University of Colorado
Boulder

NEUROMORPHIC STEREOSCOPIC IMAGING SENSORS

FINN JOEL BJERVIG

August 31, 2022

Abstract

This paper provides theory and discussion of a type of imaging sensor which mimics the neuro-biology of the human retina in terms of information acquisition and processing. It will be compared to traditional CMOS/CCD imaging sensors, where advantages of the neuromorphic *event-based imaging sensors* is presented in the context of low latency information acquisition and processing power efficiency. It will be clear that while this imaging technology is in a relatively early stage of development, and there is room for increasing efficiency and decreasing memory storage, it does provide promising results in niche areas of application such as autonomous driving and industrial manufacturing.

CONTENTS

1 INTRODUCTION

neuromorphic stereoscopic imaging sensors is a interdisciplinary field of neuromorphic engineering, first introduced by professor Carver Mead in the late 1980s [1]. The concept of neuromorphic engineering draws inspiration from the neuron architecture of the brains sensory organs, implemented into integrated circuits which generates asynchronous spiking voltage outputs similar to the brains neural signals. While the brain requires a lot of energy relative to the rest of the body, it is dwarfed by that of conventional super-computers which non-the-less falls short in carrying out the same tasks as the human brain. The outcome of neural architecture holds the promises of increasing power efficiency and lowered latency in areas as visual sensors. Conventional imaging sensors process large quantities of information, much of it redundant, both temporally and spatially. Imagine a field of green grass, and a bunny hopping across the field. The background (the grass field) remains fundamentally the same over time - temporally quasi constant [, and has the same appearance where ever you look at it - spatially quasi constant]. The bunny however, changes position over time and has distinct features, making the representing information of the bunny spatially and temporally dependent. This subset of pixels is what would actually be of interest. It is thus apparently useful for pixels in the neuromorphic imaging sensor to be aware of changes in intensity over time. When the change in intensity doesn't overcome a predefined threshold, it wouldn't send any "events" [ON/OFF - signal]. The property of being able to turn off pixels when there isn't a sufficient change, decreases power consumption, bandwidth, and storage remarkably.

2 THEORY

2.1 *Traditional versus event-based imaging sensor*

Traditional imaging sensors consists of a two-dimensional array of pixels made of MOS-transistors (Metal Oxide Semi-conductor), and works synchronously, meaning the electrical signals are transmitted by a specified rate determined by an external clock. The collection of these signals is known as a frame, and the rate of the data output is called frame rate and is measured as frames captured and processed per second. Gathering the information from the pixels synchronously poses two problems of opposite nature: Information loss and information redundancy. The frame rate is a consequence of the synchronous data extraction, and will naturally have a time constraint by the fact that it is not infinite, or rather, not continuous. Consequently, a feature in the scene might change in some property between the capturing of two frames and information from the scene is lost. On the other hand, if only a small area of a scene changes, the traditional sensor will continue capturing whole frames, with the majority of pixels not conveying any new information, spare for the changed subset of pixels which is actually of interest. This leads to information redundancy. Increasing the frame rate and applying real time [edge recognition/stereoscopic] processing to the technique of conventional frame-based data acquisition implies excess processing computations and is thus not very efficient. Additionally, the MOS technology only allows for global gain control, which narrows the dynamic range. [1] [2]

[1] Carver Mead is a Gordon and Betty Moore Professor Emeritus of Engineering and Applied Science at the California Institute of Technology

In an event-based imaging sensor each pixel works independently and asynchronously and must therefore be assigned individual wiring. Such a dense bundle solution is made possible by VLSI (Very Large Scale Integration). The pixels are sensitive to changes in light intensity against some set value at a specified time, and if the change is sufficient it will produce a spiking output, very much like human retina. The way communication works for event-based sensors is through address event representation. In essence the communication system consists of 5 parts. the transmitter (each pixel of the sensor), an address encoder, the bus system, and an address decoder for the receiver. The stream of information, called the event-stream, consists of separate events. An event in defined by a finite ordered list of elements: the pixels designated location in x-y coordinates, the timestamp t and the polarization. Polarity is a binary variable derived from the spikes generated, only taking on a negative or positive value depending on whether the intensity of the pixel increases or decreases. -1 equals [OFF] and 1 equals [ON]. Because of the asynchronous nature of the EBS (Event-Based Sensors) it has a very high temporal resolution on the scale of micro-seconds, but will also need a very sophisticated software to turn this non-unified stream of data into a synchronous image stream. Furthermore, it has a higher dynamics range than its traditional counterpart through local gain control: about 140 dB versus 60 dB, and consumes significantly less power. [1] [3]

