving transformations in the spike domain. Our model architecture consists of a spike domain switching circuit (Time Domain Processing) that channels the spikes from the sensory neurons (Time Encoding Machine) into the higher brain areas (processed with the Time Decoding Machine). Surprisingly, a simple rewiring strategy can perform a class of identity-preserving transformations such as rotations, scalings and translations, thereby giving rise to a family of invariant transformations in the spike domain. We demonstrated that this class of transformations can easily be realized with a neural circuit architecture using the same basic stimulus decoding algorithm. What changes in the architecture are only the connectivity settings of the switching matrix (i.e., the input/output "wiring") of the TDP building block. Each connectivity setting corresponds to a particular IPT.

The t-transform formalism allows us to interpret neural spiking in the language of inner product operations. This formalism also enables us to consider an architecture that has provably high performance characteristics. We note that the implementation of IPTs with a switching mechanism depends only on the structure of the encoder (in our case filterbank models of receptive fields and IAF neurons for the spiking mechanisms) and not on the precise timing of spike responses. As a result, IPTs are also realizable using the same switching mechanism for any model with the same encoding structure.

When dealing with exact IPTs, a key property we have implicitly utilized is that the receptive field outputs are invariant to transformations performed on the grid. However, a common pitfall of wavelet representation of signals is that the representation is highly dependent on the relative alignment of the input signal with the grid. For example, rotation invariance is guaranteed only for rotation that are multiples of the grid size. Therefore, the invariance is discretized as well.

Shiftable and steerable filters have been introduced as an alternative approach to these problems (Freeman & Adelson, 1991; Simoncelli et al., 1992). Shiftable and steerable filters provide efficient implementations of arbitrarily translated or oriented filters from a linear combination of a bank of basis filters. These linear transformations can be viewed as coordinate transformations in the Hilbert space; a set of coordinates correspond to a collection of filters. The coordinate transformation is achieved via a linear operation (matrix multiplication). The matrix is typically obtained by solving systems of equations. The steerable and shiftable filters are carefully designed so that the matrix can be efficiently computed analytically or expressed in closed form.

Can these desirable characteristics be translated to our setting? A critical assumption for coordinate transformations, regardless of their efficiency, is that the sampling is linear. However, temporal sampling with neural circuits is highly nonlinear and signal dependent. While in our current setting linear interpolation is not readily possible, the efficiency gain of the TDM realization is even more noteworthy.

The switching matrix of the TDP block provides a simple, yet powerful realization of spike domain processing. Although the connectivity settings of the switching matrices we presented are fixed and each corresponds to a single transformation, the rewiring capability can dramatically increase the representational and processing power of the neural assemblies.

A question that naturally arises is how can such a rewiring mechanism be implemented with neural substrates and how the visual system decides which IPT to perform. A model for dynamic regulation of the connectivity between two neural

layers was proposed by Olshausen et al. (1993), where a set of control neurons was used to dynamically modify the synaptic strengths of the inter-layer connections. Testing this hypothesis, however, would require tracking the functional switching of a large population of neurons at a very fast timescale. Given the current and near term experimental capabilities, a testable prediction about the switching connectivity matrix performed on large populations, while desirable, is out of reach.

The IPTs implemented with the switching mechanism can play an important role in view-dependent invariant recognition. In the class of invariant recognition models where transformations of the incoming visual signal are matched with an existing stored version of the image, object representations are stored as originally viewed (Bülthoff & Edelman, 1992). Recognition is achieved by transforming the input to match the view specification of the stored representation. These transformations can be achieved by interpolation (Poggio & Edelman, 1990), mental transformation (Tarr & Pinker, 1989) or alignment (Ullman, 1989). In our setting, these transforms can be readily achieved by multiple parallel readouts corresponding to multiple transforms executed in parallel.

A way of determining the transformation between the input and the stored object has been proposed by Arathorn (2002). Called the map seeking circuit (MSC), this algorithm identifies a discretely parametrized linear transformation (based on rotations, translations, dilations, etc.) that minimizes an appropriate cost functional. To do the MSC algorithm iteratively, three classes of operations have to be specified: (i) a set of linear transformations, (ii) evaluation of the similarity between a stored version of an image and the transformed input image, and (iii) arithmetic computations such as addition, and appropriate updates of the coefficients of the linear transformations. Although not specifically focussing on the implementation of the MSC algorithm, our formalism provides a methodology for the parallel construction of the set of linear transformations as required by one of the key operations of the algorithm, in the context of our spiking architecture. Efficient spike domain algorithms that evaluate the similarity between transformations of the incoming stimulus and stored in memory patterns will be pursued in future work.

