Differential response of the retinal neural code with respect to the sparseness of natural images

Cesar U Ravello, Maria-Jose U Escobar, Adrian U Palacios, Laurent U Perrinet

Differential response of the retinal neural code with respect to the sparseness of natural images *(annotated manuscript + supplementary material)*

Cesar R. Ravello
PhD Program in Neuroscience
Centro Interdisciplinario de Neurociencia de Valparaíso
Facultad de Ciencias, Universidad de Valparaiso
2360102 Valparaíso, Chile
ravello.cesar@gmail.com

Maria-Jose Escobar
Universidad Técnica Federico Santa María
Departamento de Electrónica
2390123 Valparaíso, Chile
mariajose.escobar@usm.cl

Adrian Palacios
Centro Interdisciplinario de Neurociencia de Valparaíso
Facultad de Ciencias, Universidad de Valparaiso
2360102 Valparaíso, Chile
adrian.palacios@uv.cl

Laurent U. Perrinet*
Institut de Neurosciences de la Timone (UMR7289),
CNRS / Aix-Marseille Université
27, Bd. Jean Moulin, 13385 Marseille Cedex 5, France
Laurent.Perrinet@univ-amu.fr

Abstract

Natural images follow statistics inherited by the structure of our physical (visual) environment. In particular, a prominent facet of this structure is that images can be described by a relatively sparse number of features. To investigate the role of this sparseness in the efficiency of the neural code, we designed a new class of random textured stimuli with a controlled sparseness value inspired by measurements of natural images. Then, we tested the impact of this sparseness parameter on the firing pattern observed in a population of retinal ganglion cells recorded *ex vivo* in the retina of a rodent, the *Octodon degus*. These recordings showed in particular that the reliability of spike timings varies with respect to the sparseness with globally a similar trend than the distribution of sparseness statistics observed in natural images. These results suggest that the code represented in the spike pattern of ganglion cells may adapt to this aspect of the statistics of natural images.

*See http://invibe.net/LaurentPerrinet.

Submitted to 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

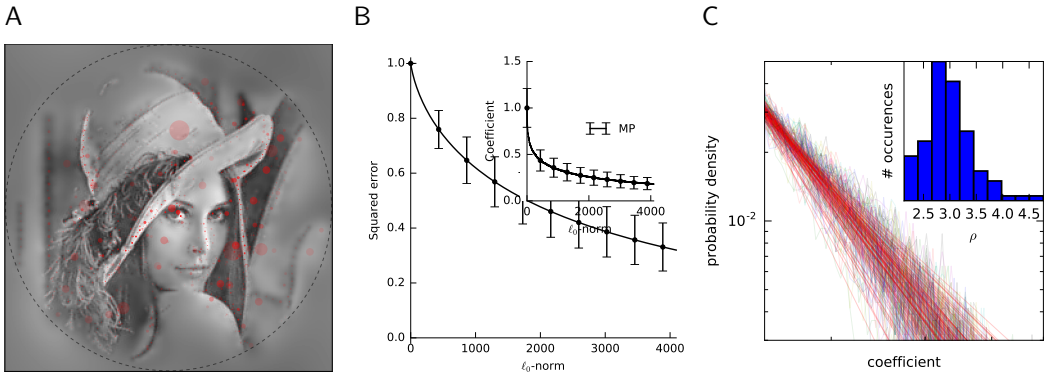

Figure 1: **Sparse coding of images in the retina follows regular statistics**: (A) An instance of the reconstruction of a natural image ("Lena") with the list of extracted image elements overlaid — see SI Section 6.1 for a full description of the algorithm. Edges outside the dashed circle are discarded to avoid artifacts. Parameters for each element are its position, scale and scalar amplitude. (B) Coefficients decrease as a function of their rank of extraction (average across images $\pm$ standard deviation). This decrease is faster when using a sparse coding mechanism (Matching Pursuit, 'MP') than when just ordering the linear coefficients ('lin'). (B-inset) For the sparse coding, we controlled the quality of the reconstruction from the edge information such that the residual energy is less than 10% over the whole set of images, a criterion met on average when identifying 8192 edges per image for images of size 256×256 (that is, a relative sparseness of $\approx 9\%$ of activated coefficients). (C) When observing on an image-by-image basis the histogram of the coefficients's amplitude, each follows a generic power-law probability density function, only controlled by a sparseness parameter ρ for each of the natural images, higher values of ρ denoting progressively sparser structures and in addition, Supplementary Figure 7 shows that once scaled, these distribution align almost perfectly. (C-inset) The distribution of these sparseness parameters shows a variety of values from the most sparsely distributed (right, $\rho \approx 4$) to those containing mostly dense textures (left, $\rho \approx 2$).

1 Motivation

Natural images, that is, visual scenes that are relevant for an animal, most generally consist of the composition of visual objects, from vast textured backgrounds to single isolated items. By the nature of the structure of the physical (visual) word, these visual items are often sparsely distributed, such that these elements are clustered and a large portion of the space is void: A typical image thus contains vast areas which are containing little information. A simplistic formalism to quantify this sparseness is to consider the coding of natural images using a bank of filters resembling the Mexican-hat profiles observed in the retina of most mammals (See Supplementary Figure 6). Then, it is possible to determine the efficiency of a near-to-optimal coding formalism [22] to show that any image from a database of static, grayscale natural images may be coded solely by a few coefficients (See Figure 1 ; see SI Section 6.1 for a full description of the algorithm). Moreover, we observed that on this database, some images were sparser than others but that globally they all fitted well a similar power-law probability distribution function. The exponent of this power-law provides an unique descriptor for sparseness which is characteristic for this property [23]. However, it is largely unclear how this property inherent to any natural image may be used in the early visual system and in particular in the retina.

Indeed, a normative explanation for the coding of sensory inflow into spike patterns in the optic nerve is the optimal representation of visual information [1, 3]. A popular technique to challenge this hypothesis is to test the system using complex stimuli [32] and in the retina, this takes the form of the use of stochastic stimuli and the inversion of a model of the neural transformation [31]. A major result is that the receptive field of ganglion cells capture information according to maximally informative directions [30]. Various aspects of the complex processing that happens in the retina has been captured in models for describing the structure of the retina with respect to spike patterns, from Spike Trigerred average [18], Spike-Trigerred Covariance and quadratic models [25] to more general techniques using variants of the Linear / Non-Linear model [31]. However, a difficulty is that

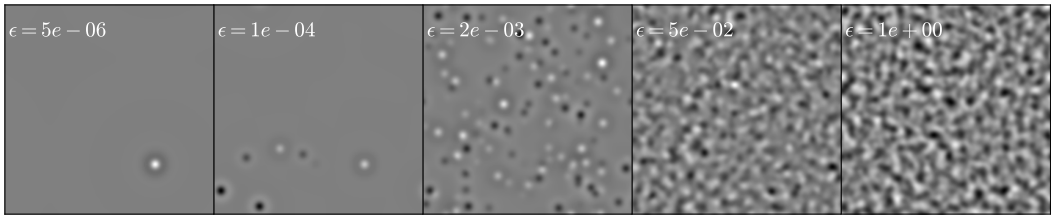

Figure 2: **DropLets stimuli:** Based on the Motion Clouds framework, we design DropLets as random phase textures similar to Motion Clouds but where, similarly to the variety observed in a set of natural images, the distribution of coefficients is parameterized by different levels of sparseness. For the sake of simplicity, sparseness is obtained here by thresholding a mask value and is given by the ratio ϵ of non-zero coefficients, from very sparse ($\epsilon = 5.10^{-6}$) to fully dense ($\epsilon = 1 = 100\%$), and with ϵ parameters chosen on a geometrical scale. We show here one single frame of each dynamical stimulus. Note that, while for $\epsilon = 1$ these textures are fully dense and correspond to a linearly filtered noise, the other versions correspond to progressively sparser textures as ϵ tends to zero. Moreover, while the textons used to synthesize these textures are arbitrary, we have chosen here to use symmetrical Mexican-hat profiles ("drops" , see Supplementary Figure 6) which are known to preferentially evoke activity in the retina.

these models often assume that the noise models are dense and that introducing sparseness results in non-linear effects which are hard to describe using these models [6]. As such, a major challenge is to refine the definition of neural coding so that they conform to the widest variety of results.

