- APROVIS3D -

Analog PROcessing of bioinspired VIsion Sensors for 3D reconstruction

Document Reference:		
Title: **D3.2: Working algorithm for coastline morphology monitoring**		
Contractor: **UCA**		
Prepared by: **Amélie Gruel and Jean Martinet**		
Document Type: **Deliverable**		
Version: **1.1**		Pages: **8**
Classification: **External document**		

Document Track

Version	Date	Remarks and Authors
1.0	25/03/2022	First draft (A. Gruel and J. Martinet – UCA)
1.1	31/03/2022	Final version (J. Martinet – UCA)

Authors

	Role / Function	Name	Organisation
Prepared by	Project Coordinator/WP0L	A. Gruel and J. Martinet	UCA
Checked by	Quality check	J. Martinet	UCA
Released by	Project Coordinator/WP0L	J. Martinet	UCA
Approved by	Project Coordinator/WP0L	J. Martinet	UCA

TABLE OF CONTENTS

FIGURES

TABLES

No table

1. Introduction

During the implementation of the APROVIS3D project, WP3 targets ML algorithms for event data based on SNN. Task 3.2 more specifically focuses on depth estimation from stereo event sensors.

1.1. Purpose

This document D3.2 deliverable of the APROVIS3D project describes the model and implementation adopted for depth estimation.

Th D3.2 deliverable consists in an event-based algorithm for depth detection. Integrating the dual sensors from the stereopsis system, we will use the above algorithms to estimate the depth as sensed from the UAV for object recognition purpose. In particular, the generic learning algorithm developed in T3.1 will allow a low-level representation of generic 3D landmarks. Comparing the spiking response to similar landmarks in both sensors from the stereopsis system will then allow for a robust and ultra-rapid depth estimate. This algorithm will then be translated into the FPAA in T2.2, with a realization of the software algorithms into the hardware, knowing their constraints (as managed in WP4 and in collaboration with WP2). In particular, we will benchmark at the software level this event-based algorithm compared to a more classical frame-based approach in terms of fidelity and energy consumption. The main objective of this task is the definition of this algorithm as software before its integration into the hardware in T2.2, which will result in a working prototype that will be used in the WP5 for evaluation. Finally, the results of this task should provide a quantitative tool to detect changes in a coastline morphology, such as after a landslide and allow for the development of the working prototype.

The work described in this document has been produced by Huiyu Han during her 6 month internship at UCA form March to August 2021, under the supervision of Jean Martinet and Amélie Gruel.

2. Documentation

2.1. Applicable and Referenced Documents

#	Id	Description	Identifier (Ed Rev)	Date
AD1	FPP	Full Project Proposal	1.0	15.01.2019

2.2. Glossary and Terminology

Acronym	Definition
DMP	Data Management Plan
WP	Work Package
RGB-D	Red-Green-Blue-Depth
ROS	Robot Operating System
CSV	Comma-Separated Value
UAV	Unmanned Aerial Vehicle
SNN	Spiking Neural Networks

3. State of the art

Three works studying depth estimation using conjointly SNN and event based camera were reviewed, in order to settle on one specific approach. Those thee works were respectively produced by Osswald et al. in 2017 [Osswald, 2017], Dikov et al. in 2017 [Dikov, 2017] and Risi et al. in 2020 [Risi, 2020]. + Chauhan, Masquelier

3.1. Spiking stereo network of 3D perception

Osswald et al. [Osswald, 2017] implemented a 3D perception model using 3 populations of spiking neurons (Figure 1):

- the retina: this population serves as input to the network, and its cells project to the coincidence detectors. The retina cells receive the projection of an object sensed by two eyes, in order to implement the stereovision.
- the coincidence detectors
- the disparity detectors: this population pools the response from the coincidence detectors via excitatory and inhibitory connexions. The final output encodes a representation of the original scene in disparity space (x, y, d).