The set of all possible IPTs and the corresponding invariant representations that can be realized with our architecture originate from the spatial structure of the neuron receptive fields. The latter form an overcomplete spatial filterbank and also exhibit specific symmetry properties that enable the implementation of IPTs with a simple switching mechanism. By employing a similar group structure for the temporal receptive fields, one can devise a space-time overcomplete filterbank at the receptive field level, similar to the ones used in space-time wavelets (Antoine et al., 2004). The use of space-time non-separable receptive fields (Lazar & Pnev-

matikakis, 2011a,b) is also possible provided the group structure is kept. Such an organization of receptive fields can facilitate a richer set of IPTs that can be realized with the same switching matrix architecture, including space-time transformations. Consequently, new forms of invariant transformations can arise such as velocity invariance. The latter suggests investigating attention models for tracking.

Our focus on invertible transformations and their implementation (and the characterization of implementable IPTs) provides a solid foundation for the theoretical capabilities of our model. In addition, some classes of IPTs, such as the ones creating the mirroring effect, are also realizable under this architecture. The switching mechanism can also be used for non-invertible transformations. For example by combining the operations of scaling and translation one can zoom into a particular region of the incoming video stream, and thereby select a particular spatial region to be propagated to the next layer. This suggests a methodology for the implementation of attention-selective mechanisms. Such properties also address the correspondence problem of identifying equivalent stimuli while constantly changing visual fixations. A further example of non-invertible transformations arises when the encoder is noisy. In the latter case, the mathematical formalism for signal recovery is based on regularization (Lazar et al., 2010; Lazar & Zhou, 2012) and can be directly adapted to our setting. However, other IPTs that frequently arise in visual recognition tasks that facilitate invariance to, for example, occlusion, clutter and illumination will require additional mechanisms.

In our architecture IPTs are realized by processing spikes, the natural language of the brain. Spike processing in the TDP block is based on a key symmetry assumption on the receptive fields. It is also based on the assumption that the spiking mechanisms of the encoding neurons are identical. Consequently, IPTs can be implemented without modifying the decoding block. To overcome the latter limitation, employing further spike processing in addition to connectivity changes may be necessary, and will be investigated elsewhere.

Acknowledgments

The work presented 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 High Performance Computing Center under NSF Grants CNS-0855217 and CNS-0958379.

Appendix A. The Architecture of Video Time Encoding Machines and Video Time Decoding Machines

In this section we provide a brief overview of the machinery of Video TEMs and Video TDMs. Full treatments can be found in the cited references.

Appendix A.1. The Architecture of Video Time Encoding Machines

Time encoding is a formal method of mapping analog signals into a time sequence (Lazar & Tóth, 2004). Formal spiking neuron models are instantiations of TEMs and encode information in the time domain. Assuming that the input signal is bandlimited and the bandwidth is known, a perfect recovery of the stimulus based upon spike times can be achieved provided that the spike density is above the Nyquist rate (Lazar & Tóth, 2004). These results hold for a wide variety of sensory stimuli, including audio and video, encoded with a population of spiking neural circuits with various spiking mechanisms such as IAF (Lazar & Pnevmatikakis, 2008), or Threshold-and-Fire (TAF) (Lazar & Pnevmatikakis, 2011b). However, even when the bandlimited assumption is dropped and noise is present, information encoded in the time domain by spiking neurons can be recovered (Lazar & Pnevmatikakis, 2009; Lazar et al., 2010), thus enhancing the representational power of spiking neural circuits.

The general architecture of a Video TEM is shown in Figure 2. The video stimulus I is sensed by a population of N neural circuits. Each circuit consists of a spatiotemporal receptive field (STRF) $D^j, j = 1, \ldots, N$, in cascade with a neural circuit. This mechanism could be either a simple abstract spiking neuron model such as IAF, TAF or a biophysical neuron model (e.g., Hodgkin-Huxley) (Kim & Lazar, 2011). Populations of M pulse-coupled neurons (e.g., On-Off neuron pairs) have also been considered (Lazar & Pnevmatikakis, 2011b).

To analyze the operation of a Video TEM we need to embed the visual stimuli into an appropriate Hilbert space. Let $\mathcal{H}$ denote the space of (real) analog video streams $I = I(x, y, t), (x, y, t) \in \mathbb{R}^3$, that are bandlimited in time, continuous in space, and have finite energy. By bandlimited in time, we mean that for every $(x_0, y_0) \in \mathbb{R}^2$, $I(x_0, y_0, t) \in \Xi$, where Ξ is the space of finite energy bandlimited functions with cutoff frequency Ω. Formally

$$\mathcal{H} = \left\{ I = I(x, y, t) | I(x_0, y_0, t) \in \Xi, \, \forall (x_0, y_0) \in \mathbb{R}^2 \text{ and } I(x, y, t_0) \in L^2(\mathbb{R}^2), \; \forall t_0 \in \mathbb{R} \right\}.$$

