nature portfolio

Peer Review File

Recurrent models of orientation selectivity enable robust early-vision processing in mixed-signal neuromorphic hardware

Corresponding Author: Professor Silvio Sabatini

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The manuscript presents a recurrent neural network (RNN) that can emulate the retinocortical visual pathway and produce Gabor-like receptive fields. The fields are tuned to visual stimuli with specific orientation and spatial frequency components. The hardware system developed to experimentally demonstrate the concept comprised of a Dynamic Vision Sensor (DVS) and a Dynamic Neuromorphic Asynchronous Processor (DyNAP).

The fundamental contribution is this work is the usage of recurrent inhibition to realize visual receptive fields in spiking neural networks. The recurrent inhibitory connections realize the Gabor-like selective spatial receptive fields with fewer neural resources than the feedforward connections. Also, the receptive fields are sharper when using RNNs.
The proposed scheme can help improve the utilization efficiency of mixed-signal neuromorphic hardware for visual tasks, by significantly reducing the number of interconnections.

The work is supported by simulation as well as hardware-based experiments. The methodology and provided details look sound.
I have the following comments/questions for the authors:
Can authors provide a schematic of the neural network showing the recurrent connections, versus the feedforward connections for realizing the receptive field? It's hard to visualize the network from the text and Figure 1.

In Table 1, a comparison between the recursive and feedforward scheme is shown in terms of the number of interconnections for a single neuron. The reduction is about 2X when considering the kernel with 5 subregions. Are there any other metrics of comparison, such as energy consumption, DyNAP chip area used, accuracy when used with a classification SNN layer, etc, which can demonstrate substantial improvement at the network level?

Reviewer #2

(Remarks to the Author)
The paper is focused on exploring hardware and algorithms for spike-based implementation of early vision signal analyzers like 2D Gabor filters. The overall content of the paper can be broadly divided into two parts:

(A) The paper describes how adding inhibitory recurrent pathways among lateral V1 cells, in addition to excitatory feedforward pathways between retina and V1 cells, faithfully emulates 2D Gabor-like filtering with the reduced number of total synaptic connections compared to the case where Gabor-like functionality is emulated using only the excitatory feedforward connections. Through network simulations, the authors show that V1 neurons have a firing rate curve that peaks at only a certain value of “orientation” and “spatial frequency” in the input, for a properly chosen set of parameters that define the Gabor-like functionality. The authors can tune the phase of the Gabor filter by performing appropriate linear superpositions of the receptive fields of neighboring V1 neurons. The authors can also tune the parameters to create banks

of v1 neurons receptive to specific orientation and radial frequency values in the input.

(B) More importantly, the authors show that using the effective firing rate of the retina (ON firing rate minus OFF firing rate) instead of both as a whole or separately, can lead to qualitatively steeper tuning curves of the V1 neurons, supposedly resulting in better selectivity to orientation & spatial frequency.

Part (B) of the paper is interesting. However, I have several qualms and reservations about the overall work presented in this paper:

1. A major chunk of the material in this paper (classified as part A in my summary) that spans everything in the Results section except for the last sub-section and all main figures except for Fig. 6 & 7 and parts of Fig. 5, have already been presented, discussed and proposed in other publications of the authors [1-2]. However, the authors present as if these are new results pertaining to this paper. For example, Fig. 1 of submission is taken from Fig. 1 of [2], Fig. 2 is the same as Fig. 1 of [1], Fig. 3 is the same as Fig. 2 of [1], Fig. 4 is the same as Fig. 4 of [2] and Fig. 5c is like Fig. 8 of [2]. The abstract, introduction, and discussion boast of already published work as contributions to this paper. Therefore, the only new result presented in this work is what is classified as part B in my summary.

2. Overall, the paper is poorly written and difficult to read. The contributions of the work in the introduction are very vaguely written, possibly due to the nominal amount of new work put forth.

3. The overall content of the paper is highly technical and specific, and in my opinion, its implications/contributions do not appeal to the general interdisciplinary audience of this journal. It is more suited to a more specific journal or conference like ISCAS, ICCV, ICONS, etc, that focuses on and welcomes such highly technical content.

4. Since the content is so highly technical, each of the topics discussed in the results section could benefit significantly from a better writing flow, and detailed explanations either in the main text or the supplementary section.

5. The authors do not explain why both ON & OFF event firing rates are important. One can appreciate the importance of a complete set of oriented filters where each filter gathers information about the signal's phase with reference to the filter's orientation and, therefore, the usefulness of equations 5-7. But it is unclear how equation 8, the key "push-pull" concept proposed by the authors, is helping/changing things for the better.

6. Although the authors highlight through Fig. 7d that the response rate becomes steeper with the inclusion of the "push-pull" technique, it is unclear how that is happening by just reading the mathematical treatment presented.

7. A graph showing how the synaptic sparsity advantage of having an inhibitory connection would scale with respect to strictly feedforward connection-based Gabor filters for different filter sizes and # of sub-regions would be helpful.

8. A description of the grating experiment in the supplementary would be good, clearly pointing out how the spatial frequency is defined and what the sinusoidal grating inputs looked like. For example, in the text corresponding to Fig. 6, the authors say they are varying the phase of the sinusoidal grating, but how exactly is that done in the input?

9. Statements like "In this paper, we have demonstrated that recurrent clustered inhibition can be successfully used in SNNs, both in simulation and on mixed-signal analog/digital neuromorphic hardware, to economically implement highly structured visual receptive fields." from the discussion section, would be more appropriate to omit as this has already been done before.

10. Another statement: "The proposed solutions meet the requirements posed by an effective cortical-like representation in terms of channel redundancy (i.e., a large number of orientations and spatial frequencies), and in terms of information efficiency (i.e., highly structured basis functions)." in the discussion section is vague and unsubstantiated. Where have the requirements been laid down? The concept of having multiple banks of Gabor filters with different parameters has already been laid down in several studies in literature, both in the non-spiking domain and in the spiking domain, many by the authors themselves. So, this so-called requirement is not met for the first time in this paper.

11. The following statement in discussion: "We verified that the linearity assumption still holds despite the high non-linearity of spiking neurons. Additionally, we showed how it is possible to combine such feature detectors to generate Gabor-like filters with arbitrary phase values, effectively implementing a full harmonic representation of the image signal." This is again misleading for the purposes of this paper, as this has already been shown in previous papers [1-2]. Similarly, the last statement of the discussion section should also be omitted: "This would pave the way to the implementation of complex bio-inspired networks for more demanding online visual tasks on neuromorphic hardware."

12. In the phrase "but inescapably discard part of the signal." mentioned in the introduction, it is unclear as to which part of the signal the authors are referring to.

13. Please add a proper reference to back up the following statement: "To assess network's performance and characterize the receptive fields of its output neurons, we used two dimensional (2D) sinusoidal drifting gratings as visual stimuli, widely used to investigate the response of cells in the primary visual cortex".

14. Fig. 3b should show the different values of the parameters d, sigma, and b used to obtain the tuning curves. The caption

provides the range in which they are varied but not the steps used.

15. The following passage from the sub-section labeled “Linearity test and feature tuning characterization” has redundant statements and can be shortened: “The best results, i.e., the narrowest tuning curves, are obtained when the size of the recurrent inhibitory clusters and their distance from the target neuron are both comparable to the width of the feed-forward excitatory kernel. More precisely, considering fixed parameters for the feed-forward kernel (in particular $\sigma_h = 3.5$, $p = 1/3$, and multiplicative weight of the feed-forward excitatory connections $a = 103$), the parameters that influence the effect of the recurrent clustered inhibition are the distance d between the target neuron and each of the inhibitory clusters, the spatial extension of the clusters and the multiplicative weight of the recurrent inhibitory connections b.”

16. For lines 170 to 179, the authors should add simulation results to support this statement or refer to work that has demonstrated this before.

17. The supplementary figure reference in line 201 is incorrect. It should be S2.

18. The phrase “substantially lower number of” in line 326 is qualitative and has to be quantified. From the numbers mentioned, I would not call it "substantial".

[1] Baruzzi, Valentina, Giacomo Indiveri, and Silvio P. Sabatini. "Emergence of Gabor-Like Receptive Fields in a Recurrent Network of Mixed-Signal Silicon Neurons." 2020 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2020.

[2] Baruzzi, Valentina, Giacomo Indiveri, and Silvio P. Sabatini. "Compact Early Vision Signal Analyzers in Neuromorphic Technology." Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020). Vol. 4. SciTePress, 2020.

Reviewer #3

(Remarks to the Author)
-In this paper, the author proposed a recurrent snn model for describing a retinocortical layer mathematically. In addition, the author implemented the model by a neuromorphic hardware system comprising a Dynamic Vision Sensor that emulates the transient pathway of real retinas and a mixed-signal Dynamic Neuromorphic Asynchronous Processor with neurons and synapses.
-I have the following concerns for the technical quality of the submitted manuscript.
-The recurrent snn model seems that it comes from the previous publication largely, as the author mentioned in the manuscript. If so, the author needs to make it clear what novelty the submitted manuscript can show in terms of model equations, etc.
-The main contribution of the paper is thought to be the neuromorphic hardware implementation that enables an economic realization of the retinocortical layer by combining the DVS camera and neuromorphic asynchronous processor with the emulation results. If so, the author needs to make it clear why the hardware implementation is more economic than the others, in addition to comparing the number of connections shown in Table 1.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Thanks to the authors’ detailed response to my previous comment regarding its ambiguity, the push-pull technique now makes sense. However, its impact on potential vision processing pipelines is still not clear. For example, the push-pull waveforms in Fig. 7 of the main text seem to be in the center of the only-ON and only-OFF waveforms, but their difference does not seem to be significant. How much improvement in the overall classification accuracy of a full-fledged vision processor based on these Gabor filters will be due to such (arguably) minor improvements in phase errors brought about by the push-pull technique?