Here, we seek to engender a better understanding of neural coding in the retina by using a novel generative model for dynamic texture synthesis. Our first contribution is to show that while the distribution of coefficients in natural images is stereotyped, it displays a distribution of sparseness parameters (see Figure 1). From that perspective, we motivate the generation of an optimal stimulation within a stationary Gaussian dynamic texture model. We base our model on a previously defined heuristic coined "Motion Clouds" [28]. Our second contribution is an efficient extension of this model which allows to parameterize the sparseness in the texture, for instance based on the measurements made on natural images. This formulation is important to provide a clear understanding of the effects of the model's parameters manipulated during neurophysiological experiments. Our third contribution in this paper is to demonstrate an effect of this sparseness on the spike pattern of retinal ganglion cells from a rodent, the *Octodon degus* as recorded *ex vivo* on an electrode grid. Our last contribution is to quantify the efficiency of the neural coding in these recordings and to show that the code is more reliable with a given sparseness level rather than with dense stimuli. These contributions shows that overall, the processing in the retina may be optimized to match the sparseness encountered in natural scenes. Beyond the contribution to the understanding of the neural code underlying image representation, this may have long ranging applications, for instance for the optimal parameterization of stimulations in a retinal implant.

2 Design of the "DropLets" stimuli

From these observations, the understanding of the neural code in the retina is linked to the definition of an adequate generative model. We propose a mathematically-sound derivation of a general parametric model of dynamic textures. This model is defined by aggregation, through summation, of a basic spatial "texton" template $\phi(x, y, t)$, eventually transformed by any arbitrary visuo-geometric transformation g_λ such as zooms or rotations, which themselves are parameterized by λ. The summation reflects a transparency hypothesis, which has been adopted for instance in [7]. In particular, this simple generative model considers independent, transparent elementary features. While one could argue that this hypothesis is overly simplistic and does not model occlusions or edges, it leads to a tractable framework of stationary Gaussian textures, which has proved useful to model static micro-textures [7] and dynamic natural phenomena [34]. In particular, the simplicity of this framework allows for a fine tuning of frequency-based (Fourier) parameterization, which is desirable for the interpretation of neuro-physiological experiments. In summary, it states that

luminance $I(x, y, t)$ for $(x, y, t) \in \mathbb{R}^2 \times \mathbb{R}$ is defined as a random field [33]:

$$I(x, y, t) = \sum_{i \in \mathbb{N}} a_i \cdot g_{\lambda_i}(\phi(x - x_i, y - y_i, t - t_i)) \tag{1}$$

where the respective random variables parameterize respectively each texton's geometric transformation parameter λ_i, scalar value a_i and positions and timings (x_i, y_i, t_i). Intuitively, this model corresponds to a dense mixing of stereotyped, static textons as in [7].

As was previously mentioned [28, 33], this set of Motion Clouds stimuli is equivalently defined by:

$$I(x, y, t) = \sum_{i \in \mathbb{N}} a_i \cdot (g_{\lambda_i}(\phi(x, y, t)) * \delta(x - x_i, y - y_i, t - t_i)) \tag{2}$$

where $*$ denotes the convolution operator. Noting the Fourier transform as $\mathcal{F}$ and $(f_x, f_y, f_t) \in \mathbb{R}^2 \times \mathbb{R}$ the frequency variables, the image I is thus a stationary Gaussian random field of covariance having the power-spectrum $\mathcal{E} = \mathcal{F}(\phi)$ (for a proof, see [33]). It comes

$$\mathcal{F}(I)(f_x, f_y, f_t) = A \cdot e^{i \cdot \Phi} \cdot \mathcal{E}(f_x, f_y, f_t)$$

where

$$A \cdot e^{i \cdot \Phi} = \sum_{i \in \mathbb{N}} a_i \cdot e^{-2i\pi(f_x \cdot x_i + f_y \cdot y_i + f_t \cdot t_i)}$$

corresponds to an iid random phase field scaled by $A \in \mathbb{R}^+$. Such a field was called a Motion Cloud (MC) and it is parameterized by the (positive-, real-valued) envelope $\mathcal{E}$. To match the statistics of natural scenes or some category of textures, the envelope $\mathcal{E}$ is usually defined as some average spatio-temporal coupling over a set of natural images. Note that it is possible to consider any arbitrary texton $\mathcal{F}^{-1}(\mathcal{E})$, which would give rise to more complicated parameterizations for the power spectrum, but we decided here to stick to a simple case which is relevant to study coding in the retina. In particular, we limited ourselves to rotationally symmetric envelopes and to spatio-temporal scale transformations (zooms).

The originality provided here is to introduce an explicit sparsity term in the summation. As such, this sparse Motion Clouds is similarly defined by equation 2 but with a sparse distribution of coefficients a_i. Mathematically, the a_i are drawn from a heavy-tailed probability distribution function p_a such as that found in natural images (See Figure 1-C). Such assumptions were previously used for generating procedural noise in computer vision [13], but we focus here on the generation on a model of sparseness in images. Such endeavor was initiated by [27] by defining a mixture of a Dirac and a Gaussian distributions, but here, we derive it from the sparse coding observed in a set of natural images. First let's define E the (sparse) event matrix corresponding to the events' scalar values, parameters, positions and timings forming the set of non-zero values $\mathcal{S}$, that is with $E = \sum_{i \in \mathcal{S}} a_i \cdot \delta(x - x_i, y - y_i, t - t_i)$. For the sake of simplicity, the sparseness parameter ϵ will thus denote the relative ratio of non-zero coefficients from a fully dense matrix with $\epsilon = 1$ to an empty matrix with $\epsilon = 0$. First note that this parameter can be trivially linked to the sparseness coefficient that we defined for natural images as $\epsilon \propto \exp(-\rho)$. We also note that we can still compute the image in Fourier space, but with a random phase field such that

$$A \cdot e^{i \cdot \Phi} = \sum_{i \in \mathcal{S}} a_i \cdot e^{-2i\pi(f_x \cdot x_i + f_y \cdot y_i + f_t \cdot t_i)}$$

Second, we note that the random phase field is simply the Fourier transform on E:

$$A \cdot e^{i \cdot \Phi} = \mathcal{F}(E) \tag{3}$$

Compared to the pure summation, a first advantage of this procedure is that the computation time does not depend on the number of events. A further advantage of equation 3 in generating these stimuli is that for any given instance of the noise, we know the position (x, y, t) of each event. As they match the shot noise image of drops of rain on water, we coined this set of stimulus as "DropLets". The implementation of this texture generation is of the same order of complexity as that of Motion Clouds and allows for the easy control of the sparseness in the stimulus (See Figure 2). Once this set of "DropLets" stimuli is defined, we may now design a protocol to stimulate the retina and explicitly study the role of sparseness.

