Spike-timing Based Image Processing: Can we Reproduce Biological Vision in Hardware

Daniel Mayer

University Leipzig - Master of Science in Bioinformatics

Abstract.   In here I present a review of research and technologies using asynchronous spike-based processing strategies as can be observed in biological vision. The research presented includes a general description of the issue [1], the presentation of a silicon retina [2] as input device as well as the current research on memristive devices [3]. while the common approach to emulate neural networks is simulation on conventional hardware as CPU's and GPU's this review is rather focussed on the efforts of neuromorphics to reproduce the property of neural structures to work vastly parallel. In special this review will look at the why the biological vision system is able to compute relevant information with just one spike per neuron, due to the relative timing of spikes and the consequences on neuromorphic systems.

1   Biological Vision and its spike-timing-based component

A lot of research considers spike frequency to be the parameter in which stimuli are encoded in the nervous system. It is obvious to come to such a conclusion, since it is comparatively easy to measure the spiking frequency of a single neuron in dependence to a stimulus. But the speed of biological vision poses a problem on conventional views of how information is processed in the brain. As will be demonstrated, the notion of spike-frequency encoding of information contradicts speed at which neurons from the highest level of the primate vision system can respond selectively to complex stimuli such as faces. Studies have shown that such processes can be accomplished in as much as 100ms, which combined with the fact, that cortical neurons seldom fire faster than every 10 ms and that there are about 10 stages of neurons involved between the photo-receptors and the high level neurons, leads to the conclusion that each neuron has in average only time to fire once. It is therefore proposed, that rather than the information being only encoded in firing frequencies corresponding to analog values, the relative timing of the spikes is also used to encode such information [4] [5] [6]. It is though harder to measure, since one has to use multiple electrodes to measure relative spike timings between neurons. Still the idea has been demonstrated experimentally using recordings from the salamander retina [7]. Simulations also show, that it is possible to do state of the art face recognition using an approach where every neuron is only allowed to fire at most once [1]. It is inherent that if stimuli are encoded in the spike timing, also the neural plasticity needs to depend on relative spike timing, which leads to the notion of spike-time-dependent-plasticity (STDP)

2 Silicon retina

Conventional digital imaging chips have the immanent property of a frame-rate, which can be interpreted as a relict from the analog imaging times. All incoming information is collected in the sensor until the next frame is read out. For every frame the imaging chip sends out a pixel map containing the corresponding absolute value of each pixel. This yields massive output data regardless of the actual information((e.g. >1 GB/s from 352x288 pixels at 10 kFPS in [8])), Biological vision on the other hand is not bound to frames but is an event-based (asynchronous) relative intensity vision, meaning that the change of an individual pixel is send as soon as it registered. There are two essential advantages of the latter approach, which are (i) the speed of reception of individual events, and (ii) a natural compression, removing redundant information.
With regard to these differences, I refer a silicon retina as proposed in [2] which is a 128x128 pixel Cmos sensor, that outputs data in Adress-event representation (AER). The device improves on prior frame based temporal difference detection imagers (e.g., [9]) by asynchronously responding to temporal contrast rather than absolute illumination.

2.1 Adress-event representation (AER)

The modular design of computer hardware leads to bandwidth bottlenecks between the individual components. Thus unlike in the nervous system, where data transmission happens through many neurons in parallel, data needs to be multiplexed onto a bus, connecting the modules. The basic idea behind AER is to encode the visual stimuli into address events that are locally generated by the pixels. In this particular silicon retina the address event consists of 2x7 bits for the x and y coordinates and one ON/OFF polarity bit [2].
There has been some other research about AER imaging devices, but apart from that, the field of AER devices is largely unexplored. One reason could be that computer vision is only becoming popular now that computer power is sufficient to deal with the incoming information, so there has been little need for asynchronous data.