Also, the authors provide information on the synaptic sparsity advantage of their technique but how much silicon area and energy (based on system level analysis) will be saved by using the sparse recurrent inhibition incorporated Gabor filter implementations compared to ones with only feedforward connections. How will the performance and hardware cost of such pipelines compare with other spike-based vision processing pipelines already studied in the literature?

Reviewer #3

(Remarks to the Author)
-The author reflected the review comments well in the revised manuscript.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I don't have further comments for the authors

Manuscript title: Recurrent models of orientation selectivity enable robust early-vision processing in mixed-signal neuromorphic hardware

Manuscript ID: NCOMMS-23-47095

Authors: Valentina Baruzzi, Giacomo Indiveri and Silvio P Sabatini

Journal: Nature Communications

Content type: Article

Submitted on October 10th, 2023, Decision received on November 15th, 2023

We would like to thank the Reviewers for the interesting comments, which helped us to improve the validity of the results, as well as the quality of the presentation. Reviewer's comments are reported in RED, our responses in BLUE and in PURPLE we report here the changed text. The same text is also highlighted in PURPLE in the manuscript.

Reviewer #1

The manuscript presents a recurrent neural network (RNN) that can emulate the retinocortical visual pathway and produce Gabor-like receptive fields. The fields are tuned to visual stimuli with specific orientation and spatial frequency components. The hardware system developed to experimentally demonstrate the concept comprised of a Dynamic Vision Sensor (DVS) and a Dynamic Neuromorphic Asynchronous Processor (DyNAP).

The fundamental contribution is this work is the usage of recurrent inhibition to realize visual receptive fields in spiking neural networks. The recurrent inhibitory connections realize the Gabor-like selective spatial receptive fields with fewer neural resources than the feedforward connections. Also, the receptive fields are sharper when using RNNs.

The proposed scheme can help improve the utilization efficiency of mixed-signal neuromorphic hardware for visual tasks, by significantly reducing the number of interconnections.

The work is supported by simulation as well as hardware-based experiments. The methodology and provided details look sound.

Remarks to the Authors

1 - Can authors provide a schematic of the neural network showing the recurrent connections, versus the feedforward connections for realizing the receptive field? It's hard to visualize the network from the text and Figure 1.

We agree with the reviewer. Accordingly, we modified the original Figure 2 of the manuscript by including top-view schemes of the feed-forward and the recurrent kernels for the different parameters' settings considered in the work (see Figure R1 here below).

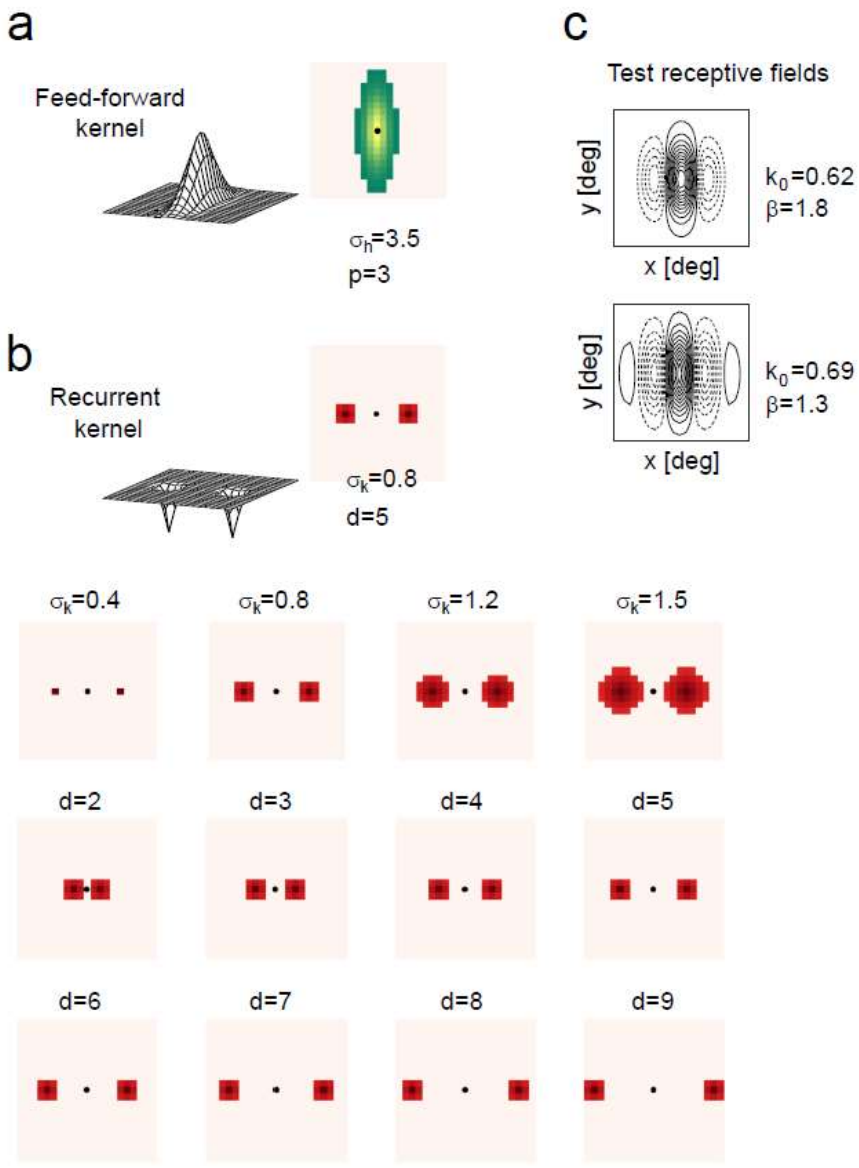

Figure R1

Furthermore, we have modified the original Figure 1 of the manuscript by including two new panels (B and C) that pictorially show the interconnections of the recurrent scheme and those for the equivalent feedforward one (see Figure R2 here below). The feedforward resolvent kernel of the recurrent integral equation (see Eq.1 in the manuscript) represents how total afferent drive at retina site affects activity at a cortical (V1) site, detecting specific characteristics present in the input pattern of excitation. By properly choosing the parameters of the kernel of recurrent inhibition, the spatial extension on which these characteristics are detected is possibly larger than that of the actual inhibitory connections. This occurs both directly, by physical local interactions, and indirectly, through propagation property of recursion. In this way, one can speak of “induced” functional couplings not directly related to the presence of corresponding specific wirings.

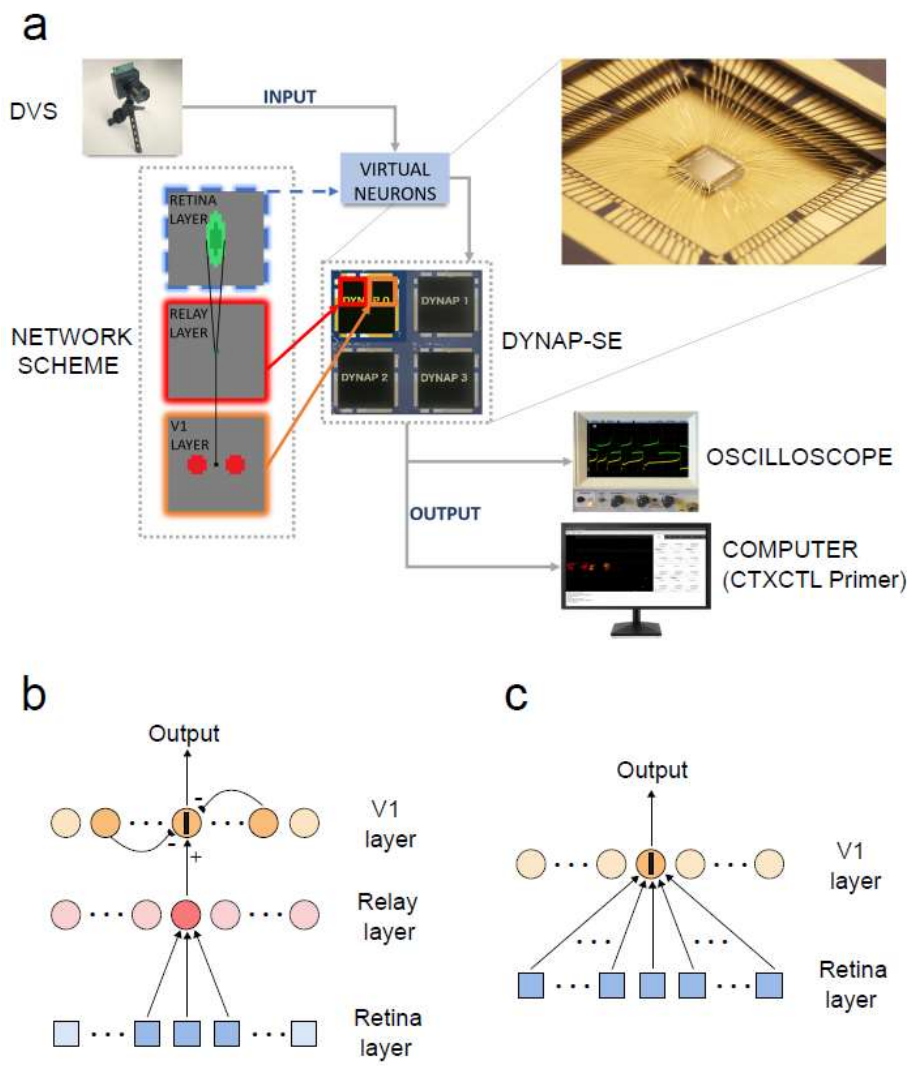

Figure R2

2 - In Table 1, a comparison between the recursive and feedforward scheme is shown in terms of the number of interconnections for a single neuron. The reduction is about 2x when considering the kernel with 5 subregions. Are there any other metrics of comparison, such as energy consumption, DyNAP chip area used, accuracy when used with a classification SNN layer, etc, which can demonstrate substantial improvement at the network level?