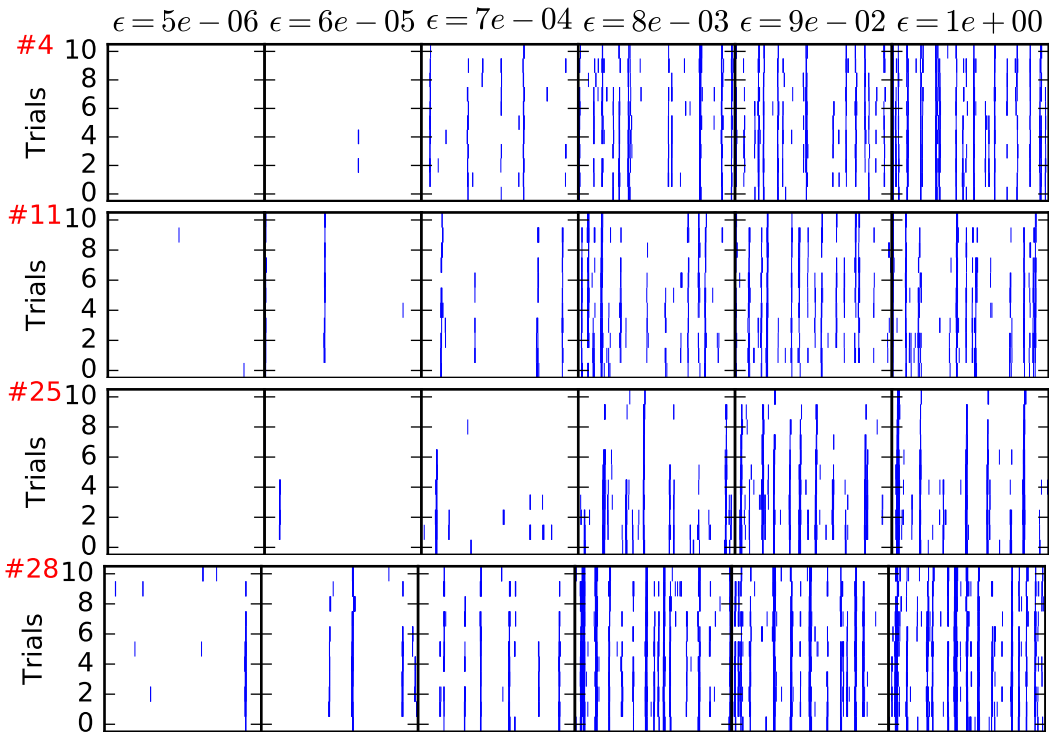

Figure 3: **Raster plots for different representative ganglion cells** (different rows, labelled by the red number) as a function of sparseness ϵ (different columns, same values of ϵ as in Figure 2). In each raster plot, we present spikes as a function of time (x-axis, the duration of a block is 24 seconds) and for 11 different trials (y-axis). First, one observes that the global, mean firing rate of each cell increases with sparseness. This was also consistent for the whole population. Second, the raster plots show that the spiking is very reproducible across trials, showing that this class of random textures could reliably evoke activity in the retina. Moreover, this pattern is very different across cells, indicating that the visual signal is well multiplexed between ganglion cells.

3 Results: neural activity in a population of ganglion cells

To characterize the response to the DropLets stimuli with respect to different sparseness parameters ϵ, we recorded *ex vivo* the electrophysiological activity of a retinal patch using a multi-electrode array (USB-MEA256, Multi Channel Systems MCS GmbH), mounted on an inverted microscope. Stimuli were projected at $60Hz$ using a conventional DLP projector with custom optics to focus the image onto the photoreceptor layer with a pixel size of $\approx 4.7\mu m$ and an average irradiance of $70nW/mm^2$. The images were presented using a custom software built upon the PsychToolbox library [2] for MATLAB. Spike sorting was performed using the procedure described in [15]. The protocol consisted of randomly interleaved 24s long, 100×100 pixels DropLets stimuli with 6 levels of sparseness from very sparse ($\epsilon = 5.10^{-6}$) to fully dense ($\epsilon = 1$), and with parameters chosen on a geometrical scale (see Figure 2). Each condition was repeated 11 times for a total duration of ≈ 1 hour (see SI Section 6.4 for a full description of the protocol). To analyze the activity evoked by the stimuli, first we computed the post-stimulus raster (PSR) for every condition tested and selected the conditions that elicited robust responses, in the sense that the response was strong and similar among trials of the same condition. The conditions that elicited the largest responses correspond to the sequences with the lowest spatial frequencies tested (data not shown). Rasters show that globally, the firing rates increases with the sparseness parameter ϵ. Qualitatively, a surprising result is that for a majority of ganglion cells, the firing pattern is heterogeneous across cells but very reproducible with different repetitions of the same (frozen) random texture (see Figure 3).

From the response to these selected sequences, we computed a simple Spike Triggered Average for image matrices of 100×100 pixels and 1440 frames for each condition, with a memory of 18 frames

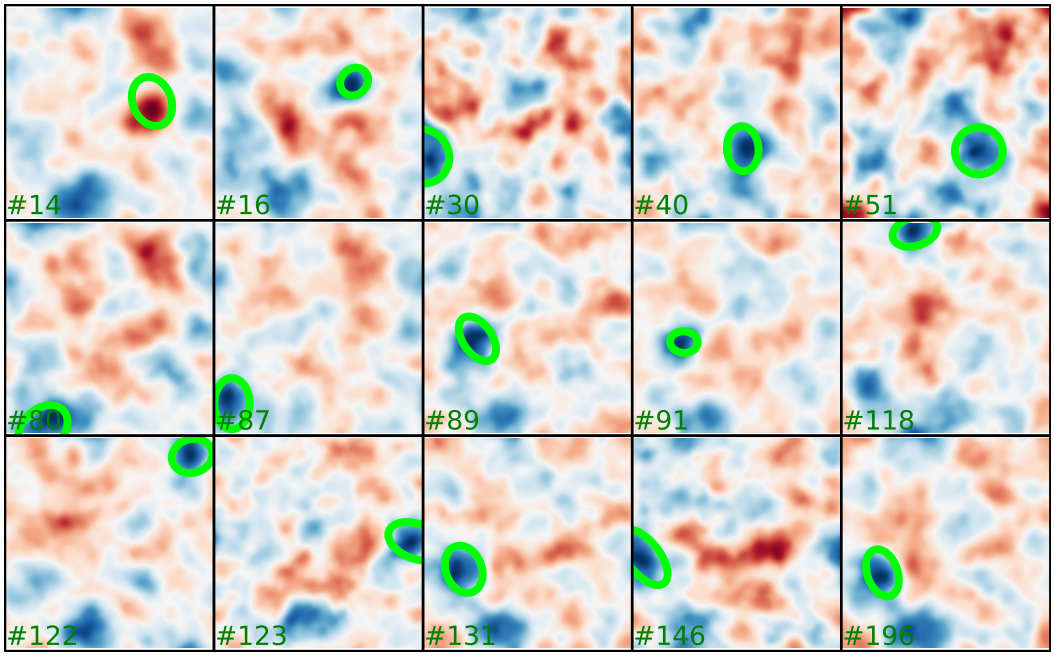

Figure 4: **Spike-Triggered Average**: Estimations of the Receptive Fields of representative retinal ganglion cells (whose number is given in green) computed from the response to the droplets (red-blue image) and compared to the fit performed on that computed using a checkerboard stimuli (green ellipse ; see Supplementary Figure 8 for a characterization of the whole population). Each image represents the average over the 18 frames ($\approx 25\text{ms}$) preceding a spike. For many cells, our estimation of the RF is very close to the one obtained by the traditional method (albeit generally with a longer temporal profile as can be seen in Supplementary Figure 9). Notice that the time of stimulation required is much shorter.