It is clear that the space $\mathcal{H}$, endowed with the standard L^2 inner product is a well defined Hilbert space. We assume that the filters describing the spatiotemporal receptive fields are Bounded-Input Bounded-Output (BIBO) stable. In full generality

we assume that each neural circuit $j, j = 1, 2, \ldots, N$, has a spatiotemporal receptive field described by the function $D^j = D^j(x, y, t), (x, y, t) \in \mathbb{R}^3$. Filtering the video stream with the receptive field of the neural circuit j gives the receptive field output $v^j(t)$. The latter serves as the main input to neural circuit j and amounts to

$$v^j(t) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}^2} D^j(x, y, s) I(x, y, t - s)\, dxdy \right) ds. \tag{A.1}$$

Note that since the filters $D^j, j = 1, \ldots, N$, are BIBO stable, the outputs $v^j, j = 1, \ldots, N$, are bounded.

The resulting output of the receptive field v^j is then mapped by the spiking neural circuit into a multidimensional sequence of spike trains $t_k^{ji}, i = 1, \ldots, M, k \in \mathbb{Z}$. Here M denotes the number of spiking components of the neural circuit. For example, in the case of an On-Off neural pair $M = 2$. Without loss of generality we assume $M = 1$ throughout the paper. This mapping can be characterized by the t-transform of the neuron which relies on the observation that for many formal spiking neuron models, spiking is equivalent to taking a generalized measurement on the input stimulus. These measurements can be expressed in the form of the inner product operations

$$\left\{ \langle v^j, \chi_k^j \rangle = q_k^j, k \in \mathbb{Z}, j = 1, \ldots, N \right\}, \tag{A.2}$$

between the receptive field output v^j and some temporal functions χ_k^j that depend on the spike times (that are in turn stimulus dependent) and the parameters of the neuron. The resulting output of this operation can be written as $q_k^j, k \in \mathbb{Z}, j = 1, \ldots, N$. For example, in the case of a TAF neuron, the functions χ_k^j are the pointwise evaluation functions of the dendritic output v^j at the spike times t_k^j, whereas for the IAF case they evaluate the integral of v^j between two consecutive spike times. Biophysical neuron models also have such a compact inner product form description (Lazar (2010)). Neural circuits with arbitrary connectivity and feedback can also be described in the same manner (Lazar & Pnevmatikakis (2011b,a)). Note that although in inner product form, the representation (sampling) of the stimulus given in equation (A.2) is non-linear and the spike times $(t_k^j), k \in \mathbb{Z}, j = 1, \ldots, N$, are its solution.

Equivalently, the inner product in (A.2) can be written as the inner product between the input stimulus I and a *sampling* function ϕ_k^j. Thus the t-transform can be written as the following set of equations

$$\left\{ \langle I, \phi_k^j \rangle = q_k^j, k \in \mathbb{Z}, j = 1, \ldots, N \right\}. \tag{A.3}$$

Both the sampling functions and the outcome of these projections is determined by the spike times and the parameters of the neural circuits. Derivation of these results and more information can be found in (Lazar & Pnevmatikakis (2011b); Lazar et al. (2010); Lazar & Zhou (2012)).

Appendix A.2. The Architecture of the Video Time Decoding Machines

In this section we assume that the Time Domain Processing block of Figure 2 faithfully transmits the incoming spike trains to the TDM block for the recovery of the encoded video stream. This will help understanding the operational power of the TDP block that will be presented in section 2.1 and in more detail in section 2.2.

A question that arises naturally is under what conditions these projections capture all the information about the input stimulus I. The key condition of perfect recovery is that the set of sampling functions $\phi = (\phi_k^j), j = 1, 2, \ldots, N, k \in \mathbb{Z}$, forms a frame for the space of interest $\mathcal{H}$ (Christensen, 2003). Intuitively, the frame property is met when two conditions are satisfied: First, the set of receptive fields spans the whole visual space and, thereby, no information is lost during the filtering of I. Second, the spike density of the neural circuits has to be above a certain threshold so that the temporal encoding of the dendritic tree outputs v^j retains all the information about I. Necessary conditions for the receptive field property were first given in (Lazar & Pnevmatikakis (2011b)). A general tight condition that involves the spiking densities as well was derived in (Lazar & Pnevmatikakis (2011a)).