Actually, with a proper choice of parameters, the RNN allows us to obtain Gabor-like receptive fields characterized by five subregions (i.e., corresponding to a relative spatial frequency bandwidth $\beta=0.8\div0.9$ octaves, calculated for a half-power cut-off frequency). A direct comparison with equivalent receptive fields obtained by a strictly feed-forward scheme shows an advantage for RNN over the former by a factor of 3.13 in terms of number of interconnections. In case we settle for receptive fields with a larger relative bandwidth (e.g., $\beta\sim1.5$ octaves, and thus a less number of subregions, e.g., three), the higher efficiency of the RNN over the FF one could drop to a factor of 1.65, yet paying the price of (accepting) a

reduced selectivity in the spatial frequency domain. Considering that typical vision applications require front-end convolutions with a huge number of receptive fields, the advantage/convenience of RNN-based solutions turns out to be so far substantial.

We have included these considerations in the manuscript.

Furthermore, following the reviewers' suggestion, we have included in the manuscript a discussion of the power consumption of the recurrent network, providing an estimate for the 21x21 hardware implementation. The average estimated power consumption for a neuron of the relay layer and for the corresponding neuron of the V1 layer, when responding to the preferred stimulus at the highest temporal frequency, is 6.54 μJ and 446 nJ, respectively. (Details on the estimation are reported in section 'Methods').

However, a comparison with equivalent feed-forward networks was not possible, as, actually, they have not been implemented on the DyNAP, and as the average number of spikes estimated by simulations significantly differ from those actually observed in hardware.

Concerning the chip area used, the reduction of area for the recurrent network implementation with respect to the equivalent feed-forward one, is still of a factor of 3, when considering a five subregion kernel, along with the reduction of the required interconnections.

Manuscript title: Recurrent models of orientation selectivity enable robust early-vision processing in mixed-signal neuromorphic hardware

Manuscript ID: NCOMMS-23-47095

Authors: Valentina Baruzzi, Giacomo Indiveri and Silvio P Sabatini

Journal: Nature Communications

Content type: Article

Submitted on October 10th, 2023, Decision received on November 15th, 2023

We would like to thank the Reviewers for the interesting comments, which helped us to improve the validity of the results, as well as the quality of the presentation. Reviewer's comments are reported in RED, our responses in BLUE and in PURPLE we report here the changed text. The same text is also highlighted in PURPLE in the manuscript.

Reviewer #2

The paper is focused on exploring hardware and algorithms for spike-based implementation of early vision signal analyzers like 2D Gabor filters. The overall content of the paper can be broadly divided into two parts:

(A) The paper describes how adding inhibitory recurrent pathways among lateral V1 cells, in addition to excitatory feedforward pathways between retina and V1 cells, faithfully emulates 2D Gabor-like filtering with the reduced number of total synaptic connections compared to the case where Gabor-like functionality is emulated using only the excitatory feedforward connections. Through network simulations, the authors show that V1 neurons have a firing rate curve that peaks at only a certain value of "orientation" and "spatial frequency" in the input, for a properly chosen set of parameters that define the Gabor-like functionality. The authors can tune the phase of the Gabor filter by performing appropriate linear superpositions of the receptive fields of neighboring V1 neurons. The authors can also tune the parameters to create banks of v1 neurons receptive to specific orientation and radial frequency values in the input.

(B) More importantly, the authors show that using the effective firing rate of the retina (ON firing rate minus OFF firing rate) instead of both as a whole or separately, can lead to qualitatively steeper tuning curves of the V1 neurons, supposedly resulting in better selectivity to orientation & spatial frequency.

Part (B) of the paper is interesting. However, I have several qualms and reservations about the overall work presented in this paper:

We thank the reviewer for having acknowledged and pointed out the novelty of Part (B). As detailed in the following, we have revised the manuscript to better analyze the results obtained by combining ON and OFF channels with respect to those for each single channel, presented in our previous preliminary works.

Remarks to the Authors

1 - A major chunk of the material in this paper (classified as part A in my summary) that spans everything in the Results section except for the last sub-section and all main figures except for Fig. 6 & 7 and parts of Fig. 5, have already been presented, discussed and proposed in other publications of the authors [1-2]. However, the authors present as if these are new results pertaining to this paper. For example, Fig. 1 of submission is taken from Fig. 1 of [2], Fig. 2 is the same as Fig. 1 of [1], Fig. 3 is the same as Fig. 2 of [1], Fig. 4 is the same as Fig. 4 of [2] and Fig. 5c is like Fig. 8 of [2]. The abstract, introduction, and discussion boast of already published work as contributions to this paper. Therefore, the only new result presented in this work is what is classified as part B in my summary.

We agree with the reviewer that *part* of the material presented in this manuscript was already presented in the two cited conference proceedings [1] and [2], which indeed we have disclosed. We believe that the manuscript has significant added value, because it integrates in an organic and complete way the body of all the results, only partially covered in previous conference proceedings, and integrates them with additional novel contributions (i.e., the part B in the reviewer's summary). Irrespective of our opinion, to address this issue we have substantially extended the new parts, and adapted the technical parts to better target the audience of Nature Communication, taking into account the valuable comment #3 of the reviewer.
Accordingly, new sections have been added, technical details have been explained better, old figures 1,2,3,6 have been revised and enriched with new panels to be more clear and explicative, and finally a new figure 4 has been introduced to substitute Table 1.

2 - Overall, the paper is poorly written and difficult to read. The contributions of the work in the introduction are very vaguely written, possibly due to the nominal amount of new work put forth.

We thank the reviewer for pointing out these issues. We did a major revision of the text throughout, to improve the clarity of the approach and better highlight the contributions of the work. Specifically in the 'Introduction', the amount of new work has been underlined and valued (see also our reply to point 1).]

3-4 - The overall content of the paper is highly technical and specific, and in my opinion, its implications/contributions do not appeal to the general interdisciplinary audience of this journal. It is more suited to a more specific journal or conference like ISCAS, ICCV, ICONS, etc, that focuses on and welcomes such highly technical content. Since the content is so highly technical, each of the topics discussed in the results section could benefit significantly from a better writing flow, and detailed explanations either in the main text or the supplementary section.

We thank the reviewer for having highlighted this point. We agree. Hence, we have revised the text throughout to make it more appealing to a general interdisciplinary audience (see also our responses to previous points).

In particular, the new parts (highlighted in purple in the manuscript) make the revised manuscript much more different from the preceding conference papers (ISCAS and VISAPP).

5 - The authors do not explain why both ON & OFF event firing rates are important. One can appreciate the importance of a complete set of oriented filters where each filter gathers information about the signal's phase with reference to the filter's orientation and, therefore, the usefulness of equations 5-7.

But it is unclear how equation 8, the key "push-pull" concept proposed by the authors, is helping/changing things for the better.

We agree with the reviewer, indeed we devoted the original Figure 7d only to support this claim.

To answer to this remark, and, in particular, to better clarify the role of equation 8, and thus the necessity of a push-pull mechanism, we have to consider that linearity is an essential property to define the phase of a signal. In order to demonstrate that our model V1 cells are capable of extracting phase of local contrast, we have first to demonstrate how we can gain linear weighting of signed contrast over the cells' receptive fields, despite the high non-linearities of their activation functions.

To this end, at the beginning of subsection "Extraction of full harmonic content through …" we added a detailed analysis on how a push-pull mechanism of ON and OFF channels determines the neuron's linear response to stimulus contrast.

We report the new text here below in purple.

In general, input-output characterization of visual RFs is based on the notion of contrast. Accordingly, we can represent the spatial image (i) as the combination of two components: one part is the average luminance of the stimulus (m), the second part is the variation of luminance about the mean, which defines the stimulus contrast (c):

$$i=(1+c)m$$

where c can be either positive or negative, and $m \geq 0$.