($\approx 25\text{ms}$). Examples of the obtained results can be seen in figure 4. This method yields a good estimation of the RF for some cells, but for many of them the result can be somewhat ambiguous or too noisy, as can be seen in the figure. This can be greatly improved by computing the STA across conditions. As can be seen in figure 4, the result is comparable to the estimation obtained with a STA computed from the response from a checkerboard stimulus. However, when analyzing the temporal response, the results are not the equivalent. In the case of the DropLets, the temporal aspect of the RF is always longer or slower than for the RF computed from the checkerboard. In the current work, we did not explore the parameter that controls the lifetime of each droplet, and it appears that the one that was chosen is higher than needed. Another issue, probably a bias caused by the properties of the stimuli, is that for some cells the size of RF is overestimated in comparison with the checkerboard. This is an interesting point that will be further analyzed in future work; if the RF is indeed smaller, then it would mean that the DropLets elicit a response even when also stimulating the periphery of the RF. Another explanation would be that the STA from the checkerboard is underestimating the size of the RF, and in this case the DropLets would yield a better estimation. Further research is already being performed to clarify this issue.

For the moment, we can conclude that it is indeed possible to estimate the RF from the response to some of the DropLets stimuli, however, there are still issues to be solved. First, and the easiest to fix, is the temporality or duration of the features of the stimuli, as apparently they last longer than needed to elicit a response from the cells. The second issue that could be solved is to optimize the size of the features. Clearly, there were a lot of sequences that did not stimulate the retina to get reliable responses, probably because the spatial frequency was too high so the features were undetectable by the retina under the experimental conditions. Finally, the most important issue would be the noise in the estimation. Even though the RF obtained from the DropLets presented a good match with the ones computed in the traditional way, in many cases the result contained additional signals that may yield conflicting interpretations. One approach to remedy this would be to keep exploring the importance of the sparseness of the stimuli, following the idea that if the features are more isolated

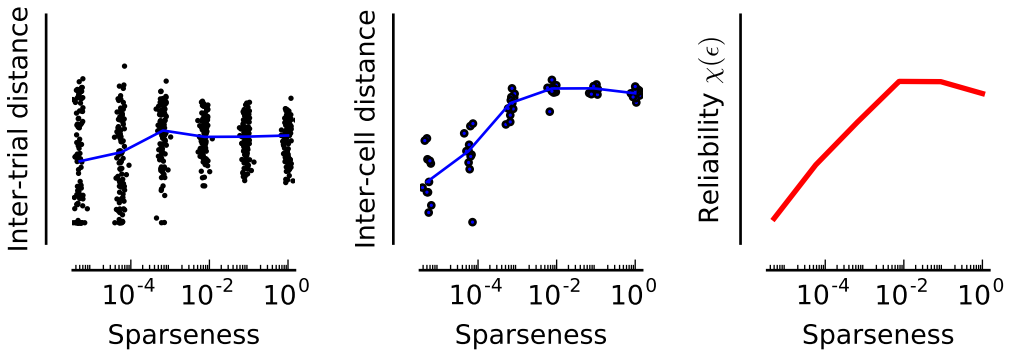

Figure 5: **Reliability as a function of sparseness.** (A) We first computed for each cell and sparseness level the inter-trial distance as the mean pair distance between two repetitions of the same stimulus, averaged across each block (in time) and for all pairs of trials. We show these measures for each cell as a black dot, and the evolution of the mean for the whole population of ganglion cells (blue line). We show here that this distance decreases for sparseness levels closer to zero: spike trains are more reproducible when the stimulus is sparser. (B) As a control, we also measured the inter-cell distance. For any given trial (a blue dot), it is equal to the mean across all pairs of cells of the temporally-averaged distance between the pair of cells. it shows that as ϵ increases, the mean over trials of the inter-cell distance also increases. (C) Finally, we compute the relative reliability as the ratio between the inter-trial distance to the inter-cell distance. This shows that reliability is optimal for a specific range of ϵ values.

in time and space, then there would be less contamination of the STA. Another approach would be to use longer sequences or many sequences with different seeds, to get more spikes triggered by different images so the noise would be reduced when averaging them.

4 Results: role of sparseness in the efficiency of retinal processing

The above computation of a STA for this class of stimuli proved that this characterization is not adapted to evaluate the inherent representation corresponding to the measured spike trains. In particular, following previous studies [31], it confirms that the receptive field estimated by the STA varied with the type of stimulus (DropLets versus checkerboard), but also with the sparseness level. As we have shown above (see Figure 1), the sparseness level may change drastically in natural images and it is thus reasonable to think that the retina integrates mechanisms acquired during natural evolution to adapt to these different levels of sparseness. In particular, it seems that qualitatively, the spike trains are very reliable across trials for a given sparseness level if it is sufficient to elicit spikes but that at a certain level, there is a larger mixing of different sources (as the cumulative effect of the shot noise model in the texture generation) as we reach a dense mixing of features ($\epsilon = 1$). Neural efficiency in the retina is mainly characterized by the capacity of neurons to reliably represent visual information in the spike timings. As such, reliability as measured by the reproducibility of a spike train when being presented the same stimulus is of course not directly related to the efficiency of the neural code, but provides a convenient information to measure "post-hoc" the precision of spike timings with respect to a given stimulus.

To estimate quantitatively such intuition, we measured reliability as the average distance between all pairs of trials for any given condition ϵ (sparseness) and cell c (over the whole population). For any pair of observed spike trains, this distance was given as the SPIKE-distance as implemented in the PySpike package [12]. This measure gives for any pair (j, k) a value $S^{j,k}$ between 0 and 1, where 0 is achieved if and only if the spike trains are identical. The average over the $N = 11$ repetitions for any given condition, and for any given cell gives an average distance (also between 0 and 1) equal to

(see Figure 5-A):

$$S^a(\epsilon, c) = \frac{1}{N(N-1)/2} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} S^{j,k}(\epsilon, c)$$

Our goal is to compare this reliability for different levels of sparseness and in particular for different average levels of firing rate. To be invariant to such properties, we used an additional measure by measuring for each trial k the average distance between all pairs of cells (also between 0 and 1) equal to (see Figure 5-B):

$$S^c(\epsilon, k) = \frac{1}{M(M-1)/2} \sum_{j=1}^{M-1} \sum_{k=j+1}^{M} S^{j,k}(\epsilon, k)$$

where $M = 101$ corresponds to the number of cells and the pair of spike trains $\{j, k\}$ is chosen for any given fixed value of ϵ and k. Ultimately, for any given sparseness value ϵ, we define the reliability $\chi(\epsilon)$ as the inverse ratio between the distance average over trials and the global distance over cells :

$$\chi(\epsilon) = \frac{< S^c(\epsilon, k) >_{k \in 1 \ldots N}}{< S^a(\epsilon, c) >_{c \in 1 \ldots M}}$$

where $< \cdot_j >_j$ denotes the average of $\cdot_j$ over the different j. In Figure 5-C, we show a plot of $\chi(\epsilon)$ for the different cells with respect to the sparseness level ϵ. It shows that ratio increases (and thus that reliability increases) for the lowest values of sparseness and then saturates at a given level and finally decreases for a fully dense stimulus.

