Neuromorphic Sensory Integration for Combining Sound Source Localization and Collision Avoidance

Thorben Schoepe*, Daniel Gutierrez-Galan†, Juan Pedro Dominguez-Morales†,
Angel Jimenez-Fernandez†, Alejandro Linares-Barranco†, Elisabetta Chicca*

* Faculty of Technology and Cognitive Interaction Technology Center of Excellence (CITEC) - Bielefeld University,
†Robotics and Tech. of Computers Lab. - University of Seville, Email: tschoepe@techfak.uni-bielefeld.de

Abstract—**Animals combine various sensory cues with previously acquired knowledge to safely travel towards a target destination. In close analogy to biological systems, we propose a neuromorphic system which decides, based on auditory and visual input, how to reach a sound source without collisions. The development of this sensory integration system, which identifies the shortest possible path, is a key achievement towards autonomous robotics. The proposed neuromorphic system comprises two event based sensors (the eDVS for vision and the NAS for audition) and the SpiNNaker processor. Open loop experiments were performed to evaluate the system performances. In the presence of acoustic stimulation alone, the heading direction points to the direction of the sound source with a Pearson correlation coefficient of 0.89. When visual input is introduced into the network the heading direction always points at the direction of null optical flow closest to the sound source. Hence, the sensory integration network is able to find the shortest path to the sound source while avoiding obstacles. This work shows that a simple, task dependent mapping of sensory information can lead to highly complex and robust decisions.**

I. Introduction

Collision free navigation in a cluttered environment requires fast and robust decision making. Animals take decisions in a timescale of tens of milliseconds to execute collision avoidance [1]. Furthermore, they take more complicated decisions based on multimodal sensory information combined with previously acquired knowledge. For example, the female budgerigar (a small Australian parrot) incorporates auditory and visual input to track down a male. The bird uses auditory cues, the Inter-aural Level Difference (ILD) and the Inter-aural Time Difference (ITD), to estimate the male's position [2]. While approaching the male, the female effectively avoids collisions thanks to visual information. First investigations indicate that the bird merges optical flow (OF) information with other visual cues to avoid obstacles [3].

A few task specific Spiking Neural Networks (SNN) which combine different sensory cues and previously acquired knowledge have already been proposed. [4] and [5] increase the localization preciseness of their sensory integration network by merging different sensory cues which point at the same target. [6] combines previously acquired knowledge with one sensory cue to reach a target direction without collision. We present a new type of SNN which mimics the behaviour of the budgerigar and other animals. Hence, the network is able to identify and follow the direction of a sound source while avoiding obstacles. The model consists of an OF encoder (OFE) network and a sound source direction (SSD) network which receive sensory input from the embedded Dynamic Vision Sensor (eDVS) [7] and the Neuromorphic Auditory Sensor (NAS) [8] respectively. The two networks feed into the sensory integration (SI) network which chooses the agent's heading direction. We evaluate the network's performance in open loop by applying different combinations of auditory and visual stimuli to the two sensors.

II. Hardware

In this section the two event-driven sensors and the neuromorphic computing system used are introduced.

A. Dynamic Vision Sensor (DVS)

The AER DVS128 retina chip [7] comprises pixels which mimic the bipolar cells present in the mammalian retina. It consists of an array of 128×128 independent pixels that respond to relative light intensity changes in real time and are intrinsically invariant to scene illumination. A pixel produces an event in response to a change in luminance over time. As soon as the event is produced, the address of the pixel (x and y coordinates, and polarity) is written on an arbitrated handshaked asynchronous bus known as the Address-Event-Representation (AER) bus. The eDVS [9] consists of a DVS128 chip connected to an ARM microcontroller. This device is intended for embedded robotics.

B. Neuromorphic Auditory Sensor (NAS)

The NAS [8] is a spike-based audio sensor inspired by Lyon's model of the biological cochlea [10], implemented on FPGA. This sensor decomposes incoming audio signals in their frequency components as the inner hair cells do in the inner ear. It was implemented using a Spike-based Low-pass Filter (SLPF) bank with a cascade topology [11]. Each SLPF represents a frequency range, and its output consists of

This research was supported by the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG) and by the Spanish grant (with support from the European Regional Development Fund) COFNET (TEC2016-77785-P). The work of D. Gutierrez-Galan was supported by a ”Formación de Personal Investigador” Scholarship from the Spanish Ministry of Education, Culture and Sport.