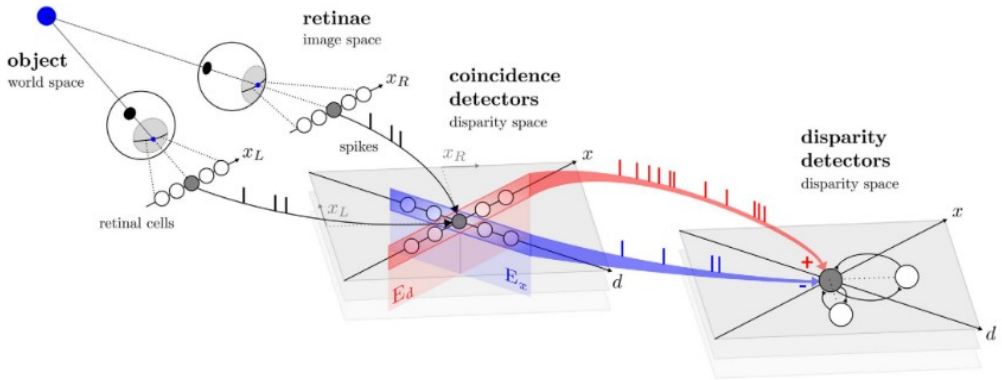

Figure 1: The spiking stereo network detailed view of a horizontal layer of the network. The spiking output of the retina cells is spatio-temporally correlated (coincidence detectors) and integrated (disparity detectors). For the sake of visibility, only a horizontal line of retinal cells, at fixed vertical cyclopean position y, is considered. The corresponding coincidence and disparity detector units, hence, lie within a horizontal plane (spanned by x and d). Only a few units are shown here whereas in the complete network, the units are uniformly distributed over the entire plane. The shaded planes indicate how the network expands vertically over y. Figure and caption taken from [Osswald, 2017],

The authors suggest that the excitation occurs in the plane of constant disparity Ed (in red on Figure 2) while the inhibition occurs in the plane of constant cyclopean position Ex (in blue on Figure 2). Thus leading to the inhibition of disparity detectors that represent spatial locations in the same line of sight.
To resolve ambiguities, we can combine the output of the “coincidence detectors” and “disparity detectors” to produce the final outcome of the network. This means that the network produces a “disparity event” only when the event produced by a disparity neuron is coincident (i.e. happens within a few milliseconds) with a

spike produced by a coincidence detection neuron at an equal (or nearby) representation in disparity space.

3.2. Spiking stereo-matching with reduced homolateral matching

Dikov et al. [Dikov, 2017] propose a spiking cooperative stereo-matching.
Their network can be understood as a three-dimensional array of cells (Figure 2), where each cell encodes a belief in matching a heterolateral pair of pixels from left and right images. Since pixel pairs have to satisfy the epipolar-geometric constraints, the size of the network grows asymptotically as the cube of the sensor's one-dimensional resolution.

In order to limit the number of false matches, assumptions based on the physical properties of common objects and geometrical configurations are made:

- Within-disparity continuity: this constraint consists in enforcing a weak potentiation between neighbouring cells representing the same disparity, since rigid bodies are cohesive. Since their surfaces are smooth, they should be represented by a smooth disparity map.
- Cross-disparity uniqueness: this second constraint translates into a strong negative interaction between certain cells in order to pair each pixel in one image to at most one corresponding pixel in the opposite image. In this case, the epipolar geometry can determine the precise patterns of inhibition.