From these results it appears that in the population of recorded cells, efficiency as inferred by our measure of reliability, varied as a function of sparseness. Indeed, the results show that there is a high variability in the measured z-scores for the different cells, but that the overall trend shows that reliability increases with sparseness until it reaches an optimal value, and then it slightly decreases for the dense texture ($\epsilon = 1$). This quantitative measure is consistent with our qualitative intuition, that is, that the efficiency of the representation as measured by this reliability measure is optimal for a value of sparseness which is lower than a dense mixing of feature and that we measure here to $\epsilon = 8e - 03$. It should be however noted that in this experiment, we sampled only a limited number of sparseness values. Moreover, it is likely that the mechanism is implemented by a neural circuit (notably the circuitry horizontal to the retinal surface) which is most certainly adaptive and would ultimately give a different response when presented over a longer term with stimuli with for instance a lower average sparseness value. Finally, it remains still to be explored if a similar behavior would be observed this time for natural images with varying levels of sparseness.

5 Discussion

In this paper, we have characterized the different levels of sparseness which are present in natural images, synthesized random textures with parameterized levels of sparseness and finally, we have shown that the neural code as recorded in ganglion cells *ex vivo* responds differentially to these different levels of sparseness. First, we have replicated the observation that natural images follow a prototypical structure for the probability density function of the coefficients that characterize them. Importantly, we have shown that on an image-by-image basis, this structure is well captured by a single parameter ρ which encompass the sparseness of a given image, from a dense texture to a highly sparse configuration. Based on these results, we designed random texture which replicate this parameterization of their sparseness using the ratio of non-zero coefficients $\epsilon \propto \exp(-\rho)$. We used that set of stimuli to evaluate the response on ganglion cells on an *ex vivo* preparation and analyzed the response of the population of neurons as a function of the sparseness parameter. Such analysis proved that the retinal neural code showed differential response with respect to the sparseness of images and preliminary measures on efficiency suggest that it is tuned for a given level of sparseness. We have shown that there is a limit in sparseness for which the retina can respond optimally; beyond that point, the response is more or less the same, meaning that the retina is still coding the features present in the sparser sequences but is not responding to the additional features of the more dense sequences. We can relate these results to the hypothesis of optimal coding of natural scenes [8], in which the visual system has a limited capacity to transmit information, it has adapted through evolution to a small subset of all the possible images and optimally code them, discarding the irrelevant information. Thus, the efficiency of the retina to code the stimuli reaches its peak at a sparseness level that would

be closer to what the system has evolved to code, and when presented with stimuli containing more signals, it discards the additional information. Much research has been performed to investigate the relationship between natural images and optimal coding, although the focus has been mainly on the spatiotemporal correlations [24, 26]. We have shown here that when keeping the spatiotemporal components constant, the modulation of the sparseness of the stimuli has an evident effect on the retinal response, and more importantly, it allows us to see the level of sparseness beyond which the efficiency does not increase.

It is important to note that these results are the outcome of an interdisciplinary convergence between image processing (to characterize sparseness in natural images), mathematical modeling (for the synthesis of textures) and neurophysiology (for the recordings and their analysis). In particular, we demonstrated here an original framework in which neural recordings are not analyzed *post-hoc*, but are instead tuned by the design of parameterized stimuli. In particular, these stimuli are defined from the analysis of natural images. One limit of this study is that we limited ourselves to a simplistic class of textons (Mexican-hat shaped profiles , see Supplementary Figure 6), both for the analysis and synthesis but that ganglion cells in the retina are known to be selective to a wider class of stimulations. However, we believe that this class of stimuli is general enough to characterize a wide range of different cell types. Indeed, by looking at local combinations of events (such as doublets), one could characterize different sub-types of ganglion cells, both static (ON, OFF), oriented or even moving, such as to characterize for instance directionally selective ganglion cells. In particular, by manipulating the statistics in the event's matrix E, one could target more specifically each of these sub-types.

Finally, the framework proposed in this paper calls for a novel mathematical characterization of the neural code. Indeed, while the Linear / Non-Linear model underlying the computation of the STA has proven to reliably predict neural responses, this has been at the price of a complex machinery. We believe that a major limitation of this model is the fact that the spiking mechanism is modeled by an inhomogeneous Poisson process. Indeed, while mathematically tractable, it is however a poor model of the spiking mechanism observed in most cells. In particular, the difference of behavior between the Poisson model and neurons is most prominent when using sparse stimuli such as the one present in the most important set of stimuli for an animal, that is, natural scenes. Such precise firing has been recently shown in area V1 and may originate from a canonical push-pull mechanism [11]. Similarly, such a mechanism is likely to be present as early as in the retina but novel types of models would be necessary to uncover this aspect of neural computations.

6 Supplementary material

6.1 Sparse coding of natural images

A method for measuring the statistics of edge co-occurrences in natural images was demonstrated by Geisler et al. [9]. Here we extend their method in two important ways. First, we use an over-complete, multi-scale representation of edges, which is more similar to the receptive fields in the retina. Second, we use a synthesis model for the edge representation, so that the edges we detect are guaranteed to be sufficient to regenerate the image with a low error.

The first step of our method involves defining the dictionary of templates (or filters) for detecting edges. We use a log-Gabor representation, which is well suited to represent a wide range of natural images [4]. This representation gives a generic model of edges parameterized by their shape, orientation, and scale. We set the range of these parameters to match what has been reported for the responses in rodents' retina. In particular, we set the bandwidth of the Fourier representation of the filters to 1 and ∞ respectively in log-frequency and polar coordinates to get a family of circular but scale-selective filters (see Fischer et al. [5] and Supplementary Figure 6 for examples of such edges). This architecture is similar to that used by Geisler et al. [9]. Prior to the analysis of each image, we used the spectral whitening filter described by Olshausen and Field [17] to provide a good balance of the energy of output coefficients [19, 4].

A linear convolution model automatically provides a rotation and translation-invariant representation. Such invariance can be extended to scalings by choosing to multiplex these sets of filters at different spatial scales. Although orthogonal representations are popular for computer vision due to their computational tractability, it is desirable in our context that we have a high over-completeness in the representation to have a detailed measure of the association field. Ideally, the parameters of edges would vary in a continuous fashion, to provide relative translation, rotation, and scale invariance. We chose to have 8 dyadic levels (that is, doubling the scale at each level) for the set of 256×256 images. Tests with a range of different numbers ofscales yielded similar results. Finally, each image is transformed into a pyramid of coefficients. This pyramid consists of approximately $4/3 \times 256^2 \approx 8.7 \times 10^4$ pixels multiplexed on 8 scales, that is, approximately $.7 \times 10^6$ coefficients, an over-completeness factor of about 11.

This transform is linear and can be performed by a simple convolution repeated for every edge type. Following Fischer et al. [5], convolutions were performed in the Fourier (frequency) domain for computational efficiency. The Fourier transform allows for a convenient definition of the edge filter characteristics, and convolution in the spatial domain is equivalent to a simple multiplication in the frequency domain. By multiplying the envelope of the filter and the Fourier transform of the image, one may obtain a filtered spectral image that may be converted to a filtered spatial image using the inverse Fourier transform. We exploited the fact that by omitting the symmetrical lobe of the envelope of the filter in the frequency domain, the output of this procedure gives a complex number whose real part corresponds to the response to the symmetrical part of the edge, while the imaginary part corresponds to the asymmetrical part of the edge (see Fischer et al. [5] for more details). More generally, the modulus of this complex number gives the energy response to the edge (comparable to the response of complex cells in area V1), while its argument gives the exact phase. Such a representation is implemented Python scripts available at `XXXanonymousXXX`. This property further expands the richness of the representation.