978-1-5090-0617-5/19/$31.00 ©2019 IEEE

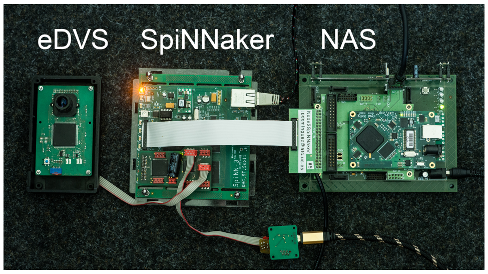

Fig. 1. eDVS and NAS feeding spikes into a SpiNN-3 board through the SpiNNlink connectors. The eDVS can directly process visual input while the NAS receives stereo audio input through an audio jack.

a stream of Address Events (AEs). In this work, we used a 64-channel binaural NAS generated with OpenNAS[1].

C. Spiking Neural Network Architecture (SpiNNaker)

SpiNNaker [12] is a massively-parallel multicore computing system designed for modeling very large SNNs in real time. We used a 4-node SpiNNaker machine, which consists of 72 ARM processor cores. It has a 100 Mbps Ethernet connection for the communication between the computer and the board and two SpiNNaker links. The latter were used to connect the eDVS and the NAS as input to the SNN (see Fig. 1).

III. Neural Network

In this section the three main sub-networks are presented.

A. Optical Flow Encoder Network

The speed of an object moving in the visual field of a translationally moving agent is inversely proportional to its relative distance. Bees, flies and some bird species use this visual cue called optical flow (OF) to safely navigate through densely cluttered environments [3]. Since the discovery of OF, various OF encoding algorithms and models have been designed [13], based on the Hassenstein-Reichardt-Detector [14]. One very recent OF detector model is the spiking elementary motion detector (sEMD) proposed by Milde et al. [15]. It encodes the time difference between two spikes from adjacent pixels in both the number of output spikes and the inter-spike-interval (ISI). In this paper we feed filtered data from the eDVS into a sEMD population to encode the spatial distribution of OF (see Fig. 2). A spatio-temporal filter population between the eDVS and the sEMDs reduces the noise and the spatial resolution of the visual information (See [15] for further information). The OFE network's output provides topographically arranged relative distance information in form of OF to the SI network. The whole OFE network was implemented on SpiNNaker.

B. Sound Source Direction Network

Birds use the ILD and the ITD to perform the sound source localization task [2]. While ILD achieves better accuracy with high-frequency sounds, ITD performs better when low-frequency sounds are present [16]. The ITD can be estimated by calculating the correlation between the input stimuli from both ears to determine the position of a sound source. According to [17], the correlation can be calculated by using an array of spike-based coincidence detector neurons. The excitatory output spikes from the cochlear nucleus (CN) are fed into those detectors through delay-lines. Depending on the sound source position, the sound waves arrive earlier to one ear than the other. Those input stimuli coincide in a specific coincidence detector, which identifies the estimated position. Note that the time difference is directly related with the distance between the ears. Since the ears' distance in birds amounts to just a few centimetres, time differences are in the tens of microseconds range. These fine temporal delays are not calculable on SpiNNaker due to limitations in temporal resolution. Because of that, a spike-based Jeffress model was designed as a real-time VHDL module to be added along with the NAS. However, the coincidence detector neurons project onto a winner-take-all (WTA) network [18] implemented on SpiNNaker to decrease the noise in the SSD network's output. The WTA network feeds spikes into the lateral sound transmitter (LST) population explained in the next section (see Fig. 2).