3 Memristive Neural Networks

In current Computer hardware information storage and computation are done in separate units, whereas in nature a neuron acts as a computational unit that keeps information stored in the synapses to other neurons. When modeling neural behaviour a spike signal has to be simulated on hardware that is foreign to the notion of spikes and relative timing, but abides in the realm of clocked binary logic. In fact an emulation of a part of the cat brain has been done on the BlueGene/P using 147 456 CPUs and 144 TB of main memory and a power consumption in the MW regime, but still being 83 times slower than the cat [10]. This ineficiency among other things due to the so called Neuman bottleneck[10]. To overcome the these constraints, the recent discovery of the yet

foreseen [11] fourth two-terminal canonical circuit element named memristor has lead computational neuroscientsts to dream of implementing a neural network directly on chip.

3.1 Memristor

Proposed theoretically in 1971 by Chua [11] and eventually identified in 2008 by Strukov et al, the memristor is a electrical component that relates charge q to magnetic flux φ. It is supposed, that it took nearly 40 years to experimentally prove Chuas prediction, because people were simply searching in the wrong place, mislead by the erroneous assumption that it had to be something dealing with magnetism, infered from that it related the magnetic flux. The type of memristor found by Strukov had in fact nothing to do with magnetism, but the effect is created by dopant ions drifting in a thin layer of titanium dioxide [12]. What makes the memristor exceptionally interesting, is that it retains the property of 'memorizing' the charge that has flown through. this happens by moving the dopant ions through the titanium dioxide according to the voltage across it, thus changing its resistance by changing the size of the volume that is made more conductive through the dopant. The memristor stores its history in its resistance, giving it the name memristor as a contraction of the words 'memory' and 'resistor'. The effect is reversed when reversing the voltage across the memristor leaving it unchanged by a symmetric alternating signal giving the possibility to read the value of the memristor without changing it. Because of the analog nature it is at the moment possible to reliably store 3 bit in a single memristor [13]. Apart from that memristor arrays can be manufactured much denser than NAND flash or magnetic hard drives, leading to an even higher storage density than what is physically possible with present-day storage techniques.

3.2 Memristor Arrays for Neuromorphic Systems

The property of keeping track of its past is very similar to the plasticity of synapses. Hence it has been proposed to use memristor arrays as synapses for neuromorphic systems. The proposed structure consists of crossbar arrays of memristors that connect artificial spiking neurons made up of a few transistors [14]. A network of this form wouldn't be trained by back propagation or other rather sophisticated techniques, but would be self learning Thus it is even nearly impossible to supervise or debug the process at a neural level not to mention extracting the learned information and putting it into a repository, usable by other computers. Since the memristor value is changed upon a bias voltage along the two terminals it is obvious, that plasticity can only come from relative spike timing and non-linearity of the memristic behaviour. Two neurons spiking synchronously at high frequency would leave the memristor value unchanged, as the potential difference across the memristor would stay very small. Spikes would have to be symmetric, for reasons pointed out above and the only chance of a memristor reacting to two well timed spikes is, that its change of resistance

is nonlinear to the bias voltage [15][13] $dR = V^\alpha dt$ with $\alpha > 1$.

4 Discussion

Spike based hardware is still under heavy development but the research done so far shows promising results for computer vision and neural networks. Most of the devices built today are purely academic. But it is unavoidable to research on STDP if we truly want to understand the dynamics of biological neural networks, and it could well lead to ways of building sensory implants that interact in a more natural way with the central nervous system.

Memristors being a fairly recent discovery there are still some problems to overcome. Even though the potential of memristive devices in neuromorphics is evident, it is yet to be seen if memristors will be available with resistances high enough and operating voltages low enough to keep power consumption low enough. Assuming wafer scale dies of 100 nm pitch which could be reached soon, it is well possible to put 10^6 neurons with 10^{10} synapses on a cm^2, but at 1 V and $1 M\Omega$ a spike frequency of 10 Hz with a spike duration of about 20ms that would lead to a power dissipation of 2 kW by the memristors alone.

In the biological case transmission of stimuli takes place at a mere 1-2 m/s, which combined with the length of axons yields a time base for relative and frequency timing. On a neuromorphic system data transmission happens nearly at speed of light with connections being even shorter than the axons so there is no relevant immanent latency and hence the timing of signals is going to have little context.