Because this dictionary of edge filters is over-complete, the linear representation would give a inefficient representation of the distribution of edges (and thus of edge co-occurrences) due to *a priori* correlations between coefficients. Therefore, starting from this linear representation, we searched for the most sparse representation. Minimizing the ℓ_0 pseudo-norm (the number of non-zero coefficients) leads to an expensive combinatorial search with regard to the dimension of the dictionary (it is NP-hard). As proposed first by Perrinet et al. [20], we may approximate a solution to this problem using a greedy approach.

In general, a greedy approach is applied when finding the best combination is difficult to solve globally, but can be solved progressively, one element at a time. Applied to our problem, the greedy approach corresponds to first choosing the single filter Φ_i that best fits the image along with a suitable coefficient a_i, such that the single source $a_i\Phi_i$ is a good match to the image. Examining every filter

Φ_j, we find the filter Φ_i with the maximal correlation coefficient, where:

$$i = \text{argmax}_j \left(\left\langle \frac{\mathbf{I}}{\|\mathbf{I}\|}, \frac{\Phi_j}{\|\Phi_j\|} \right\rangle \right), \tag{4}$$

$\langle \cdot, \cdot \rangle$ represents the inner product, and $\| \cdot \|$ represents the ℓ_2 (Euclidean) norm. Since filters at a given scale and orientation are generated by a translation, this operation can be efficiently computed using a convolution, but we keep this notation for its generality. The associated coefficient is the scalar projection:

$$a_i = \left\langle \mathbf{I}, \frac{\Phi_i}{\|\Phi_i\|^2} \right\rangle \tag{5}$$

Second, knowing this choice, the image can be decomposed as

$$\mathbf{I} = a_i \Phi_i + \mathbf{R} \tag{6}$$

where $\mathbf{R}$ is the residual image. We then repeat this 2-step process on the residual (that is, with $\mathbf{I} \leftarrow \mathbf{R}$) until some stopping criterion is met. Note also that the norm of the filters has no influence in this algorithm on the choice function or on the reconstruction error. For simplicity and without loss of generality, we will thereafter set the norm of the filters to 1: $\forall j, \|\Phi_j\| = 1$. Globally, this procedure gives us a sequential algorithm for reconstructing the signal using the list of sources (filters with coefficients), which greedily optimizes the ℓ_0 pseudo-norm (i.e., achieves a relatively sparse representation given the stopping criterion). The procedure is known as the Matching Pursuit (MP) algorithm [14], which has been shown to generate good approximations for natural images [21].

For this work we made two minor improvements to this method: First, we took advantage of the response of the filters as complex numbers. As stated above, the modulus gives a response independent of the phase of the filter, and this value was used to estimate the best match of the residual image with the possible dictionary of filters (Matching step). Then, the phase was extracted as the argument of the corresponding coefficient and used to feed back onto the image in the Pursuit step. This modification allows for a phase-independent detection of edges, and therefore for a richer set of configurations, while preserving the precision of the representation.

Second, we used a "smooth" Pursuit step. In the original form of the Matching Pursuit algorithm, the projection of the Matching coefficient is fully removed from the image, which allows for the optimal decrease of the energy of the residual and allows for the quickest convergence of the algorithm with respect to the ℓ_0 pseudo-norm (i.e., it rapidly achieves a sparse reconstruction with low error). However, this efficiency comes at a cost, because the algorithm may result in non-optimal representations due to choosing edges sequentially and not globally. This is often a problem when edges are aligned (e.g. on a smooth contour), as the different parts will be removed independently, potentially leading to a residual with gaps in the line. Our goal here is not to get the fastest decrease of energy, but rather to provide a good representation of edges along contours. We therefore used a more conservative approach, removing only a fraction (denoted by α) of the energy at each pursuit step (for MP, $\alpha = 1$). We found that $\alpha = 0.5$ was a good compromise between rapidity and smoothness. One consequence of using $\alpha < 1$ is that, when removing energy along contours, edges can overlap; even so, the correlation is invariably reduced. Higher and smaller values of α were also tested, and gave classification results similar to those presented here.

In summary, the whole learning algorithm is given by the following nested loops in pseudo-code:

1. draw a signal $\mathbf{I}$ from the database; its energy is $E = \|\mathbf{I}\|^2$,
2. initialize sparse vector $\mathbf{s}$ to zero and linear coefficients $\forall j, a_j =< \mathbf{I}, \Phi_j >$,
3. while the residual energy $E = \|\mathbf{I}\|^2$ is above a given threshold do:
   (a) select the best match: $i = \text{ArgMax}_j |a_j|$, where $|\cdot|$ denotes the modulus,
   (b) increment the sparse coefficient: $s_i = s_i + \alpha \cdot a_i$,
   (c) update residual image: $\mathbf{I} \leftarrow \mathbf{I} - \alpha \cdot a_i \cdot \Phi_i$,
   (d) update residual coefficients: $\forall j, a_j \leftarrow a_j - \alpha \cdot a_i < \Phi_i, \Phi_j >$,
4. the final non-zero values of the sparse representation vector $\mathbf{s}$, give the list of edges representing the image as the list of couples (i, s_i), where i represents an edge occurrence as represented by its position, orientation and scale.

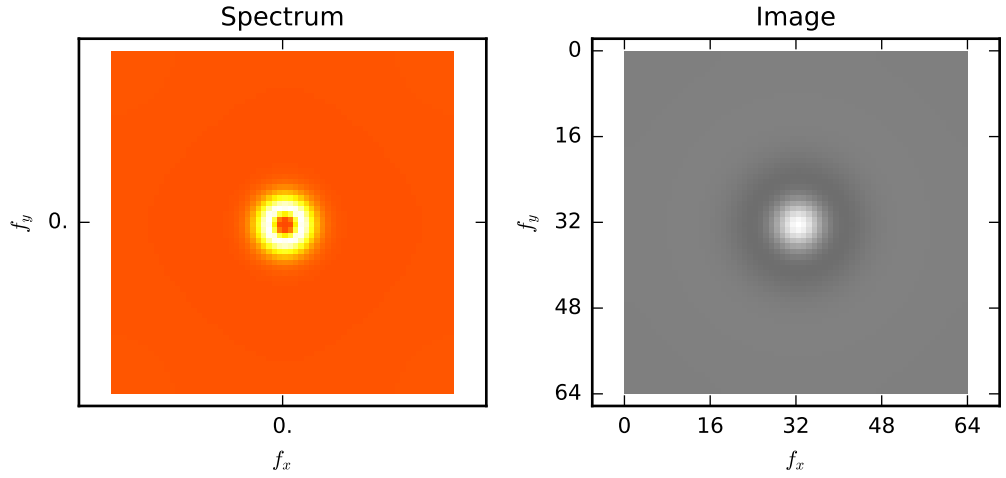

Figure 6: **Filters used in this paper.** To mimic the average profile of receptive fields in the retina, we use rotationally symmetric filters. These are characterized by a donut-shaped spectrum (Left) and resemble the Laplacian of Gaussian or the Difference of Gaussian filters in Image coordinate (Right)

This class of algorithms gives a generic and efficient representation of edges, as illustrated by the example see Figure 1. The performance of the algorithm can be measured quantitatively by reconstructing the image from the list of extracted edges. Measuring the ratio of extracted energy in the images, $N = 9192$ edges were enough to extract an average of 90% of the energy of 256×256 images on all sets of images. with packages NumPy (version 1.6.2) and SciPy (version 0.7.2) [16] on a cluster of Linux computing nodes. Visualization was performed using Matplotlib (version 1.1.0) [10]. These python scripts are available at `XXXanonymousXXX`.