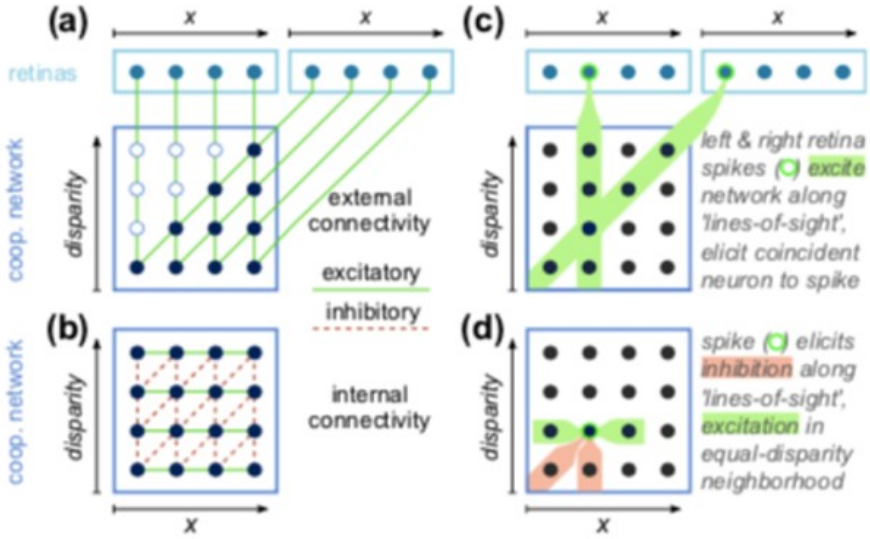

Figure 2: Network topology of Dikov's cooperative stereo-matching.
(a - b) Connectivity in terms of excitatory and inhibitory connections: neurons/cells (dots) are framed into populations and interconnected by synaptic projections (lines).
(c - d) Inhibition and excitation pattern for an exemplary pair of retinal input spikes (c) leading to a match, i.e. spike, in the cooperative network (d) which in turn triggers internal inhibition and excitation. Figure and caption taken from [Dikov, 2017].

"Homolateral matching" happens when a neuron in one retina is allowed to trigger the disparity neurons without any corresponding stimulus from the other retina. It can lead to a false matching in case the excitation of one retina by one pixel at high frequency.
To prevent this behaviour, the authors develop a neural mechanism to ensure only a pair of heterolateral pixel events could be considered as corresponding (Figure 3). The micro-ensemble consists of one

disparity-sensitive collector neuron C, and two blocker neurons B_l, B_r which are inserted between the retina cells R_l, R_r and C. The blocker neurons B are inhibitory interneurons. In the ipsilateral pathways $R_l \rightarrow B_l$ and $R_r \rightarrow B_r$, retina spikes are relayed from R_l, R_r to inhibit C, whereas retina spikes bypassing B_l, B_r excite C. The respective input weights w of C are matched such that the inhibitory (B → C) and excitatory (R → C) inputs cancel out:

$$| W_{B \rightarrow C} | = | W_{R \rightarrow C} |$$

The respective delays d is matched such that the spike propagation along the R → B and R pathways take the same amount of time:

$$d_{R \rightarrow B} + \tau_B + d_{B \rightarrow C} = d_{R \rightarrow C}$$

where τ_B represents the neuronal spike propagation time inside B.

Consequently, a single ipsilateral retina spike, e.g. from R_l, has zero net effect on C's membrane potential. A concurrent spike from the contralateral retina R_r, however, suppresses the blocker neuron B_l via an inhibitory contralateral projection $R_r \rightarrow B_l$ and therefore disinhibits C.

All of the previously described synaptic connections within the cooperative network are between C neurons, i.e. a C neuron spike excites neighbouring C neurons in the same disparity layer and inhibits other C neurons along with the two retinal input projections.