Provided that the frame condition is satisfied, there are a number of algorithms that can be used to recover the stimulus. In (Lazar et al., 2010) the recovered signal was represented by an orthogonal basis of $\mathcal{H}$. The coefficients of this orthogonal basis were determined by solving a (in general overcomplete) system of linear equations, such that the recovered stimulus satisfies the measurements provided by the t-transform. This algorithm which relies on a matrix inversion can also be realized with neural components (Lazar & Zhou, 2012). Here we use the algorithm of (Lazar & Pnevmatikakis (2011b)) which enables the fast realization (decoding) of the identity preserving transformations. The algorithm is summarized below and is schematically depicted in Figure 2.

Algorithm Appendix A.1. *If the frame condition holds, then the video stream I, encoded with a Video TEM (Figure 2), can be recovered as*

$$I(x, y, t) = \sum_{j=1}^{N} \sum_{k \in \mathbb{Z}} c_k^j \psi_k^j(x, y, t), \tag{A.4}$$

where $\psi_k^j(x,y,t)$ is a set of suitable recovery functions that span $\mathcal{H}$ and $c_k^j,\ j = 1,2,\ldots,N, k \in \mathbb{Z}$, are suitable coefficients. Let $[\mathbf{c}^j]_k = c_k^j$ and $\mathbf{c} = [\mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^N]^T$. The coefficients $\mathbf{c}$ can be computed as

$$\mathbf{c} = \mathbf{G}^+\mathbf{q}, \tag{A.5}$$

where T denotes the transpose, $\mathbf{q} = [\mathbf{q}^1, \mathbf{q}^2, \ldots, \mathbf{q}^N]^T$, $[\mathbf{q}^j]_k = q_k^j$ and $\mathbf{G}^+$ denotes the pseudoinverse of $\mathbf{G}$. The entries of the block-matrix $\mathbf{G}$ are given by $\mathbf{G} = [\mathbf{G}^{ij}]$ where $[\mathbf{G}^{ij}]_{kl} = \langle \phi_k^i, \psi_l^j \rangle, k,l \in \mathbb{Z},\ i,j = 1,2,\ldots,N$.

Intuitively, the recovered stimulus is expanded upon by the set of recovery functions $\psi_k^j, k \in \mathbb{Z}, j = 1,\ldots,N$. This representation is feasible because the frame condition holds (Lazar & Pnevmatikakis, 2011b) and the set or recovery functions is an overcomplete basis for $\mathcal{H}$. The vector of coefficients $\mathbf{c}$ can be evaluated by solving the system of linear equations (A.5).

Appendix B. The SIM(2) Group

Adapting the group notation, we define an element of SIM(2) $g = ([x_0, y_0], \alpha, \theta)$, where $x_0, y_0 \in \mathbb{R}, \alpha \in \mathbb{R}^+$ and $\theta \in [0, 2\pi)$. Each of $g \in$ SIM(2) is an elementary transformations (translations, dilations and rotations) on the $\mathbb{R}^2$ plane, where the transformation is given by

$$[x', y'] = ([x_0, y_0], \alpha, \theta)[x, y] = \alpha r_\theta [x, y] + [x_0, y_0].$$

The group law (which consists by the action $\circ$ of an element of the group on another element, the identity element, and the inverse element) is given by

$$\begin{aligned} g' \circ g = ([x_0', y_0'], \alpha', \theta') \circ ([x_0, y_0], \alpha, \theta) &= ([x_0', y_0'] + \alpha' r_{\theta'}[x_0, y_0], \alpha'\alpha, \theta' + \theta), \\ e &= ([0,0], 1, 0), \\ g^{-1} = ([x_0, y_0], \alpha, \theta)^{-1} &= (-\alpha^{-1} r_{-\theta}[x_0, y_0], \alpha^{-1}, -\theta), \end{aligned} \tag{B.1}$$

and the associativity can be easily verified.

Appendix C. Proof of Theorem 2.4

Proof of Theorem 2.4:
Since the filterbank is structured, we will use the compact notation in Section 2.2.3

as (k, m, n, l) with $k \in \mathbb{N}, m \in \mathbb{Z}, n \in \mathbb{Z}/N, l \in \mathbb{Z}/L$. Moreover, since the discretized subset is countable, the elements of the filterbank can be ordered in one dimension with an injective mapping $y : \mathbb{N} \times \mathbb{Z} \times \mathbb{Z}/N \times \mathbb{Z}/L \mapsto \mathbb{Z}$, where $\mathbb{Z}/N$ and $\mathbb{Z}/L$ denote the set of integers modulo N and L, respectively. The above transformations correspond to a connectivity setting of the switching matrix such that

$$y^{-1}(\sigma(y(k, m, n, l))) - (k, m, n, l) = const. \tag{C.1}$$

The interpretation of (C.1) is that for any operator $\mathcal{T}$ described above, i.e., $\mathcal{T} \in H_p$, acting on the input video I, it is sufficient to channel the spikes coming from the neural circuit j to the l-th entry of the next level, i.e., $\sigma(j) = l$, where l satisfies $D^j = \mathcal{T} D^l$. Since the set of spatial receptive fields is invariant under any $\mathcal{T} \in H_p$, such an l always exists and is given by (13) and (C.1) depending on the transformation. Moreover each l is unique, i.e., the permutation σ is injective and σ^{-1} exists.