6.2 Defining natural images

Our goal is to study how the statistics of contrast occurrence vary across a set of natural images. It consists of the image databases (600 images each)[2] used by Serre et al. [29], which contain either animals at different close-up views in a natural setting (which we call "animal image"), or natural images without animals, which we call "non-animal natural images".

6.3 DropLets

6.4 Designing the protocol

Designing an experiment is mainly constrained by the time available for recording and we will estimate it here (optimistically) to < 2 hours. The protocol consists in showing a set of motion clouds with different parameters. Spatial frequency will be tuned from the literature for the given eccentricity in retina, while average speed (as in the above example) is always nil. The remaining parameters are the precision in spatial frequency (B_f), the precision in time (B_V) which is inversely proportional to the average life-time of the textons, and precision for orientation ($B_\theta = \infty$, orientation θ is arbitrary).

- Presentation of stimuli at 60.0 (frame/second) on the 400x400 (pxl x pxl) array during 24.0 s
- fixed parameters:
  - $sf_0 = 0.05$ mean spatial frequency tuned for optimal neural tuning,
  - $B_sf = 0.075$ is the spatial frequency bandwidth tuned for optimal neural tuning, proportional to $sf_0 = 0.05$,
  - $B_V = 0.5$ is the temporal frequency bandwidth tuned for optimal retinal tuning,
  - $B_theta = \infty$ miplements circular symmetry, as such, θ is arbitrary

[2]Publicly available at *http://cbcl.mit.edu/software-datasets/serre/SerreOlivaPoggioPNAS07*.

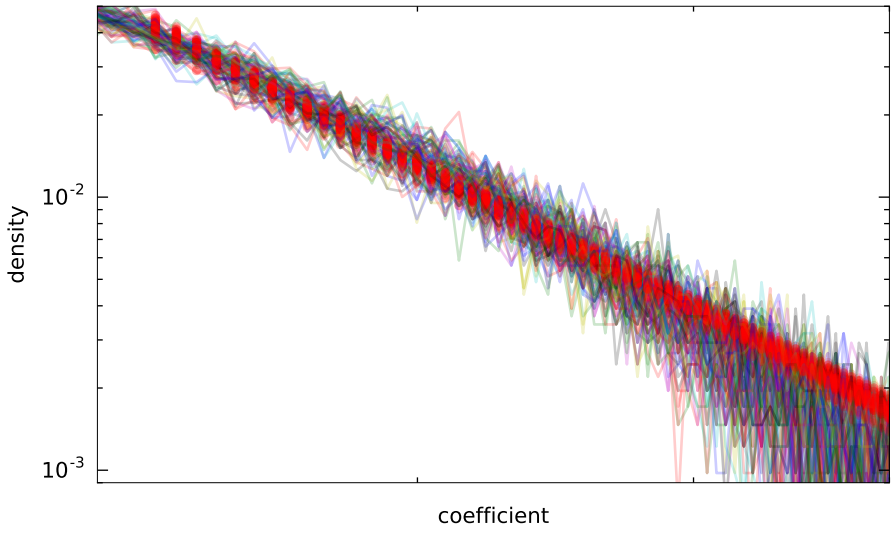

Figure 7: **Histogram of scaled sparseness coefficients.** To show the generality of the power-law description of the distribution of coefficients in natural images, we show the distribution of scaled coefficients for each image i, when knowing the estimated coefficient ρ_i. This shows that scaled coefficients align on a similarly shaped distribution.

- parameters:
  - $N_{sparse} = 6$ degrees of sparseness, distributed on a logscale of base $sparse_{base} = 200000.0$, resulting in the sparseness vector $sparseness = [5.00000000e - 065.74349177e - 056.59753955e - 047.57858283e - 038.70550563e - 021.00000000e + 00]$ - resulting in the following number of components: $[72, 827, 9500, 109131, 1253592, 14400000]$
  - $N_{trial} = 11$ different repetitions

References

[1] Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 3(2):213–52.

[2] Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial vision*, 10(4):433–436.

[3] Doi, E., Gauthier, J. L., Field, G. D., Shlens, J., Sher, A., Greschner, M., Machado, T. A., Jepson, L. H., Mathieson, K., Gunning, D. E., Litke, A. M., Paninski, L., Chichilnisky, E. J., and Simoncelli, E. P. (2012). Efficient Coding of Spatial Information in the Primate Retina. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(46):16256–64.

[4] Fischer, S., Redondo, R., Perrinet, L. U., and Cristóbal, G. (2007a). Sparse approximation of images inspired from the functional architecture of the primary visual areas. *EURASIP Journal on Advances in Signal Processing*, 2007(1):090727–122.

[5] Fischer, S., Sroubek, F., Perrinet, L. U., Redondo, R., and Cristóbal, G. (2007b). Self-invertible 2D log-Gabor wavelets. *International Journal of Computer Vision*, 75(2):231–246.

[6] Fournier, J., Monier, C., Pananceau, M., and Frégnac, Y. (2011). Adaptation of the simple or complex nature of v1 receptive fields to visual statistics. *Nat Neurosci*, 14(8):1053–60.

[7] Galerne, B., Gousseau, Y., and Morel, J. M. (2011). Micro-Texture synthesis by phase randomization. *Image Processing On Line*, 1.

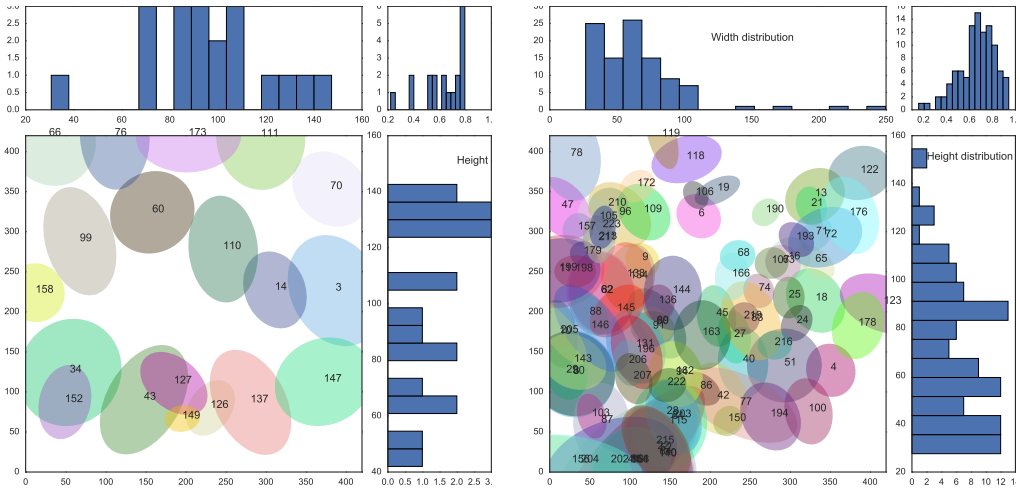

Figure 8: **Characterizing cells using the checkerboard stimulus:** Estimations of the Receptive Fields of representative retinal ganglion cells using a checkerboard stimulus. On the left we show the ON cells and on the right the OFF cells. The large plot is the map of the location of each receptive field. We show the ellipses corresponding to the best fits on the RFs' shapes. Adjacent to them are the histograms of the distribution of height and width, and in the corner is the distribution of the eccentricity of the ellipses fitted to the RF.