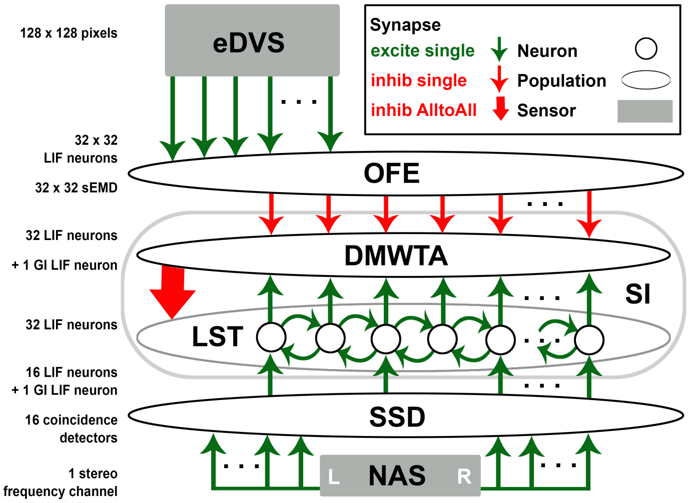

Fig. 2. Complete network. The OFE population consists of the spatio-temporal correlation (SPTC) leaky integrate-and-fire (LIF) population and the sEMD population. The SSD population includes coincidence detector neurons and an additional hard WTA network. Each WTA consists of an excitatory LIF neuron population and one global inhibitory (GI) neuron. All excitatory LIF neurons are connected to the GI neuron. The GI neuron projects back onto the excitatory LIF neurons.

C. Sensory Integration Network

The OFE network's retinotopic output map and the SSD network's tonotopic output map are arranged topographically. Both networks project (directly or indirectly) onto the SI network's decision making winner take all (DMWTA) map (see Fig. 2). This type of mapping seems to be quite efficient since it has been found in different vertebrates which have been optimized over millions of years [19]. The different sensory maps have to be correctly aligned to each other to

[1] https://github.com/RTC-research-group/OpenNAS

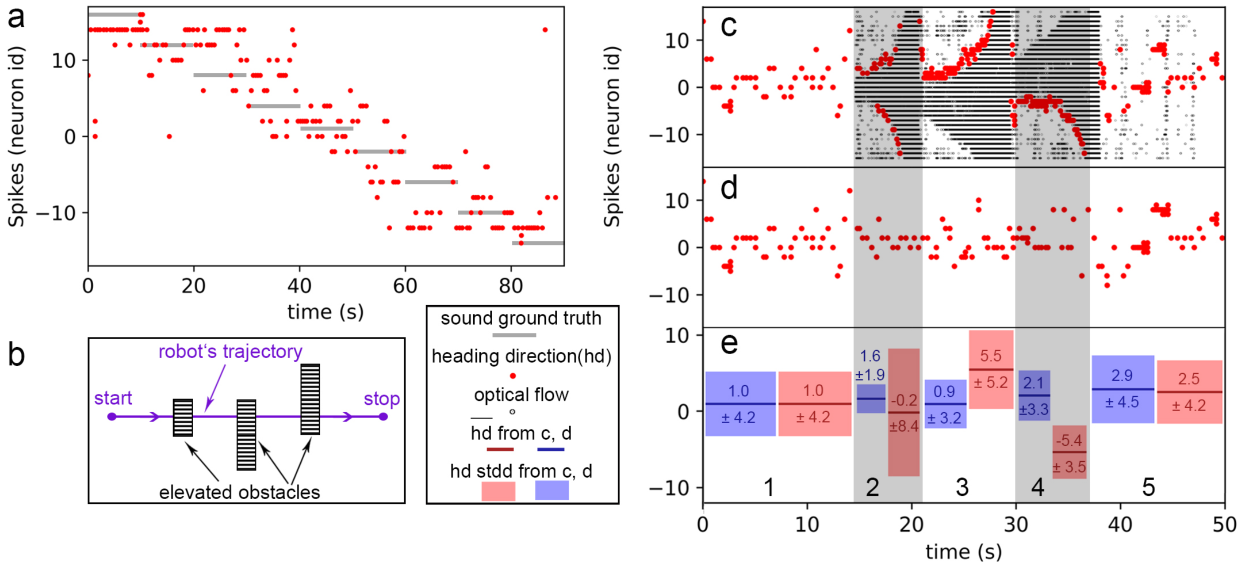

Fig. 3. (a) SI network's heading direction response to a 180 degrees sound source direction sweep. (b) Setup to record OF data used in Figure 3c. An eDVS is mounted on top of a robotic platform which drives purely translational with a speed of ~0.8 m/s through the scene. The first obstacle is located in the middle of the visual field, the second one on the right side and the last one on the left side. All obstacles are positioned at least 40 cm above the ground so that the robot can drive underneath them. Heading directions for a centered sound source with OF (c) and without OF (d). (e) Heading direction mean and heading direction standard deviation (stdd) with OF (red) and without OF (blue) for five different time periods with different visual scenarios

combine multimodal sensory cues in one network. In human beings the spatially more reliable visual input teaches the adaptation of the auditory input map [20]. Such an alignment adaptation has been simulated in neuromorphic systems [4] [5]. Given the current open loop configuration, there is no need for an adaptive alignment of the visual and auditory maps. Therefore, we simply map corresponding positions in all three networks.