In early stages of the visual system, for each contrast polarity channel, local changes of contrast in a cell's receptive field yield to changes of that cell rate of response (r) :

$$\Delta r^{ON} = r^{ON} - r_0 \qquad \Delta r^{OFF} = r^{OFF} - r_0,$$

where r_0 is the neuron's spontaneous firing rate that we can assume equal for both ON and OFF channels. In order to gain equivalent a linear summation response to a signed contrast pattern within the overall neuron's receptive field (composed of ON and OFF subregions), a push-pull mechanism is usually advocated [Tolhurst & Dean, 1990; Hirsch & Martinez, 2006; Jo et al., 2023], that collects positive (i.e., excitatory) contribution from relay cells of preferred polarity and negative (i.e., inhibitory) contribution from relay cells of opposite polarity. ON and OFF event detectors in the retina-like DVS camera cannot *per se* encode negative responses. Yet, assuming a push-pull configuration, events provided by DVS camera can be conceptually combined to obtain positive or negative changes of response on the basis of the sign of contrast. As a result, stimulating an ON neuron by a not appropriate contrast polarity results in a decrease of its response, due to inhibition from the corresponding OFF neuron, which, conversely, has received the appropriate stimulus in its RF:

$$- \Delta r^{ON} \stackrel{\rm def}{=} \Delta r^{OFF} .$$

In other words, we take the excitatory response of the OFF channel as the estimate of the inhibitory response of the ON channel, and the combined response can be written as:

$$\Delta r = \Delta r^{ON} - \Delta r^{OFF} = r^{ON} - r^{OFF} .$$

To prove the efficacy of the push-pull mechanism we can test the superposition property. Suppose we have two contrast stimuli $c_1>0$ and $c_2<0$. The response variations of the ON and OFF channels will be:

\begin{equation}

$$\Delta r^{ON} = h_1 c_1^{+} + h_2 c_2^{+} = h_1 c_1 + 0 = h_1 c_1$$

$$\Delta r^{OFF} = h_1 c_1^{-} + h_2 c_2^{-} = 0 - h_2 c_2 = h_2 |c_2|$$

respectively, where $c^{+} = \max\{0,c\}$ and $c^{-} = - \min\{0,c\}$ and h_1, h_2 denote the values of the receptive field profile. By combining the two responses we obtain:

$$\Delta r = \Delta r^{ON} - \Delta r^{OFF} = h_1 c_1 - h_2 |c_2| = h_1 c_1 + h_2 c_2$$

which proves the linearity of the response, provided we model as negative the weights of the OFF subregions of the receptive field. In this way, the receptive field properly acts as a linear filter by mapping a weighted sum of the signed input contrast of the stimulus to the neural response.

In addition, we have conducted a comparative analysis on the accuracy and reliability of the phase obtained with the push-pull mechanism and without it (i.e., for the ON-only and OFF-only conditions).

Figures R3 and R4 here below show the results of such a comparison for the three stimulus gratings considered in Fig.6 of the original manuscript. For all the simulated conditions, the push-pull combination of the ON and OFF channels does not show any bias, thus resulting in a better estimate.

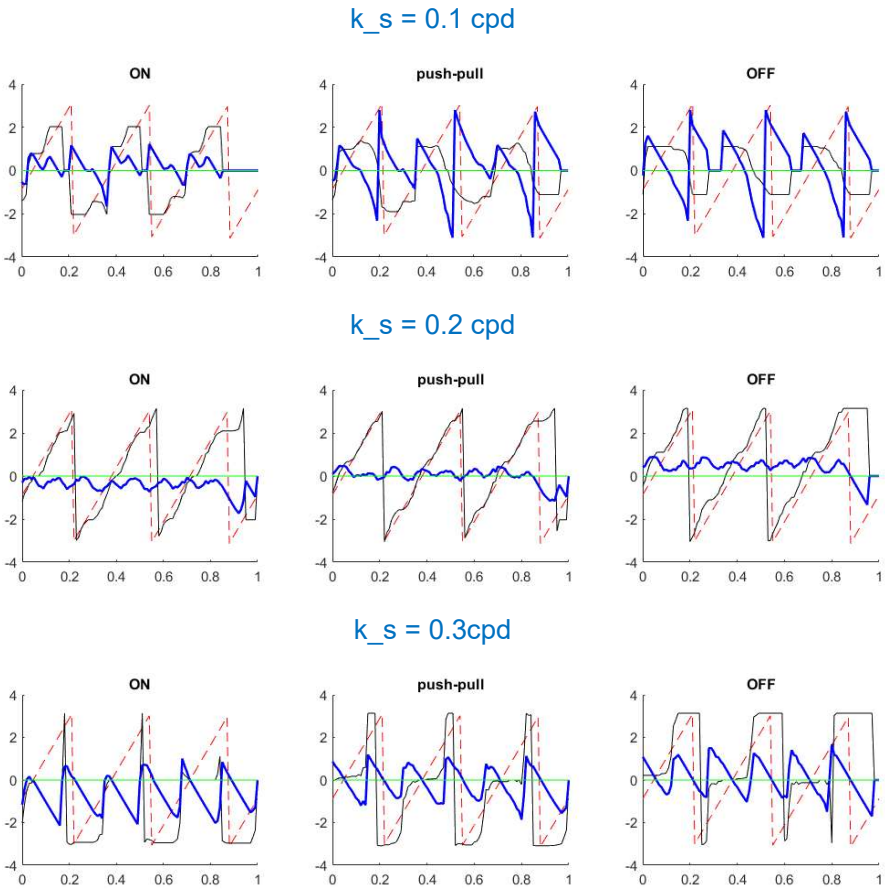

Figure R3: Comparative results of the phase estimate for different spatial frequencies of the stimulus grating and for single ON- and OFF- channels (left and right columns) and their push-pull combination (central column). In general, the estimate obtained by the push-pull configuration is more accurate than the corresponding estimates when a single channel is used. Red dashed lines represent the actual phase of the stimulus as a function of time, black thin lines represent the estimated phase, thick blue lines represent the phase errors.

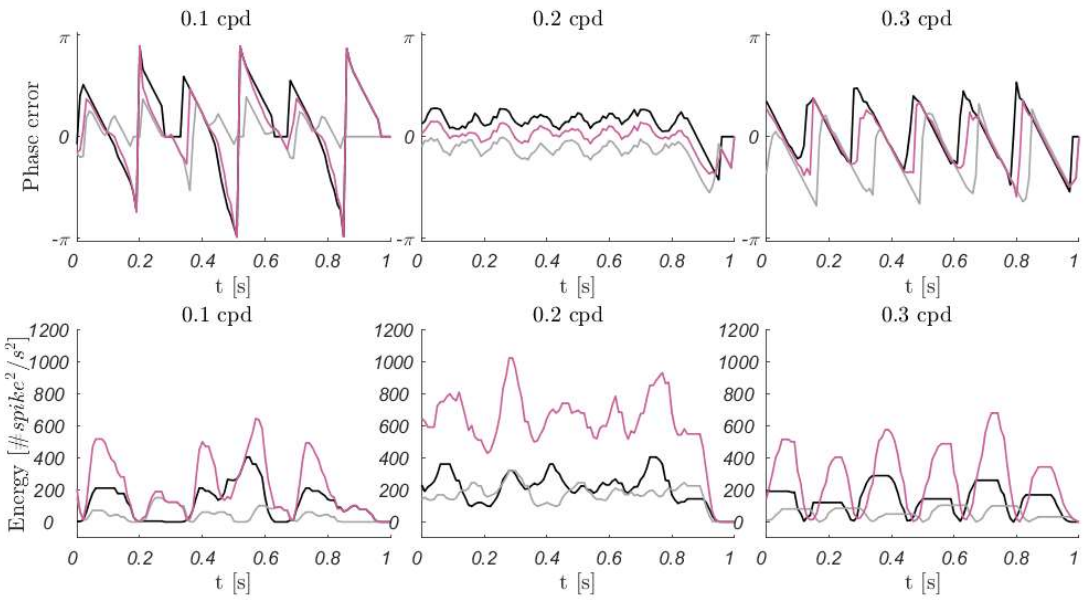

Figure R4: Comparative results of the error of phase estimation, and its reliability as a function of time, for different spatial frequencies of the stimulus grating (k_s=0.1, 0.2, and 0.3 cpd), and for single ON- (light gray), OFF- channel (black), and their push-pull combination (purple).

In the revised manuscript, we have included a graph with the comparative results only for the case in which the stimulus' spatial frequency matched the peak frequency of the Gabor-like receptive field (i.e., k_s = 0.2cpd ≃ k_0), see new panel (d) of the new Fig.7. For the sake of clarity, the actual (i.e., ground truth) phase signal was also displayed in the plots as reference.

In relation to that, in the 'Methods' section, we have included the methodology used for such a comparison:

In order to analyze the advantages of the push-pull combination of ON and OFF channels, we computed the capacity of the recurrent network to provide an effective estimate (\phi_est) of the local phase of the input stimulus (\phi_act) in terms of accuracy and reliability. To this end, the phase error \Delta \phi(n,t) = \phi_est(n,t) – \phi_s(n,t) was directly computed in the complex plane by using the following identity:

\Delta \phi(n,t) = atan2(C_s(n,t)*S(n,t) - C(n,t)*S_s(n,t), C(n,t)*C_s(n,t) + S(n,t)*S_s(n,t)), where C(n,t) and S(n,t) are the responses of a quadrature pair of neurons with Gabor-like receptive fields centered in a fixed spatial position (for the sake of convenience, to minimize the border effect, we considered the center of the layer, i.e., n=0), whereas C_s(n,t) e S_s(n,t) are the actual quadrature components of the stimulus drifting grating s(n,t) characterized by a spatial frequency k_s:

$$s(n,t) = \sin(k_s n + \phi_s(n,t)) = \sin(k_s n)\cos(\phi_s(n,t)) + \cos(k_s n)\sin(\phi_s(n,t)) = C_s(n,t) + j\, S_s(n,t).$$

In this way, since the four-quadrant inverse tangent atan2 function returns values in the closed interval [–π, π], we avoided the attendant problem of phase unwrapping of the angle difference. The reliability of the phase estimate was obtained by the associated response energy associated response energy/amplitude C^2(n,t) + S^2(n,t).