A particularly interesting fact is, as mentioned above, that neuromorphic systems are going to be very different to machines that we know today. Because programming will be similar to biological learning, unless trained in exactly the same way, the knowledge of one device will be physically bound to that device impeding possible 'gleichschaltung' in the sense that we attribute to software being installed on computers.

Spike timing dependent cameras with a small neuronal network attached to it might soon be used let's say for self driving cars, and they will not be objective any more, but will actively filter data according to their own(very limited) knowledge before transmitting it to the CPU or anything else.

To answer the question of the title: Steps have beend taken to make computer vision more nature like and they have been effectively used in particular applications. More nature like computer vision will have to be used in future, when conventional hardware reaches its limit considering speed or time resolution producing an inconsumable data stream. But for now there are no commercial solutions, as the topic is mainly still under development.

References

[1] Simon J. Thorpe. Spike-Based Image Processing: Can We Reproduce Biological Vision in Hardware? In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *Computer Vision – ECCV 2012. Workshops and Demonstrations*, number 7583 in Lecture Notes

in Computer Science, pages 516–521. Springer Berlin Heidelberg, October 2012. DOI: 10.1007/978-3-642-33863-2_53.

[2] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128 times; 128 120 dB 15 #956;s Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, February 2008.

[3] Sung Hyun Jo, Kuk-Hwan Kim, and Wei Lu. High-Density Crossbar Arrays Based on a Si Memristive System. *Nano Letters*, 9(2):870–874, February 2009.

[4] Simon J. Thorpe and Michel Imbert. Biological constraints on connectionist modelling. In *Connectionism in Perspective*, pages 63–92. Elsevier, 1989.

[5] W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland. Reading a neural code. *Science (New York, N.Y.)*, 252(5014):1854–1857, June 1991.

[6] W. Gerstner, R. Ritz, and J. L. van Hemmen. Why spikes? Hebbian learning and retrieval of time-resolved excitation patterns. *Biological Cybernetics*, 69(5-6):503–515, 1993.

[7] Tim Gollisch and Markus Meister. Rapid Neural Coding in the Retina with Relative Spike Latencies. *Science*, 319(5866):1108–1111, February 2008.

[8] S. Kleinfelder, SukHwan Lim, Xinqiao Liu, and A. El Gamal. A 10000 frames/s CMOS digital pixel sensor. *IEEE Journal of Solid-State Circuits*, 36(12):2049–2059, December 2001.

[9] U. Mallik, M. Clapp, E. Choi, G. Cauwenberghs, and R. Etienne-Cummings. Temporal change threshold detection imager. In *ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005.*, pages 362–603 Vol. 1, February 2005.

[10] Sung Hyun Jo, Ting Chang, Idongesit Ebong, Bhavitavya B. Bhadviya, Pinaki Mazumder, and Wei Lu. Nanoscale Memristor Device as Synapse in Neuromorphic Systems. *Nano Letters*, 10(4):1297–1301, April 2010.

[11] L. Chua. Memristor-The missing circuit element. *IEEE Transactions on Circuit Theory*, 18(5):507–519, September 1971.

[12] Dmitri B. Strukov, Gregory S. Snider, Duncan R. Stewart, and R. Stanley Williams. The missing memristor found. *Nature*, 453(7191):80–83, May 2008.

[13] Nima TaheriNejad, P. D. Sai Manoj, and Axel Jantsch. Memristors' Potential for Multi-bit Storage and Pattern Learning.

[14] Bernabé Linares-Barranco and Teresa Serrano-Gotarredona. Memristance can explain Spike-Time-Dependent-Plasticity in Neural Synapses. *Nature Precedings*, (713), March 2009.

[15] Zdenek Biolek, Dalibor Biolek, and Viera Biolková. Spice Model of Memristor With Nonlinear Dopant Drift. *Radioengineering*, pages 210–214.