Besides the alignment, the importance of the different sensory information with respect to decision making has to be taken into account. Visual information always dominates the proposed network since collision avoidance is an essential task to guarantee the agent's damage-free navigation. To achieve that, the OFE network's output (visual information) strongly inhibits the SI network's DMWTA population (see Fig. 2). This guarantees that the agent never drives into the direction of a nearby object because the DMWTA population's output defines the heading direction.

Whenever the visual field is object free the heading direction equals the sound source direction created by the SSD network. That means that the SSD network's output could directly be mapped onto the DMWTA population. Still, when an object appears directly in the sound source direction the corresponding DMWTA neuron is strongly inhibited so that it can not win. In this condition the LST population comes into play (see Fig. 2). The LST neuron positioned at the SSD excites the two adjacent neurons. This lateral excitation further spreads through the LST population until a position with zero OF input is reached. At that position the DMWTA population releases a spike. Since the DMWTA population consists of a hard WTA network [18] the winning neuron inhibits all other decision making neurons. At each instant the network can only decide for one specific heading direction. The selected heading direction always appears at the position of null OF closest to the sound source direction. This is caused by the fact that the lateral excitation wave in the LST population reaches the closest position with null OF with the smallest delay and with the highest excitation. The lateral excitation decreases with increasing lateral sound source distance given that the lateral connections are weak. This WTA structure matches with findings in mammals supporting the hypothesis that competing alternatives switch off each other through inhibition [1]. When a spike is released by the DMWTA population, it also sets back the LST population by inhibition.

IV. Experiments and Results

Two experiments were conducted to characterize the network's performance. In the first experiment only auditory information was fed into the network to verify that the SI network's heading direction follows the sound source direction. In the second experiment OF was added to investigate the network's behaviour when it tries to follow a sound source in a cluttered environment.

A. Sound Source Tracking

In this experiment, a synthetic audio file was generated by using a python script along with the Room Impulse Responses (RIR) generator library. In this script a 500 Hz sound source was swept from left to right at 2 meters receptor distance. This recording was fed into the NAS in order to check the SI network's sound source following behaviour. For all tests in this paper, events from one of the 64 NAS channels with a center frequency close to the sound source frequency of 500

Hz were used. As shown in Fig. 3a, the heading direction (identified as neuron id) follows the sound sweep from left (high id) to right (low id). The correlation between expected and achieved heading direction amounts to 89% (Pearson correlation coefficient) [21].

B. Sound Source Tracking and Obstacle Avoidance

In this experiment a synthetic audio file with a centered sound source generated similarly as in subsection IV-A was fed into the NAS. Additionally eDVS recordings were projected into the OFE network. These recordings were done with a robot executing pure translational motion through the environment shown in Figure 3b.

As long as only sound information is fed into the whole network the mean of the heading direction lies as expected close to neuron id zero, which corresponds to the sound source direction (Fig. 3d,e). The same accounts for region one in Figure 3c,e because the robot isn't moving. In Figure 3c, in region two, three and four a high amount of OF changes the heading direction. In region two, the heading direction's standard deviation is very high. As explained in subsection III-C, the heading direction always points to the direction of null OF closest to the sound source direction. Since the obstacle is located in the middle, there is no clear closest direction with zero OF and the heading direction fluctuates a lot between both sides. This could be seen as a problem but in case of a closed loop experiment the first laterally located spike will cause a turn of the robot so that the object is not centrally located anymore. What happens in case of a laterally positioned object can be seen in region three. The heading direction points significantly to the left. This can be explained by the fact that the obstacle is located at the right side. This makes the path at the left side around the obstacle the shorter one. In region four, the same effect can be shown but with the obstacle on the left side. After avoiding the obstacles the heading direction goes back to the middle. This is almost identical to the behaviour without OF (Region 5). As expected, the SI network always points at the direction of null OF closest to the sound source.

V. Conclusions

The proposed sensory integration SNN shows the expected behaviour: it adjusts its heading direction to the sound source direction with a correlation of 89%. When OF is introduced into the network the heading direction always points at the direction of null OF closest to the sound source. Hence, the sensory integration network is able to find the shortest path to the sound source while avoiding obstacles under well defined test conditions. These findings will be further investigated on a closed loop robotic platform.