In the revised manuscript, we have also included a graph that shows, as violin plots, the distributions around the mean of the estimation errors of the phase of the drifting grating, for the case in which the stimulus' spatial frequency matched the peak frequency of the Gabor-like receptive field (i.e., k_s = 0.2cpd ≃ k_0), see Figure R5 here below and the new panel (d) of the new Fig.7.

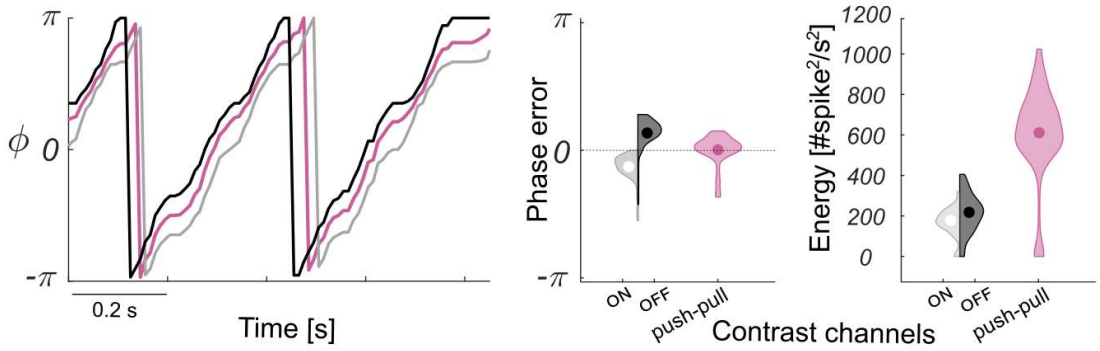

Figure R5: (Left) Comparison of the stimulus phase estimates for the ON-channel only (gray), the OFF-channel only (black), and their push-pull combination (purple). (Middle, Right) The distributions of the phase errors and phase estimate reliability for the three conditions considered. The minor bias in the error and the higher energy make the phase estimate by the push-pull configuration more accurate and reliable than those attainable by single channels.

6 - Although the authors highlight through Fig. 7d that the response rate becomes steeper with the inclusion of the "push-pull" technique, it is unclear how that is happening by just reading the mathematical treatment presented.

We have included a thorough mathematical treatment for justifying the linearity gained by the push-pull mechanism (see our response to previous point).

If c(x) is a sinusoidal contrast pattern, $c^+(x)$ and $c^-(x)$ are its complementary half-wave rectified patterns. By combining in push-pull the r^{ON} and r^{OFF} responses, and by modeling as negative the weights of OFF subregions of the RF h(x), we can gain a linear response to the full (signed) contrast pattern c(x).

7 - A graph showing how the synaptic sparsity advantage of having an inhibitory connection would scale with respect to strictly feedforward connection-based Gabor filters for different filter sizes and # of sub-regions would be helpful.

Following the suggestion of the Reviewer, we have substituted Table 1 with the new Fig.4 of the revised manuscript (see also Figure R6 here below) that comparatively shows the number of interconnections required for the recursive and feedforward scheme, as the receptive field is scaled from 13x13 to 61x61 pixels, which also directly relate to the DyNAP chip area used.

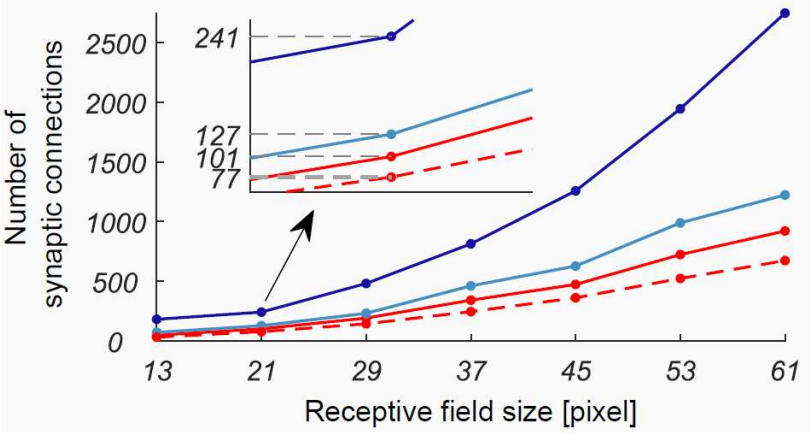

Figure R6: Comparison between recursive and feed-forward scheme in terms of required interconnections.

The different curves show the synaptic sparsity advantage of the recurrent implementation of Gabor-like receptive fields (red curves) over strictly feed-forward ones (blue curves), and how it scales for different sizes and number of sub-regions. Dashed and solid red lines represent the number of total interconnections used for a recurrent implementation of a five sub-region receptive field when the width of the inhibitory cluster (\sigma_k) was equal to 0.8 and 1.2 deg, respectively, corresponding to a relative spatial frequency bandwidth β=0.8÷0.9 octaves; the size of the feed-forward kernel (\sigma_h) was kept fixed to 3.5 deg. Light and dark blue lines represent the number of interconnections required by equivalent strictly feed-forward receptive fields of three and five sub-regions, respectively. The inset details the numerical comparison for a five pixel size of the central sub-region. Rescaling was done by maintaining the same proportions among kernels and by flooring to the greatest odd integers for obtaining the resulting sizes in pixels.

The advantage of the RNN over strictly feed-forward schemes is up to more than 3x for a five sub-region receptive field with a size of 21x21, and progressively increases with the rescaling of the filter's size. When the clusterization of the inhibitory kernel with respect to the size of the feed-forward (excitatory) one is chosen above an optimal value (e.g., \sigma_k=0.16*d and \sigma_h=0.7*d, where d is the distance of the lateral recurrent inhibition), Gabor-like receptive fields reach the highest possible number of sub-regions by acting on the strength of inhibition b, which can be increased up to the limit of network instability, with no impact on the number of the required synaptic interconnections.

We have included these considerations in the manuscript.

8 - A description of the grating experiment in the supplementary would be good, clearly pointing out how the spatial frequency is defined and what the sinusoidal grating inputs looked like. For example, in the text corresponding to Fig. 6, the authors say they are varying the phase of the sinusoidal grating, but how exactly is that done in the input?

As suggested by the Reviewer, we added a section in the Supplementary Notes in which we detailed the definition of the moving sinusoidal gratings used in the experiment, including also explicative figures (e.g., see figure R7 here below). The mathematical expression of the traveling grating has been also included in the 'Methods' section (see also point 5 above).

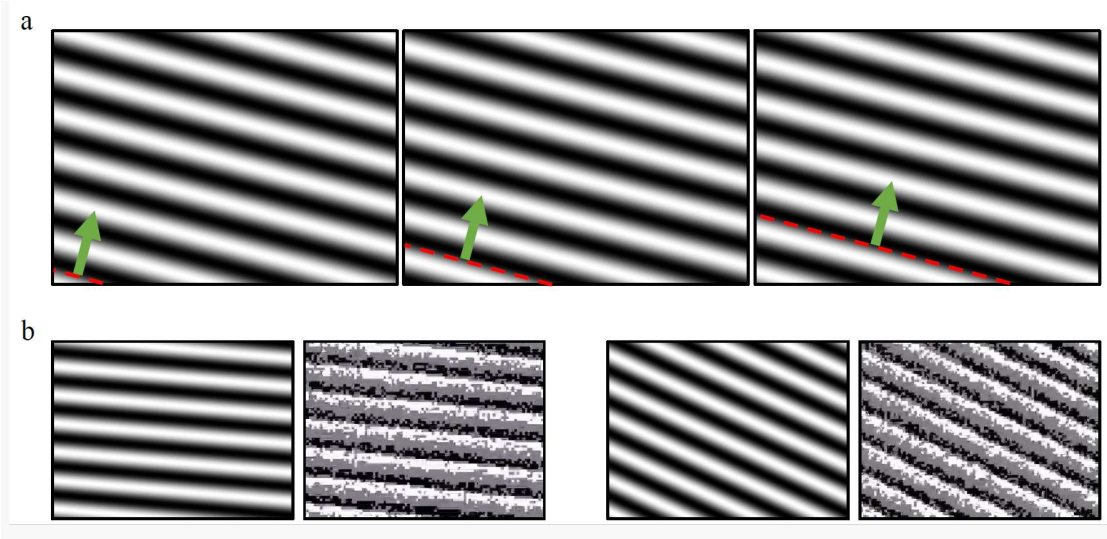

Figure R7: a) Consecutive snapshots of a moving sinusoidal grating with a temporal frequency of 3 Hz captured at intervals of 0.2 s. The red dashed line highlights the same wavefront in all snapshots, whereas the green arrow indicates the direction of movement, perpendicular to the wavefront. **b)** Pairs of snapshots of moving sinusoidal gratings as reproduced on the screen and of the corresponding DVS recordings as visualized through the jAER interface.