It remains to prove that the set of "switched" spike trains

$$\sigma(\boldsymbol{\tau}) = \left\{ (t_k^{\sigma^{-1}(j)}), k \in \mathbb{Z}, j = 1, 2, \ldots, N \right\}$$

represents the video input $\mathcal{T} I$. From the t-transform representation of (A.2) we see that the receptive field output v^j produces the spike train (t_k^j). Similarly to (A.1) let us define $v_{\mathcal{T}}^j(t)$ as

$$v_{\mathcal{T}}^j(t) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}^2} D_s^j(x, y) D_\tau(t) (\mathcal{T} I)(x, y, t - s)\, dxdy \right) ds, \tag{C.2}$$

i.e., the output of the j-th receptive field when the input is $\mathcal{T} I$ (here we also used the space-time separability of the receptive fields). Since $\mathcal{T}$ is a unitary operator we have that

$$v_{\mathcal{T}}^j(t) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}^2} (\mathcal{T}^{-1} D_s^j(x, y)) D_\tau(s) I(x, y, t - s)\, dxdy \right) ds = v^{\sigma^{-1}(j)}(t). \tag{C.3}$$

In other words, the input $\mathcal{T} I$ at the j-th spiking circuit produces the spike train $(t_k^{\sigma^{-1}(j)})$ and therefore the set $\sigma(\boldsymbol{\tau})$ represents the video input $\mathcal{T} I$. □

References

Anderson, C. H., & Essen, D. C. V. (1987). Shifter Circuits: A Computational Strategy for Dynamic Aspects of Visual Processing. *Proceedings of the National Academy of Sciences*, *84*, 6297–6301.

Antoine, J.-P., Murenzi, R., Vandergheynst, P., & Ali, S. T. (2004). *Two-Dimensional Wavelets and their Relatives*. Cambridge University Press.

Arathorn, D. W. (2002). *Map-Seeking Circuits in Visual Cognition*. Stanford, CA: Stanford University Press.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical Support for a Two-Dimensional View Interpolation Theory of Object Recognition. *Proceedings of the National Academy of Sciences*, *89*, 60–64.

Christensen, O. (2003). *An Introduction to Frames and Riesz Bases*. Applied and Numerical Harmonic Analysis. Birkhäuser.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*, 333–341.

Dodwell, P. (1983). The Lie transformation group model of visual perception. *Attention, Perception, & Psychophysics*, *34*, 1–16.

Field, G. D., & Chichilnisky, E. J. (2007). Information Processing in the Primate Retina: Circuitry and Coding. *Annual Review of Neuroscience*, *30*, 1–30.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*, 891–906.

Hoffman, W. (1966). The Lie algebra of visual perception. *Journal of Mathematical Psychology*, *3*, 65–98.

Jones, J., & Palmer, L. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*, 1233–1258.

Kim, A. J., & Lazar, A. A. (2011). Recovery of stimuli encoded with a hodgkin-huxley neuron using conditional prcs. In N. Schultheiss, A. Prinz, & R. Butera (Eds.), *Phase Response Curves in Neuroscience*. Springer.

Kravitz, D., Vinson, L., & Baker, C. (2008). How position dependent is visual object recognition? *Trends in cognitive sciences*, *12*, 114–122.

Kuffler, S. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, *16*, 37.

Lazar, A. A. (2006). A Simple Model of Spike Processing. *Neurocomputing*, *69*, 1081–1085.

Lazar, A. A. (2010). Population encoding with Hodgkin-Huxley neurons. *IEEE Transactions on Information Theory*, *56*, 821–837.

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. (2009). Reconstruction of sensory stimuli encoded with integrate-and-fire neurons with random thresholds. *EURASIP Journal on Advances in Signal Processing*, *2009*, 13 pages. Special Issue on Statistical Signal Processing in Neuroscience.

Lazar, A. A., & Pnevmatikakis, E. A. (2011a). Encoding of multivariate stimuli with MIMO neural circuits. In *Proceedings of the ISIT 2011*. Saint Petersburg, Russia: IEEE.

Lazar, A. A., & Pnevmatikakis, E. A. (2011b). Video Time Encoding Machines. *IEEE Transactions on Neural Networks*, *2*, 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. Special Issue on Mathematical Models of Visual Coding.

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.

Lee, T. S. (1996). Image Representation Using 2D Gabor Wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*, 959–971.