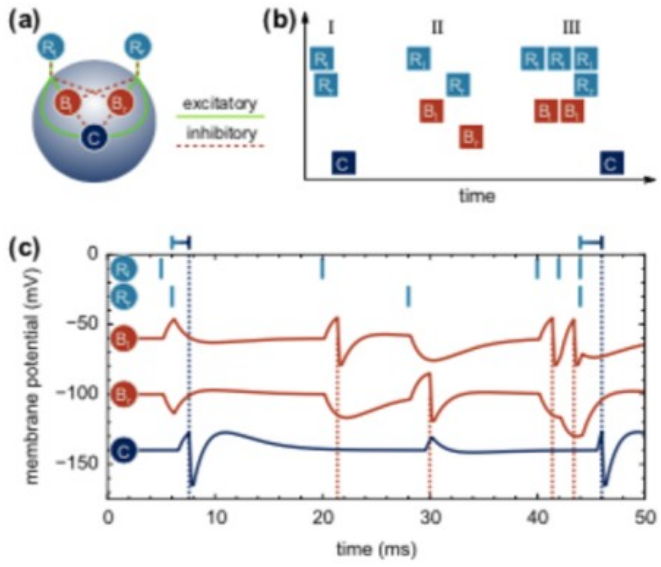

Figure 3: The working principle of neural micro-ensembles to prevent homolateral self-matching. (a) Inhibitory interneurons "Blockers" B_l, B_r are inserted between retina pixels Rl, R_r and the disparity sensitive collector neuron C.
(b) 3 typical spike sequences. (I) Left and right retina spikes arrive in quick succession and cause C to spike. (II) If R_r spikes too late after R_l , C does not spike. (III) Repeated retinal stimulation from R_l activates blockers B_l which annihilate R_l's excitatory input to C. Only once a contralateral spike from R_r adds additional excitatory C input and inhibits B_l, C spikes.
(c) Measured membrane potentials of B_l, B_r, C (traces offset vertically for clarity) during cases I, II, III in that temporal order. Vertical bars indicate retina events, dashed lines mark neuronal spikes. The rulers on top of the graph mark the 2 ms delay between retina input and network output.
Figure and caption taken from [Dikov, 2017].

3.3. Spike based architecture of stereo-vision

Risi et al. [Risi, 2020] adapts the structure developed by Osswald et al. [Osswald, 2017] presented above, by adding two biologically inspired mechanisms (Figure 4).

The first mechanism consists in connecting each coincidence detector to both of the corresponding input retina cells, once via a slow NMDA-like synapse and once via a fast AMPA-like synapse circuit block. This allows for the firing of the coincidence detector only when both synapses are stimulated in rapid succession.

The second addition to Osswald's model is the micro-ensemble developed by Dikov et al. [Dikov, 2017], also described above, in order to reduce the effect of high frequency homolateral excitation.

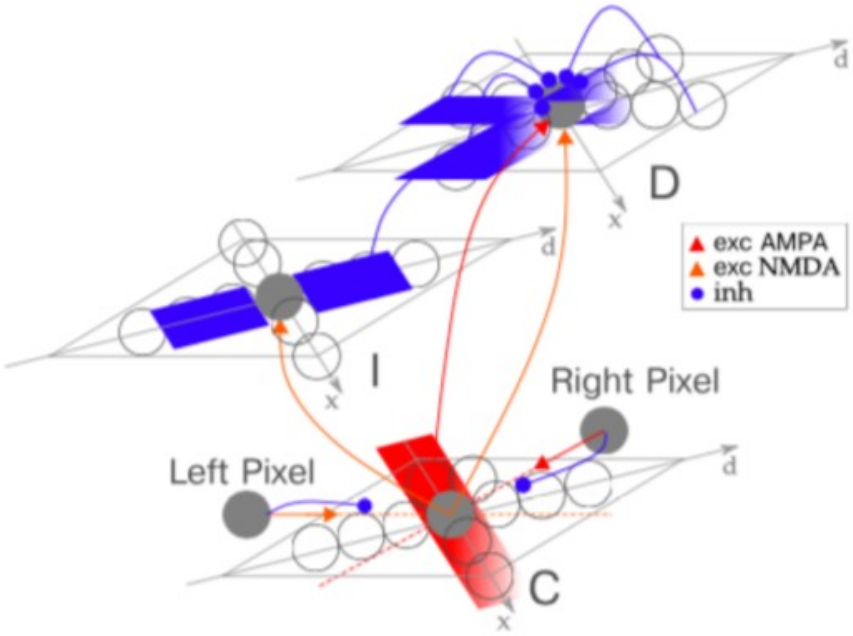

Figure 4: Detailed view of synaptic kernels for one cross-section of the network along the vertical dimension: coincidence detectors, excitatory neurons (C), coincidence detectors, inhibitory neurons (I) and disparity neurons (D). Figure and caption taken from [Risi, 2020].