9 - Statements like "In this paper, we have demonstrated that recurrent clustered inhibition can be successfully used in SNNs, both in simulation and on mixed-signal analog/digital neuromorphic hardware, to economically implement highly structured visual receptive fields." from the discussion section, would be more appropriate to omit as this has already been done before.

We agree with the reviewer, for added clarity we have rephrased these statements and added appropriate references:

In previous works [VISAPP20 and ISCAS20], we indicatively demonstrated that recurrent clustered inhibition can be successfully used in SNNs, both in simulation and on mixed-signal analog/digital neuromorphic hardware, to economically implement highly structured visual receptive fields. The results of this paper corroborate those preliminary findings, specifically extending the analysis of the linearity of the resulting receptive fields when using the effective firing rate of the retina (ON firing rate minus OFF firing rate) instead of both as a whole or separately. Such a push-pull combination of the complementary ON and OFF channels led to more reliable and unbiased representation of the harmonic content (see phase and energy in Fig.~6d) which would eventually lead to steeper tuning curves of the V1 neurons, resulting in better selectivity to the local orientation, spatial frequency and phase of the visual input.

10 - Another statement: “The proposed solutions meet the requirements posed by an effective cortical-like representation in terms of channel redundancy (i.e., a large number of orientations and spatial frequencies), and in terms of information efficiency (i.e., highly structured basis functions).” in the discussion section is vague and unsubstantiated. Where have the requirements been laid down? The concept of having multiple banks of Gabor filters with different parameters has already been laid down in several studies in literature, both in the non-spiking domain and in the spiking domain, many by the authors themselves. So, this so-called requirement is not met for the first time in this paper.

We agree with the reviewer. Indeed, employing multiple banks of Gabor filters at the front-end of a bio-inspired vision system is not a novel concept *per se* [Daugman 1984; Watson, 1987; Riesenhuber & Poggio, 2000; Carandini et al., 2005; Dapello et al., 2020]. Several band-bass filters characterized by different orientations and bandwidths are usually employed in the linear stages of early vision for extracting compact and complete information about the local structure in the visual signal, on which to build higher-order image descriptors. However, the novelty here is in having proposed an economic implementation of nearly linear receptive fields that can be efficiently scaled with the kernel size. Although examples of hardware implementation of Gabor filters can be found in the literature [Shi 1999; Raffo et al., 1999; Cheung et al. 2005; Choi et al., 2005; Shimonomura & Yagi, 2008; Pauwels et al., 2012], to the best of our knowledge, this is the first time that linear Gabor-like receptive fields are implemented in hardware by a spiking neural network.

We have modified the text in the ‘Discussion’ accordingly, and included therein these considerations.

11 - The following statement in discussion: “We verified that the linearity assumption still holds despite the high non-linearity of spiking neurons. Additionally, we showed how it is possible to combine such feature detectors to generate Gabor-like filters with arbitrary phase values, effectively implementing a full harmonic representation of the image signal.” This is again misleading for the purposes of this paper, as this has already been shown in previous papers

[1-2]. Similarly, the last statement of the discussion section should also be omitted: “This would pave the way to the implementation of complex bio-inspired networks for more demanding online visual tasks on neuromorphic hardware.”

In the revised ‘Discussion’ we have deleted those statements, as suggested by the Reviewer.

12 - In the phrase “but inescapably discard part of the signal.” mentioned in the introduction, it is unclear as to which part of the signal the authors are referring to.

We agree that the statement is unclear. Accordingly, also considering that we get back on this concept further ahead in the ‘Introduction’, we have deleted the statement.

13 - Please add a proper reference to back up the following statement: “To assess network’s performance and characterize the receptive fields of its output neurons, we used two dimensional (2D) sinusoidal drifting gratings as visual stimuli, widely used to investigate the response of cells in the primary visual cortex”.

We have added the following references to support the statement:

- Graham, N. Spatial-frequency channels in human vision: Detecting edges without edge detectors. In Harris, C. (ed.) Visual coding and adaptability, 215–262 (Psychology Press, New York, NY,1981.
- Jones, J., Stepnoski, A. & Palmer, L. The two-dimensional spectral structure of simple receptive fields in cat striate cortex. J. Neurosci. 58, 1212–1232 (1987).
- De Valois, R. & De Valois, K. Spatial vision (Oxford University Press, 1988).

14 - Fig. 3b should show the different values of the parameters d, sigma, and b used to obtain the tuning curves. The caption provides the range in which they are varied but not the steps used.

The Reviewer is right, we have added the missing information in the figure caption.

15 - The following passage from the sub-section labeled “Linearity test and feature tuning characterization” has redundant statements and can be shortened: “The best results, i.e., the narrowest tuning curves, are obtained when the size of the recurrent inhibitory clusters and their distance from the target neuron are both comparable to the width of the feed-forward excitatory kernel. More precisely, considering fixed parameters for the feed-forward kernel (in particular $\sigma h = 3.5$, $p = 1/3$, and multiplicative weight of the feed-forward excitatory connections

a = 103), the parameters that influence the effect of the recurrent clustered inhibition are the distance d between the target neuron and each of the inhibitory clusters, the spatial extension of the clusters and the multiplicative weight of the recurrent inhibitory connections b."

We agree with the reviewer. We have shortened the passage accordingly:

The narrowest tuning curves are obtained when recurrent inhibitory connections cluster at a distance (d) comparable to the width of the feed-forward excitatory kernel (\sigma_h). Other parameters that play a key role in shaping the periodicity of the resulting receptive field profiles are the spatial extension of the clusters (\sigma_k) and the strength of the recurrent inhibitory connections (b).

16 - For lines 170 to 179, the authors should add simulation results to support this statement or refer to work that has demonstrated this before.

As suggested by the Reviewer, we added the results in the Supplementary Notes (see Supplementary Fig.3).

17 - The supplementary figure reference in line 201 is incorrect. It should be S2.

We thank the reviewer for having pointed it out. Yet, since we have removed that figure from the Supplementary Notes, we have removed that reference in the manuscript, accordingly.

18 - The phrase "substantially lower number of" in line 326 is qualitative and has to be quantified. From the numbers mentioned, I would not call it "substantial".

We have removed the adverb 'substantially'. See also response to point 7.

Manuscript title: Recurrent models of orientation selectivity enable robust early-vision processing in mixed-signal neuromorphic hardware

Manuscript ID: NCOMMS-23-47095

Authors: Valentina Baruzzi, Giacomo Indiveri and Silvio P Sabatini

Journal: Nature Communications

Content type: Article

Submitted on October 10th, 2023, Decision received on November 15th, 2023

We would like to thank the Reviewers for the interesting comments, which helped us to improve the validity of the results, as well as the quality of the presentation. Reviewer's comments are reported in RED, our responses in BLUE and in PURPLE we report here the changed text. The same text is also highlighted in PURPLE in the manuscript.

Reviewer #3

In this paper, the author proposed a recurrent snn model for describing a retinocortical layer mathematically. In addition, the author implemented the model by a neuromorphic hardware system comprising a Dynamic Vision Sensor that emulates the transient pathway of real retinas and a mixed-signal Dynamic Neuromorphic Asynchronous Processor with neurons and synapses.

I have the following concerns for the technical quality of the submitted manuscript.

Remarks to the Authors

1 - The recurrent SNN model seems that it comes from the previous publication largely, as the author mentioned in the manuscript. If so, the author needs to make it clear what novelty the submitted manuscript can show in terms of model equations, etc.

According to the Reviewer's comment, in the revised manuscript we have substantially extended the novel contributions, with respect to previous publications. As a whole, the revised manuscript presents, in an organic and complete way, the body of all the results, only partially covered in our previous conference papers, and integrates them with:

(1) an additional modeling and analysis of the combination of ON and OFF channels provided by the DVS sensor;

(2) a thorough mathematical treatment of the role of the push-pull mechanism on the generation of almost linear response to luminance contrast visual stimuli;

(3) a detailed assessment of the accuracy and reliability of the local phase information extracted by a hypercolumn of V1 modeled neurons.

Furthermore, we have:

- modified the original Figure 2 of the manuscript by including top-view schemes of the feed-forward and the recurrent kernels for the different parameters' settings considered in the work;
- included in the same Figure two new panels (B and C) that pictorially show the interconnections of the recurrent scheme and those for the equivalent feedforward one;
- included considerations about other metrics of comparison between the recursive and feedforward scheme, such as power consumption and DYNAP-SE chip area;
- added a section in the Supplementary Note that details the definition of the moving sinusoidal gratings used in the experiment;
- improved the readability of the text, particularly in the 'Introduction' and 'Discussion' sections.
- added several missing important references to support statements, where necessary.

2 - The main contribution of the paper is thought to be the neuromorphic hardware implementation that enables an economic realization of the retinocortical layer by combining the DVS camera and neuromorphic asynchronous processor with the emulation results. If so, the author needs to make it clear why the hardware implementation is more economic than the others, in addition to comparing the number of connections shown in Table 1.