Logothetis, N., & Sheinberg, D. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621.

Nattel, E., & Yeshurun, Y. (2000). An efficient data structure for feature extraction in a foveated environment. In *Biologically Motivated Computer Vision* (pp. 139–165). Springer.

Olshausen, B. A., Anderson, C. H., & Essen, D. C. V. (1993). A Neurobiological Model of Visual Attention and Invarinant Pattern Recognition Based on Dynamic Routing of Information. *The Journal of Neuroscience*, *13*, 4700–4719.

Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999). *Discrete-Time Signal Processing*. (2nd ed.). Prentice Hall.

Poggio, T., & Edelman, S. (1990). A Network that Learns to Recognize Three-Dimensional Objects. *Nature*, *343*, 263–266.

Rodieck, R. (1965). Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research*, *5*, 583–601.

Simoncelli, E., Freeman, W., Adelson, E., & Heeger, D. (1992). Shiftable multiscale transform. *IEEE Transactions on Information Theory*, *38*, 587–607.

Tarr, M. J., & Pinker, S. (1989). Mental Rotation and Orientation-Dependence in Shape Recognition. *Cognitive Psychology*, *21*, 233–282.

Ullman, S. (1989). Aligning Pictorial Descriptions: An Approach to Object Recogition. *Cognition*, *32*, 193–254.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*, 600–612.

Weber, C., & Triesch, J. (2009). Implementations and implications of foveated vision. *Recent Patents on Computer Science*, *2*, 75–85.

Westheimer, G., & Beard, B. (1998). Orientation dependency for foveal line stimuli: detection and intensity discrimination, resolution, orientation discrimination and vernier acuity. *Vision research*, *38*, 1097–1103.

Wohrer, A., & Kornprobst, P. (2009). Virtual retina: A biological retina model and simulator, with contrast gain control. *Journal of computational neuroscience*, *26*, 219–249.

Wolfrum, P., & von der Malsburg, C. (2007). What is the Optimal Architecture for Visual Information Routing? *Neural Computation*, *19*, 3293–3309.

The Power of Connectivity: Identity Preserving Transformations on Visual Streams in the Spike Domain
List of Supplementary Materials *

Aurel A. Lazar
Department of Electrical Engineering
Columbia University, New York, NY 10027
aurel@ee.columbia.edu

Eftychios A. Pnevmatikakis
Department of Statistics and Center for Theoretical Neuroscience
Columbia University, New York, NY 10027
eftychios@stat.columbia.edu

Yiyin Zhou
Department of Electrical Engineering
Columbia University, New York, NY 10027
yiyin@ee.columbia.edu

January 23, 2013

*The work presented here was supported by AFOSR under grant # FA9550-09-1-0350 and, in part, by a grant of computer time from the City University of New York High Performance Computing Center under NSF Grants CNS-0855217 and CNS-0958379. The author names are listed alphabetically.

The supplementary materials consists of a total of six videos and two figures.

1. **video1.mov:**
   Recovery and Rotations on the Log-Polar Grid. The video consists of 4 parts. Starting from left to right, they are: **(A)** Original visual stream. **(B)** Recovery with a null setting switching matrix. **(C)** Recovery with a switching matrix setting realizing a rotation of 45 degrees counter-clockwise. **(D)** Recovery with a switching matrix setting realizing a rotation of 171 degrees clockwise. SSIM index for (B-D) is, respectively, $0.97, 0.97, 0.97$.

   Difference of Gaussian receptive fields were used. The parameters of the Log-Polar grid are the same as in Section 3.1:

   - $\alpha_0 = 2, m \in \{-3, -2, -1, 0, 1\}$,
   - $\theta_0 = 2\pi/L, L = 120, l \in \{0, 1, \cdots, 119\}, \omega_0 = 2\pi, N = 1$,
   - $b_0 = 0.8$.

2. **video2.mov:**
   Dilations and Rotations on the Log-Polar Grid. The video consists of 4 parts. Starting from left to right, they are: **(A)** Original visual stream. **(B)** Recovery with a switching matrix setting realizing a dilation by a factor of 2. The result was multiplied by 2 before being displayed. The region of reconstruction corresponds to the region inside the dashed circle in the original visual stream. **(C)** Recovery with a switching matrix setting realizing a dilation by $1/2$. The result was multiplied by $1/2$ before being displayed. The region of reconstruction corresponds to the region inside the solid circle in the original visual stream. **(D)** Recovery with a switching matrix setting realizing a simultaneous dilation by $1/2$ and rotation by 63 degree counter-clockwise. The result was multiplied by $1/2$ before being displayed. The region of reconstruction corresponds to the region inside the solid circle in the original visual stream. SSIM index for (B-D) is, respectively, $0.97, 0.90, 0.91$.