Acknowledgment

Preliminary results were obtained at the 2019 CNE Workshop by D.G-G and T.S. Their attendance to the workshop was supported by the NeuroTech fellowship. Part of this work was carried out during research internships of D.G-G, J.P.D-M and A.L-B (Salvador de Madariaga program) in the Faculty of Technology (Bielefeld University). We are grateful to the SpiNNaker project [12] for the use of the SpiNN-3 board.

References

[1] A. B. Barron, K. N. Gurney, L. F. S. Meah, E. Vasilaki, and J. A. R. Marshall, "Decision-making and action selection in insects: inspiration from vertebrate-based theories," in *Front. Behav. Neurosci.*, 2015.

[2] C. C. Amagai, S. and R. Dooling, "Brainstem auditory time-coding nuclei in budgerigars: Physiology," *ARO Abstracts*, vol. 19, p. 191, 1996.

[3] D. L. Altshuler and M. V. Srinivasan, "Comparison of visually guided flight in insects and birds," *Frontiers in Neuroscience*, vol. 12, p. 157, 2018.

[4] V. Chan, C. Jin, and A. van Schaik, "Neuromorphic audio-visual sensor fusion on a sound-localising robot," *Frontiers in Neuroscience*, vol. 6, p. 21, 2012.

[5] H. Finger, S. Liu, P. Ruvolo, and J. R. Movellan, "Approaches and databases for online calibration of binaural sound localization for robotic heads," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2010, pp. 4340–4345.

[6] T. K. Horiuchi, "A spike-latency model for sonar-based navigation in obstacle fields," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 56, no. 11, pp. 2393–2401, Nov 2009.

[7] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor," *IEEE journal of solid-state circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[8] A. Jiménez-Fernández *et al.*, "A Binaural Neuromorphic Auditory Sensor for FPGA: A Spike Signal Processing Approach," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 28, no. 4, pp. 804–818, 2017.

[9] G. R. Mller and J. Conradt, "A miniature low-power sensor system for real time 2d visual tracking of led markers," in *2011 IEEE International Conference on Robotics and Biomimetics*, Dec 2011, pp. 2429–2434.

[10] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1119–1134, 1988.

[11] A. Jimenez-Fernandez, A. Linares-Barranco, R. Paz-Vicente, G. Jiménez, and A. Civit, "Building blocks for spikes signals processing," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–8.

[12] E. Painkras, L. A. Plana, J. Garside, S. Temple, F. Galluppi, C. Patterson, D. R. Lester, A. D. Brown, and S. B. Furber, "SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, 2013.

[13] T. Brosch, S. Tschechne, and H. Neumann, "On event-based optical flow detection," *Frontiers in Neuroscience*, vol. 9, p. 137, 2015. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2015.00137

[14] B. Hassentstein and W. Reichardt, "Systemtheoretische analyse der zeit-, reihenfolgen- und vorzeichenauswertung bei der bewegungsperzeption des rüsselkäfers chlorophanus," *Z. Naturforsch.*, vol. 11b, pp. 513–524, 01 1956.

[15] M. B. Milde, O. J. N. Bertrand, H. Ramachandran, M. Egelhaaf, and E. Chicca, "Spiking elementary motion detector in neuromorphic systems," *Neural Computation*, vol. 30, no. 9, pp. 2384–2417, 2018, pMID: 30021082. [Online]. Available: https://doi.org/10.1162/neco_a_01112

[16] Q. Liu, C. Patterson, S. Furber, Z. Huang, Y. Hou, and H. Zhang, "Modeling populations of spiking neurons for fine timing sound localization," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2013, pp. 1–8.

[17] L. A. Jeffress, "A place theory of sound localization." *Journal of comparative and physiological psychology*, vol. 41, no. 1, p. 35, 1948.

[18] M. Oster, R. Douglas, and S.-C. Liu, "Computation with spikes in a winner-take-all network," *Neural computation*, vol. 21, pp. 2437–65, 07 2009.

[19] J. H. Kaas, "Topographic maps are fundamental to sensory processing," *Brain Research Bulletin*, vol. 44, no. 2, pp. 107–112, 1997.

[20] Z. Shi and H. Mller, "Multisensory perception and action: Development, decision-making, and neural mechanisms," *Frontiers in integrative neuroscience*, vol. 7, p. 81, 11 2013.

[21] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.