To make more complete the analysis, we have substituted Table 1 with the new Fig.4 of the revised manuscript that comparatively shows the number of interconnections required for the recursive and feedforward scheme, as the receptive field is scaled from 13x13 to 61x61 pixels, which also directly relate to the DyNAP chip area used.

The advantage of the RNN over strictly feed-forward schemes is up to more than 3x for a five sub-region receptive field with a size of 21x21, and progressively increases with the rescaling of the filter's size.

Furthermore, following the Reviewer's suggestion, we have included in the manuscript a discussion of the power consumption of the recurrent network, providing an estimate for the 21x21 hardware implementation.

To calculate the power consumption of the neuron of the relay layer and of the central neuron of the V1 layer we considered the following equation, which approximates the power consumption of a silicon neuron on the DYNAP-SE including spike generation and routing as primitive operations [Risi et al., 2020]:

$$P_n = r_{inp} (E_{spike} + E_{pulse}) + r_{out} (E_{en} + E_{br} + RT \cdot E_{rt})$$

where r_inp and r_out are the average input firing rate and average output firing rate, respectively; E_spike is the energy required to generate one spike, corresponding to 883 pJ; E_pulse is the energy required by the pulse extender circuit, corresponding to 324 pJ; E_en is the energy required to encode one spike and append destination, corresponding to 883 pJ; E_br is the energy required to broadcast one event to the same core, corresponding to 6.84 nJ; E_rt is the energy required to route the event to a different core, corresponding to 360 pJ; RT is set to 1 if the spike is sent to a different core, and is set to 0 otherwise.

(These details on the estimation are reported in section ‘Methods’ of the revised manuscript).

The average estimated power consumption for a neuron of the relay layer and for the corresponding neuron of the V1 layer, when responding to the preferred stimulus at the highest temporal frequency, is 6.54 µJ and 446 nJ, respectively.

However, a comparison with equivalent feed-forward networks was not possible, as, actually, they have not been implemented on the DYNAP-SE, and as the average number of spikes estimated by simulations significantly differ from those actually observed in hardware.

Concerning the chip area used, the reduction of area for the recurrent network implementation with respect to the equivalent feed-forward one, is still of a factor of 3, when considering a five sub-region receptive field, along with the reduction of the required interconnections.

We have included these considerations in the manuscript.

Manuscript title: Recurrent models of orientation selectivity enable robust early-vision processing in mixed-signal neuromorphic hardware

Manuscript ID: NCOMMS-23-47095B

Authors: Valentina Baruzzi, Giacomo Indiveri and Silvio P Sabatini

Journal: Nature Communications

Content type: Article

Submitted on 10th October, 2023. Decision received on 15th November, 2023

Submitted in revised form on 8th July, 2024. Decision received on 15th August, 2024

We would like to thank the Reviewers again for having acknowledged the improvements brought on the revised manuscript steered by their comments, and for the further observations, which helped us to better clarify the validity of the results, and their impact in the field of Neuromorphic Vision Processing.

Reviewer #1

Reviewer's comments are reported in RED, our responses in BLUE and in PURPLE we report the changed text, which appears also highlighted in PURPLE in the revised manuscript, with respect to the first revision. New references, highlighted in **bold,** are listed in progressive order at the end of this document, and appear with different numbers in the full bibliography of the revised manuscript.

1. The push-pull technique now makes sense. However, its impact on potential vision processing pipelines is still not clear. For example, the push-pull waveforms in Fig. 7 of the main text seem to be in the center of the only-ON and only-OFF waveforms, but their difference does not seem to be significant. How much improvement in the overall classification accuracy of a full-fledged vision processor based on these Gabor filters will be due to such (arguably) minor improvements in phase errors brought about by the push-pull technique?

We certainly agree with the reviewer that the accuracy of *classification* of the original images convolved with the proposed Gabor-like filter would be only marginally affected when considering the responses of the only-ON or only-OFF components with respect to the case of considering their push-pull combination. This because – generally speaking – image classification can well rely upon intrinsically 1D (i1D) properties, like edges and contours, which are often sufficient to obtain a compact and complete feature description that enables a similarity measure to be applied to the different samples of popular image dataset (e.g., N-MNIST and N-Caltech101 **[1]**, HOTS **[2]**, MNIST_DVS **[3]**, the event-based UCF-50 **[4]**, Plane Dropping Dataset **[5]**).

However, advancing the front-end stage of an image classification processor is not our goal. The effort due for implementing the proposed filtering stage can be appreciated only when one considers more complex machine vision problems.

Actually, the advantage of the push-pull configuration (over single polarity, either ON or OFF) becomes prominent, not to say crucial, when we compare the efficacy and stability of the associated wavelet-based feature maps in reconstructing the 3D properties of the objects in the scene, or their relative motion. The difference can indeed be negligible for extracting i1D objects' properties, like edges and contours, attributable to (sparse) local image energy peaks, disregarding phase information (but see **[6]**). This is not the case for unveiling intrinsically 2D (i2D) properties, like textures, which convey quantitative and dense information about the scene's 3D structure, and require extracting precise relations among the phases of the various harmonics **[7]**. In particular, accurate phase detection depends on an ideal quadrature pair of bandpass filters to obtain the analytic signal. The *dc* sensitivity of even-symmetry filter components (arisen from the locality of the basis functions of the wavelet transform) is a well-recognized issue, which we must deal with (**[8] [9]**), e.g. by correcting for, or constraining their shape **[7]**. The push-pull configuration automatically cancels the *dc* sensitivity, which otherwise would dramatically affect the reliability and stability of the local phase measurements and thus those of the derived visual features.

To substantiate this statement, we refer to facts and evidence reported in the literature for classical, and neuromorphic computer vision. An actual benchmarking would require the design of large-scale cortical networks of spiking neurons for depth or motion perception, which evades the scope of the paper.

In order to clarify this point (which indeed was not sufficiently discussed in the original manuscript), we have included:

- Specific comments in the "Results" section on how to interpret the comparative assessment of the different local phase estimates in view of what discussed herewith above.

  The zero mean (i.e., zero *dc*) feature of the combined response, differently from the others, yields to almost unbiased and reliable phase estimates (see Fig. 7d leftmost panel). The phase error and energy violin plot distributions underline this conclusion, pointing out the overall higher efficiency of the push-pull response compared to those of the ON and OFF channels, separately. Certainly, these differences would have only negligible effect on the (eventual) classification accuracy achieved from the band-passed images obtained by convolving the original images with the three filters. This because, typically, image classification can well rely upon local image energy peaks, which are sufficient for characterizing the different samples of popular image dataset (e.g., N-MNIST and N-Caltech101 **[1]**, HOTS **[2]**, MNIST_DVS **[3]**, the event-based UCF-50 **[4]**) used for benchmarking. However, the advantage of implementing the proposed filtering stage in the push-pull configuration becomes prominent, and even crucial, when we compare the efficacy of the associated phase-based feature maps in more complex machine vision problems. Accurate phase detection depends on ideal quadrature pair of bandpass filters to obtain the analytic signal. The *dc* sensitivity of the real (symmetric) part of the Gabor kernel is therefore an issue, which we must deal with **[8] [9]**, e.g. by correcting for, or

constraining their shape **[7]**. The push-pull configuration automatically cancels the *dc* sensitivity, which otherwise introduces a positive bias in the real part of the response that would affect the reliability and stability of local phase measurements and thus those of the derived visual features.

It is worth noting that, although in principle the value of the phase associated to each orientation channel is correct, its confidence decreases as far as the symmetry axis of the image structure deviates from the orientation axis of the filter. We can thus state that the energy value of the associated wavelet-like transform is not isotropic, since it is not invariant under rotations of the signal. The isotropy of the representation is yet regained when one considers the whole set of oriented channels (i.e., the whole hypercolumn [31]).

- A paragraph in the "Discussion" section to highlight the importance of stable feature extraction from the incoming visual signal.

Discussion

Neuromorphic sensing modules - Today's neuromorphic systems represent a promising alternative to conventional von Neumann architectures for both understanding and reproducing the properties of biological sensory processing systems, as they are subject to similar constraints in terms of noise, variability, and parameter resolution [32]. Reproducing the dynamics of biological neural systems using subthreshold analog circuits and asynchronous digital ones make these systems ideal computational substrate for testing and validating hypotheses about models of sensory processing for a wide range of application domains [33, 14]. In addition, their real-time response properties allow us to test these models in closed-loop sensory-processing hardware setups and to get immediate feedback on the effect of different parameter settings.

From pixels to features - Assuring sufficient resources to enable complex transformations - from pixels to features - and to implement the corresponding computational models for such transformations sets a specific challenge for such systems, being the amount of data and operations in visual processing intrinsically high. Indeed, front-end early vision modules have to construct high-dimensional quantitative representations of image properties, referable to local contrast variations across different orientations, and according to different spatial frequencies. Subsequent stages eventually combine these properties in various ways, to come up with categorical qualitative descriptors, in which information is used in a non-local way to formulate more global spatial and temporal predictions (e.g., see [34]). However, it is worth remarking that only rarely classical (i.e., frame-based) computational theories can be applied directly to event-based sensory data. More properly, the adopted solutions for object detection, pattern recognition, and scene property reconstruction rely upon algorithms and computational procedures that well conform to the peculiar properties of the sensory data representation. Considering specifically image classification tasks **[2][10][11]**, intrinsically 1D properties, like edges and contours, are often sufficient to obtain a compact and complete feature description that enables a similarity measure to be applied to the different samples of popular image dataset. Other applications, like depth perception, optic flow, or simultaneous localization and mapping (SLAM), more decisively rely upon the timings of events **[12][13][14]**.