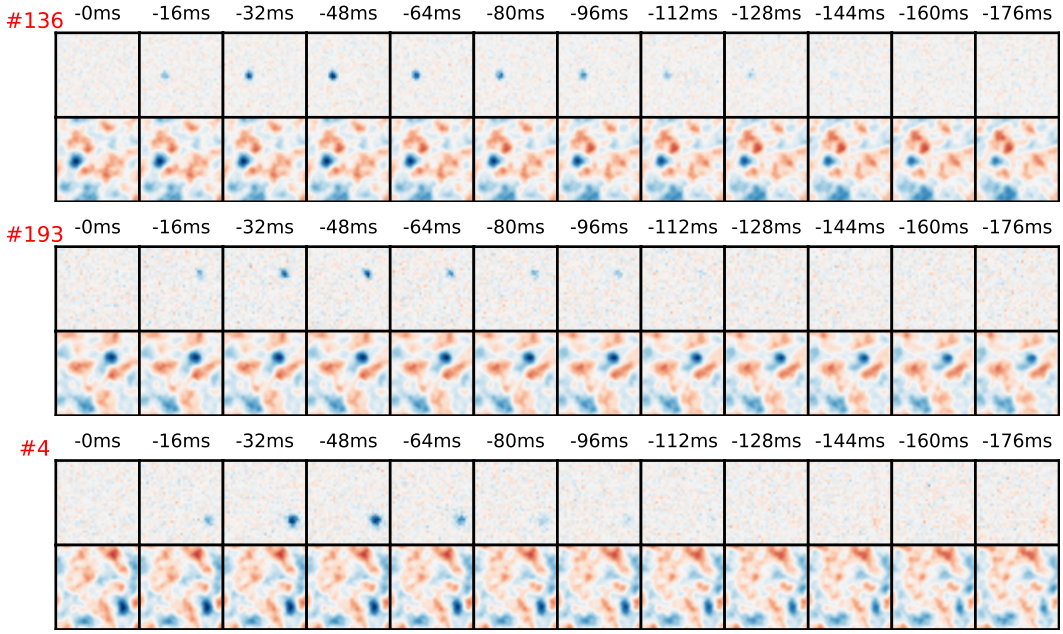

Figure 9: **Temporal profiles of the estimated Receptive Fields:** Each pair of rows represent the STA of a representative cell. Each image is the average frame preceding each spike, at the time indicated above. The upper row is the STA computed from the response to the checkerboard stimulus and the lower row the STA computed from the response to the DropLets. As can be seen, the temporal course is much shorter for the checkerboard, mainly due to the temporal correlations of the DropLets. Ideally, the parameters will be tuned to get an equivalent response from both. For many cells, our estimation of the RF is very close to the one obtained by the traditional method (albeit generally with a longer temporal profile). Notice that the time of stimulation required is much shorter.

[8] Geisler, W. S. (2008). Visual Perception and the Statistical Properties of Natural Scenes. *Annual Review of Psychology*, 59(1):167–192.

[9] Geisler, W. S., Perry, J. S., Super, B. J., and Gallogly, D. P. (2001). Edge co-occurence in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–24.

[10] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3):90–95.

[11] Kremkow, J., Perrinet, L. U., Monier, C., Alonso, J.-M., Aertsen, A., Frégnac, Y., and Masson, G. S. (2016). Push-Pull Receptive Field Organization and Synaptic Depression: Mechanisms for Reliably Encoding Naturalistic Stimuli in V1. *Frontiers in Neural Circuits*, 10.

[12] Kreuz, T., Chicharro, D., Houghton, C., Andrzejak, R. G., and Mormann, F. (2013). Monitoring spike train synchrony. *Journal of Neurophysiology*, 109(5):1457–1472.

[13] Lagae, A., Lefebvre, S., Drettakis, G., and Dutré, P. (2009). Procedural noise using sparse gabor convolution. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2009)*, 28(3):54–64.

[14] Mallat, S. and Zhang, Z. (1993). Matching Pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3414.

[15] Marre, O., Amodei, D., Deshmukh, N., Sadeghi, K., Soo, F., Holy, T. E., and Berry, M. J. (2012). Mapping a Complete Neural Population in the Retina. *Journal of Neuroscience*, 32(43):14859–14873.

[16] Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science and Engineering*, 9(3):10–20.

[17] Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325.

[18] Paninski, L. (2006). The spike-triggered average of the integrate-and-fire cell driven by gaussian white noise. *Neural computation*, 18(11):2592–616.

[19] Perrinet, L., Samuelides, M., and Thorpe, S. (2004a). Coding static natural images using spiking event times: Do neurons cooperate? *IEEE Transactions on Neural Networks*, 15(5):1164–1175. Special issue on 'Temporal Coding for Neural Information Processing'.

[20] Perrinet, L., Samuelides, M., and Thorpe, S. (2004b). Sparse spike coding in an asynchronous feed-forward multi-layer neural network using matching pursuit. *Neurocomputing*, 57:125–134. Special issue: New Aspects in Neurocomputing: 10th European Symposium on Artificial Neural Networks 2002 - Edited by T. Villmann.

[21] Perrinet, L. U. (2010). Role of homeostasis in learning sparse representations. *Neural Computation*, 22(7):1812–36.

[22] Perrinet, L. U. (2015). Sparse models for computer vision. In Cristóbal, G., Perrinet, L., and Keil, M. S., editors, *Biologically Inspired Computer Vision*, chapter 13. Wiley-VCH Verlag GmbH & Co. KGaA.

[23] Perrinet, L. U. (2016). Biologically-inspired characterization of sparseness in natural images. *IEEE Xplore*.

[24] Pitkow, X. and Meister, M. (2012). Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, 15(4):628–635.

[25] Rajan, K. and Bialek, W. (2013). Maximally informative "stimulus energies" in the analysis of neural responses to natural signals. *PLoS ONE*, 8(11):1–16.

[26] Rikhye, R. V. and Sur, M. (2015). Spatial Correlations in Natural Scenes Modulate Response Reliability in Mouse Visual Cortex. *Journal of Neuroscience*, 35(43):14661–14680.

[27] Sallee, P. and Olshausen, B. A. (2003). Learning Sparse Multiscale Image Representations. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in neural information processing systems*, volume 15, pages 1327–34. The MIT Press, Cambridge, MA.

[28] Sanz-Leon, P., Vanzetta, I., Masson, G. S., and Perrinet, L. U. (2012). Motion clouds: model-based stimulus synthesis of natural-like random textures for the study of motion perception. *Journal of Neurophysiology*, 107(11):3217–3226.

[29] Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429.

[30] Sharpee, T., Rust, N. C., and Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural computation*, 16(2):223–250.

[31] Simoncelli, E. P., Paninski, L., Pillow, J., and Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. *Cognitive Neurosciences Iii, Third Edition*, pages 327–338.

[32] Touryan, J. (2001). Analysis of sensory coding with complex stimuli. *Current Opinion in Neurobiology*, 11(4):443–448.

[33] Vacher, J., Meso, A. I., Perrinet, L. U., and Peyré, G. (2015). Biologically inspired dynamic textures for probing motion perception. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1918–1926. Curran Associates, Inc.

[34] Xia, G. S., Ferradans, S., Peyré, G., and Aujol, J. F. (2014). Synthesizing and mixing stationary gaussian texture models. *SIAM Journal on Imaging Sciences*, 7(1):476–508.