4. Stereo sensing system designed

A stereo event camera sensing demonstrator has been designed at UCA during the summer 2021, based on the work of [Dikov, 2017]. Its network has been implemented using the PyNN interface by combining the coincidence and disparity detectors into one population. Since its architecture is the simplest one in terms of number of neurons and connexions, it was chosen in order to be run on one SpiNN-3 board.

Han's internship produced an event stereo system prototype at UCA, generating disparity maps with one SpiNN-3 board (resolution 22 x 22).

4.1. Spiking stereo neural network model

To better understand the neural aspect of the method, a unique horizontal and vertical cyclopean coordinate x and y, as well as a disparity coordinate d are assigned to every neuron (similarly to [Osswald, 2017]).
With D the one-dimensional disparity space, the three coordinates represent a point in the 3D disparity space D^3 which corresponds to the neuron's cognitive representation of a location in 3D space.
We can thus define the map M which transforms retinal image coordinates to disparity space as :

$$M : \quad N^3 \rightarrow D^3$$
$$M : \quad (x_l, x_r, y) \rightarrow (x, y, d) = (x_l + x_r, \ y, \ x_r - x_l)$$

where (x_r, y_r) and (x_l, y_l) are rectified pixel coordinates of the retinal input neurons.

The network uses leaky-integrate-and-fire (LIF) neuronal dynamics. The membrane potential $v_c(t)$ of a LIF coincidence neuron is described by the following equation:

$$\begin{cases} \tau_c \frac{d_{v_c(t)}}{d_t} = -v_c(t) + I_c(t), & v_c(t) < \theta_c \\ v_c(t) = 0, & v_c(t) \geq \theta_c \end{cases}$$

where the time constant τ_c determines the neuron's leak and θ_c the threshold at which the neuron fires.

A Collector (Dikov's micro-ensemble [Dikov, 2017]) firstly receives input from a pair of epipolar retinal cells, which can be described as a sum of spikes (assume the collector is only one neuron) :

$$I_c(t) = \omega \sum_i \delta_{x_l}(t - t_i) + \omega \sum_j \delta_{x_r}(t - t_j) \mid c$$

where the indices i and j indicate the spike times of the retinal cells (x_l, y_l) and (x_r, y_r) respectively. Then this Collector will aggregate evidence from responses the connected collector neurons with a distinct time constant τ_d and a firing threshold θ_d:

$$I_c(t) = \omega_{ext} \sum_{c \in c^+} \sum_k \delta_c(t - t_k) - \omega_{inh} \sum_{c \in c^-} \sum_k \delta_c(t - t_k)$$

where k represents the index of the spike times of c, while ω_{ext} and ω_{inh} are constant excitatory and inhibitory weights. The regions c^+ are neighborhood neurons with the same coordinate d (equal-disparity neighborhood, see Figure 2) and neighborhood neurons in the vertical cyclopean (different coordinate y) with the same coordinate x(x_r, x_l). The regions c^- are neighborhood neurons in the horizontal cyclopean (same coordinate y), with different coordinate d, but same coordinate x r or x l (along "lines-of-sight", see Figure 2).

4.2. Network simulation

The event stereo system prototype developed during the internship do not consider the polarity of events, and treat them all equally. Only positive disparity values are used, with d = 0 corresponding to distant objects.