   The same receptive fields and grid parameters as in video1 above were used.

3. **video3.mov:**
   Recovery and Translations on the Cartesian Grid. The video consists of 4 parts. Starting from left to right, they are: **(A)** Original visual stream. **(B)** Recovery with a null setting switching matrix. **(C)** Recovery with a switching matrix setting realizing a translation by 2 units to the left and by

2 units downwards. **(D)** Recovery with a switching matrix setting realizing a translation by 4 units upwards. SSIM index for (B-D) is, respectively, $0.97, 0.96, 0.96$.

Gabor receptive fields were used. The parameters of the Cartesian grid are the same as in Section 3.2:

- $\alpha_0 = 2, m \in \{0, 1, 2, 3\}$
- $b_0 = 2$,
- $\omega_0 = 2\pi/N, N = 8, n \in \{0, 1, 2, 3, 4, 5, 6, 7\}$,

4. **video4.mov:**
Dilations and Translations on the Cartesian Grid. The video consists of 4 parts. Starting from left to right, they are: **(A)** Original visual stream. **(B)** Recovery with a switching matrix setting realizing a dilation by a factor of 2. The result was multiplied by 2 before being displayed. The region of reconstruction corresponds to the region inside the dashed rectangular frame in the original visual stream. **(C)** Recovery with a switching matrix setting realizing a dilation by a factor of $1/2$. The result was multiplied by $1/2$ before being displayed. The region of reconstruction corresponds to the region inside the solid rectangular frame of the original visual stream. **(D)** Recovery with a switching matrix setting realizing a simultaneous dilation by a factor of $1/2$ and a translation by 1 unit to the right. The result was multiplied by $1/2$ before being displayed. SSIM index for (B-D) is, respectively, $0.90, 0.90, 0.90$.

The same receptive fields and grid parameters as in video2 above were used.

5. **video5.mov:**
Approximate Rotations on the Cartesian Grid. The video consists of 4 parts. Starting from left to right, they are: **(A)** Original visual stream. **(B)** Recovery with a switching matrix setting realizing an approximate rotation by 46 degrees clockwise. **(C)** Recovery with a switching matrix setting realizing an approximate rotation by 92 degrees clockwise. **(D)** Recovery with a switching matrix setting realizing an approximate rotation by 138 degrees clockwise. SSIM index for (B-D) is, respectively, $0.90, 0.90, 0.90$.

Difference of Gaussian receptive fields were used. The parameters of the Cartesian grid are the same as in Section 3.3:

- $\alpha_0 = 2, m \in \{0, 1, 2, 3\}$,
- $\omega_0 = 2\pi, N = 1$,
- $b_0 = 0.8$.

6. **video6.mov:**
   Approximate Translations on the Log-Polar Grid. The video consists of 4 parts. Starting from left to right, they are: **(A)** Original visual stream. **(B)** Recovery with a switching matrix setting realizing an approximate translation by 0.5 units to the right. **(C)** Recovery with a switching matrix setting realizing an approximate translation by 1 unit to the right. **(D)** Recovery with a switching matrix setting realizing an approximate translation by 1.5 units to the right. SSIM index for (B-D) is, respectively, $0.97, 0.94, 0.94$.

   The same receptive fields and grid parameters as in video1 above were used.

7. **Figure S1:**
   Figure S1 illustrates an example of discretization of the SIM(2) group to Log-Polar Grid, as defined by (15) in Section 2.2.3, where $N = 8, L = 30, \alpha_0 = 2, b_0 = 0.8$, and the spatial grid of three levels of $m \in \{0, -1, -2\}$ is depicted in each of the subfigures, corresponding to three different dilation levels (dilation parameter of the group elements). Each element in the grid is represented by a small oriented line segment. The spatial position of the center of a line segment corresponds to the translation parameter of group element, and the angle between a line segment and x axis corresponds to the rotation parameter of the group element. The color of each line segment corresponds to a particular value of n in (15) modulo 4, where blue, magenta, red and black correspond to $n = 0, 1, 2, 3$, respectively. Note that at the origin, only one of the L elements are shown for clarity.