Although fully exploiting the time coding of spikes trains can be extremely efficient, we cannot disregard extracting the information conveyed by the spatial structure (i.e., the texture) of the luminance pattern, which depends on precise relations among the phases of the various harmonics **[6] [7]**. We must ensure that such information is not lost. The latter indeed plays a pivotal role in gaining dense feature maps potentially informative for several machine vision applications. Extracting stable spatial image structure requires local operations to regularize the information contained in spike trains. This can be done afterwards, on the result of the interpretation of the event stream (as mostly adopted by event-based machine vision algorithms, e.g., see **[15]**), or concurrently with picking-up sensory signals. Having such an early stage dedicated to the extraction of general-purpose regularized features brings about enormous advantages in terms of adaptability and versatility for compositionally building or learning a variety of higher-order visual descriptors. At a first level of abstraction, it is thus important that the rate coding model of network's neuronal firing replicates the known encoding properties of the cells in the primary retinocortical pathway, according to a linear filtering model with appropriate kernels (i.e., receptive fields) **[16]**. It is well acknowledged that Gabor wavelets are a powerful tool to gain an efficient regularized representation of the information contained in frame-based visual signals, in terms of local amplitude, phase and orientation maps of the transformed signal.

Progress beyond state-of-art - In previous works [9, 10], we indicatively demonstrated that recurrent clustered inhibition can be successfully used in SNNs, both in simulation and on mixed-signal analog/digital neuromorphic hardware, to economically implement highly structured Gabor-like RFs. The results of this paper corroborate those preliminary findings, specifically extending the analysis of the linearity of the resulting RFs when using the net firing rate of the retina (ON firing rate minus OFF firing rate) instead of both as a whole, or separately. Such a push-pull combination of the complementary ON and OFF channels led to more reliable and unbiased representation of the harmonic content (see phase and energy in Fig. 7d) which would eventually lead to steeper tuning curves of the V1 neurons, resulting in better selectivity to the local orientation, spatial frequency and phase of the visual input. Employing multiple banks of Gabor filters at the front-end of a bio-inspired vision system is not a novel concept per se [35] [36] [37] [38] [39]. ~~Several band-bass filters characterized by different orientations and bandwidths are usually employed in the linear stages of early vision for extracting compact and complete information about the local structure in the visual signal, on which to build higher order image descriptors.~~ However, the novelty here is in having proposed an economic implementation of nearly linear RFs that can be efficiently scaled with the kernel size. Although examples of hardware implementation of Gabor filters can be found in the literature [40] [41] [42] [43] [44] [45], to the best of our knowledge, this is the first time that linear Gabor-like RFs are implemented in hardware by a spiking neural network. The resulting RFs are characterized by spatial profiles and by tuning curves that are typically sharper than the ones obtained using equivalent feed-forward schemes. Yet, RFs obtained through a recursive scheme use a lower number of interconnections than that required when using an exclusively feed-forward approach. The advantage of the recurrent network over strictly feed-forward schemes is up to more than 3× for a five sub-region RF with a size of 21 × 21, and increases with the rescaling of the filter's size. This is an important feature when dealing with the limitations in terms of available synaptic connections posed by neuromorphic processors.

In summary, the solution proposed in this work demonstrates that an early vision filtering stage can be implemented in mixed-signal neuromorphic hardware in a relatively economic way, with adequate accuracy and stability. Particularly, exploiting both ON and OFF channels – through their push-pull combinations – shows to be an appropriate approach to remove the undesired effect of *dc* component sensitivity (see Section "Results"), and thus obtain highly informative phase-based features. The implemented units act as multiple oriented bandpass frequency channels, well supporting a compact and reliable representation of position, orientation and phase of local image patches. As a whole, the resulting harmonic signal description provided by the proposed neuromorphic circuit could be potentially used for a complete characterization of the 2D local structure of the visual signal in terms of phase relationships from all the available oriented channels. The amplitude (i.e., energy) information can be used as an indicator for the likelihood of the presence of a certain structure, while the orientation of contrast transitions and their spatial symmetry (i.e., phase, **[7] [17]**) can be used as an attribute of the visual descriptor.

2. Also the authors provide information on the synaptic sparsity advantage of their technique but how much silicon area and energy (based on system level analysis) will be saved by using the sparse recurrent inhibition incorporated Gabor filter implementation compared to ones with only feedforward connections. How will the performance and hardware cost of such pipelines compare with other spike-based vision processing pipelines already studied in the literature?

The number of interconnections of a feedforward network that can produce receptive fields comparable to those obtained with the proposed recurrent connectivity scheme is approximately 2.5 times higher (e.g., a feedforward network for a 7-pixel wide Gabor patch with 5 subregions requires 481 connections, while its recurrent equivalent requires 191 connections). As a consequence, the number of wires required in an equivalent feedforward architecture is at least 2.5 times longer, assuming a best case scenario in which one interconnection requires just one unity wire element (i.e. a square metal layout block). In this case both area usage and power consumption would increase by at least a factor of 2.5. In practice, area is likely to increase significantly more (because VIAs need to be taken into account and additional wire lengths to optimize routing). The increase in power consumption will depend on the activity (i.e., voltage changes) on those wires. Assuming sparse activations, the factor of 2.5 is a good estimate. As, to the best of our knowledge, there are no other feedforward equivalent architectures present in the literature, it is impossible to make comparisons with past designs.

We have added these considerations in the "Results" section.

References

[1] Orchard, Garrick and Jayawant, Ajinkya and Cohen, Gregory K. and Thakor, Nitish. Converting Static Image Datasets to Spiking Neuromorphic Datasets Using Saccades. Frontiers in Neuroscience vol. 9, 2015

[2] [Lagorce, X.; Orchard, G.; Gallupi, F.; Shi, B.E.; Benosman, R. HOTS: A Hierarchy Of event-based Time-Surfaces for pattern recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 8828.

[3] Serrano-Gotarredona, T., and Linares-Barranco, B. (2015). Poker-DVS and MNIST-DVS. Their history, how they were made, and other details. Front. Neurosci. 9:481. doi: 10.3389/fnins.2015.00481

[4] Hu Y, Liu H, Pfeiffer M, Delbruck T. DVS Benchmark Datasets for Object Tracking, Action Recognition, and Object Recognition. Front Neurosci. 2016 Aug 31;10:405. doi: 10.3389/fnins.2016.00405.

[5] Afshar, S.; Hamilton, T.J.; Tapson, J.; van Schaik, A.; Cohen, G. Investigation of Event-Based Surfaces for High-Speed Detection, Unsupervised Feature Extraction, and Object Recognition. Front. Neurosci. **2018**, 12, 1047

[6] M. Morrone, D. Burr. Feature detection in human vision: a phase-dependent energy model. Proc. Roy. Soc. Lond. B, 235 (1988), pp. 221-245

[7] Silvio P. Sabatini, Giulia Gastaldi, Fabio Solari, Karl Pauwels, Marc M. Van Hulle, Javier Diaz, Eduardo Ros, Nicolas Pugeault, Norbert Krüger, A compact harmonic code for early vision based on anisotropic frequency channels, Computer Vision and Image Understanding, Volume 114, Issue 6, 2010, Pages 681-699,

[8] D. J. Fleet and A. D. Jepson, "Stability of phase information," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 12, pp. 1253-1268, Dec. 1993, doi: 10.1109/34.250844.

[9] Crespi, B., Cozzi, A., Raffo, L. *et al.* Analog computation for phase-based disparity estimation: continuous and discrete models. *Machine Vision and Applications* **11**, 83–95 (1998).

[10] B. Ramesh, H. Yang, G. Orchard, N. A. Le Thi, S. Zhang, and C. Xiang, “Dart: distribution aware retinal transform for event-based cameras,” TPAMI, 2020.

[11] N. Messikommer, D. Gehrig, A. Loquercio, and D. Scaramuzza, “Event-based asynchronous sparse convolutional networks,” in Proc. ECCV, 2020.

[12] Osswald, M., Ieng, SH., Benosman, R. et al. A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems. Sci Rep 7, 40703 (2017).

[13] S. Shiba, Y. Klose, Y. Aoki and G. Gallego, "Secrets of Event-based Optical Flow, Depth and Ego-motion Estimation by Contrast Maximization," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2024.3396116.

[14] J. Jiao, H. Huang, L. Li, Z. He, Y. Zhu, and M. Liu, “Comparing Representations in Tracking for Event Camera-based SLAM,” in Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2021.

[15] G. Gallego, H. Rebecq and D. Scaramuzza, "A Unifying Contrast Maximization Framework for Event Cameras, with Applications to Motion, Depth, and Optical Flow Estimation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 3867-3876, doi: 10.1109/CVPR.2018.00407.

[16] Adelson, E. H., & Bergen, J. R. (1991). The plenoptic function and the elements of early vision. In M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 3–20). The MIT Press.

[17] Zhitao Xiao, Zhengxin Hou, Changyun Miao, Jianming Wang. Using phase information for symmetry detection. Pattern Recognition Letters. Volume 26, Issue 13, 2005, Pages 1985-1994.