The network size N (number of neurons in the network) depends on the number of retinal input pixel x_{max} and y_{max} as well as on the maximum disparity to be detected d_{max} :

$$N = (x_{max} - \frac{d_{max}}{2}) \times (d_{max} + 1) \times 3 \times y_{max} + 2 \times x_{max} \times y_{max}$$

In order to fit the network on the locally available system of 6 SpiNN-5 boards, the authors in [Dikov, 2017] performed experiments with values as low as :

- $x_{max} = y_{max} = 106$
- $d_{max} = 32$

4.3. Adjustment of the structure to be run on the SpiNN-3 board

A SpiNN-3 is composed of 4 chips, each containing 18 cores. Each core can simulate around 255 neurons. The total number of neurons that can be simulated on a SpiNN-3 equals : n = 4 x 18 x 255 = 18630

According to the equation presented above and the information about the total number of neurons that we can simulate, we need to ensure that N ≤ 18360.

If we assume that $x_{max} = y_{max} = d_{max} + 1$, we have at most $x_{max} = y_{max} = 22$ and $d_{max} = 21$.
One solution would be to crop one section of the experimental event data when run on the Spinn-3 board, but there is still not enough resources to run the data at the calculated dimension.

Only one population can be simulated per core on SpiNN-3 board, and one board is composed of 72 cores. For retina cells, the pixels with the same x coordinate are placed in the same population. Hence, the number of populations is x_{max} and the number of neurons per population is y_{max}. As for the micro-ensemble, all the neurons with the same x coordinate (xr, x_l) are part of the same population.

Therefore, the number of populations involved in this architecture equals $(x_{max} - d_{max} / 2) \times (d_{max} + 1) \times 2$ (once for the Blocker, once the Collector) and the number of neurons per population is dim_y and $dim_y \times 2$ respectively for the Collector and the Blocker.

Thus :

$$(x_{max} - \frac{d_{max}}{2}) \times (d_{max} + 1) \times 2 + 2 \times x_{max} \leqslant 72$$

Assuming that $x_{max} = y_{max} = d_{max} + 1$, we have $x_{max} \leq 7$.

According to the parameters above, only event data of small dimension can be simulated due to the network structure. Only $x_{max} = 7$ populations can be simulated.

Following this assesment, the network structure was adjusted in order to unit more neurons per population, in order to simulate event data with bigger dimensions. The theoretical calculation method is kept, to maintain the excitatory and inhibitory connections between the neurons.

The structure adjustments are the following :

- The pixels with the same y coordinate are grouped in the same retina population. The number of populations equals now to y_{max} with x_{max} neurons per population ; there are thus only 2 populations for the retina cells.
- The neurons with the same y coordinate are grouped in the same Collector population. There are now y_{max} and y_{max} *x 2* populations respectively for the Collector and the Blockers, with $(x_{max} - d_{max} / 2) \times (d_{max} + 1) \times 2$ neurons per population.

Since ymax x 3 + 2 ≤ 72 needs to be respected in order to run this architecture on one board, we have now y ≤ 23 with the new adjustments.

Assuming that $x_{max} = y_{max} = d_{max} + 1$, we have $x_{max} = y_{max} = 22$ and $d_{max} + 1 = 21$.

This allow for the simulation of a bigger section of the experimental event data, but not the whole scene. Future work will include the use of spatial event downscaling [Gruel, 2022] to reduce the dimension and allow the whole structure to be run on SpiNN-3,

5. Bibliography

[Dikov, 2017] G. Dikov, M. Firouzi, F. Röhrbein, J. Conradt, and C. Richter, 'Spiking Cooperative Stereo-Matching at 2 ms Latency with Neuromorphic Hardware', 2017, doi: 10.1007/978-3-319-635378_11.

[Gruel, 2022] A. Gruel, J. Martinet, T. Serrano-Gotarredona and B. Linares-Barranco, 'Event data downscaling for embedded computer vision', VISAPP, 2022.

[Osswald, 2017] M. Osswald, S.-H. Ieng, R. Benosman, and G. Indiveri, 'A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems', Scientific reports, 2017, doi:10.1038/srep40703.

[Risi, 2020] N. Risi, A. Aimar, E. Donati, S. Solinas, and G. Indiveri, 'A Spike-Based Neuromorphic Architecture of Stereo Vision', Front. Neurorobot., vol. 14, 2020, doi: 10.3389/fnbot.2020.568283.