8. **Figure S2:**
   Figure S2 illustrates an example of discretization of the SIM(2) group to Cartesian Grid, as defined by (22) in Section 2.2.4, where $N = 8, \alpha_0 = 2, b_0 = 0.8$, and the spatial grid of three values of $m \in \{0, 1, 2\}$ is depicted in each of the subfigures, corresponding to three different dilation levels (dilation parameter of the group elements). Each element in the grid is represented by a small oriented line segment. The spatial position of the center of a line segment corresponds to the translation parameter of group element, and the angle between a line segment and x axis corresponds to the rotation parameter of the group element. The color of each line segment

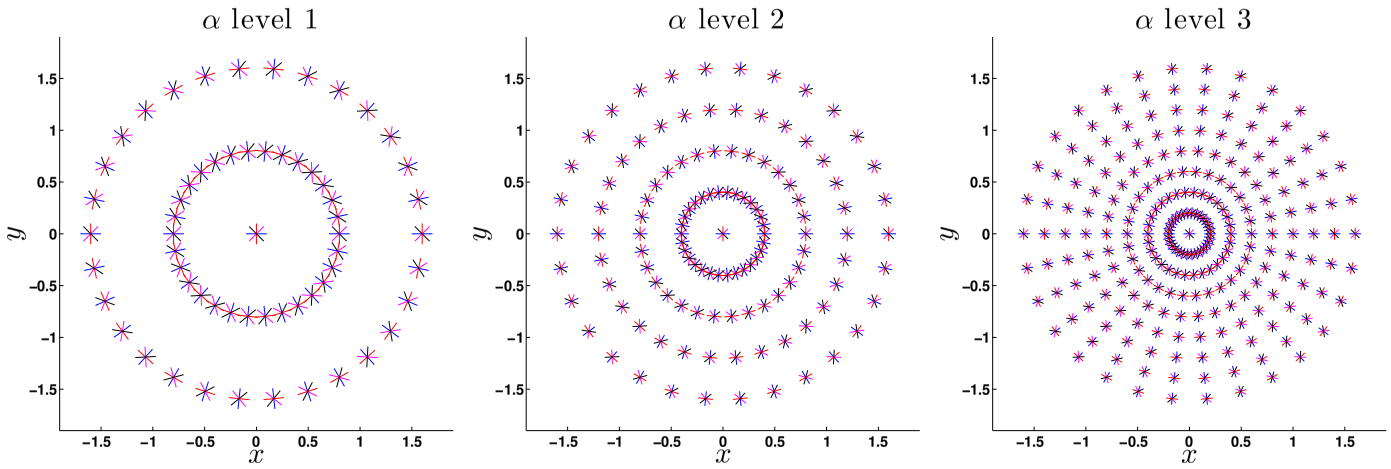


Figure 1: An example of Log-Polar Grid.

corresponds to a particular value of n (22) modulo 4, where blue, magenta, red and black correspond to $n = 0, 1, 2, 3$, respectively.

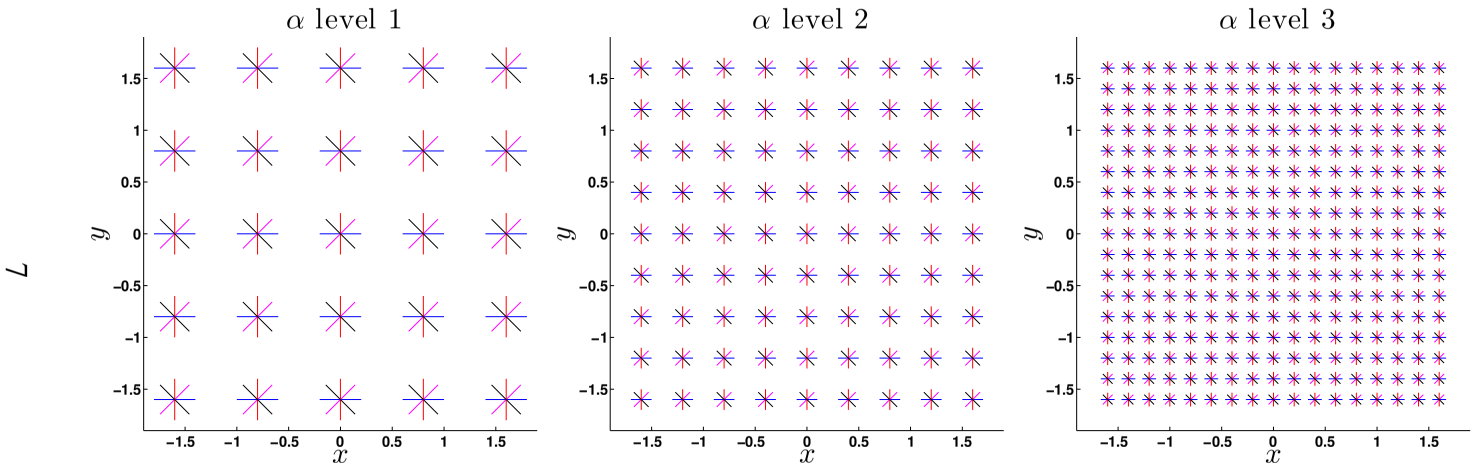


Figure 2: An example of Cartesian grid.