$$Event = Event(x, y, t, p) \tag{1}$$

$$polarity = p \in \{-1, 1\} \tag{2}$$

2.2 *Human retina*

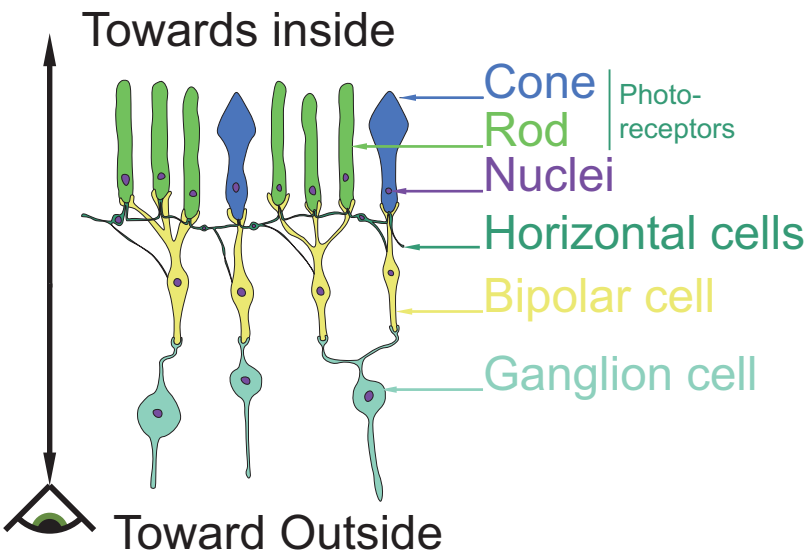

Figure 1: The retinal structure appears backwards, with the photoreceptor layer behind the bipolar and ganglion cell layer. The incident light propagates *towards the inside* of the eye, passing the two first layers, getting absorbed by the photoreceptors. From there, electrical signals is transmitted *towards the outside* of the eye, which then takes a turn towards the inside again through the optical nerve (not visible in figure)

In our interest of mimicking the fundamental process of the eye, only three main layers of the eye is taken into account, see figure 1. The first layer of the retina consists of photo receptors, which transmit electrical signals when excited by incident photons. To acquire a high dynamic range, two different types of receptors are present: cones and rods. Rods are sensitive to low intensity light, while cones are sensitive to high intensity light. Cones can vary their range of which they get excited by temporally averaging the light intensity in the scene. This adaptive mechanism is the reason of when walking out from a dim room out to the bright outdoors, the scene seems overexposed. There are different types of cones that absorbs different wavelengths of light - giving rise to color vision. Since cones require more light and are responsible for color vision, it explains why you can't see colors as well in a dimly lit setting. The next layer consists of horizontal cells, which connects to the photo receptors and the bipolar cells in a triad synapse. The horizontal cells are also laterally connected to their neighbors, so that the potential of one cell is equal to the weighted average of its neighboring cells. Roughly speaking closer neighbors has a stronger contribution than those further away. The weight distribution for each horizontal cell is closely approximated by the Laplacian of the Gaussian distribution function, which takes the shape of a Mexican hat, see figure 2.

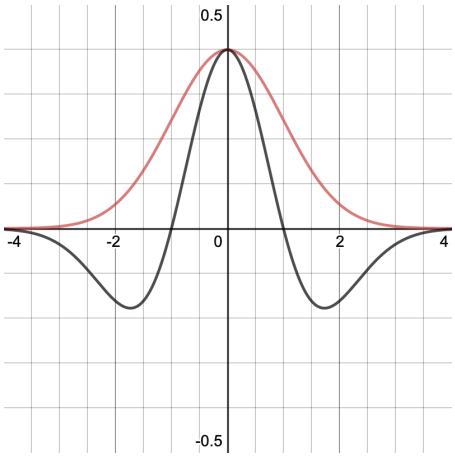

Figure 2: The black curve is the negative of the Laplacian of the Gaussian distribution function with the general equation 3 and the red curve is the Gaussian distribution function found inside the square brackets of equation 3

The general equation for the Laplacian of the Gaussian distribution function.

$$-\nabla^2 f_g = -\nabla^2 \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \right] \tag{3}$$

Next in line is the bipolar cells which receives signals from both the photo receptor and the horizontal cells and outputs a potential proportional to the difference between the two.

$$V_{bp} \propto V_{pr} - V_h \tag{4}$$

V_{bp} is the bipolar voltage potential, V_{pr} is the photo receptors potential and V_h is the horizontal cells potential. The potential of the photo receptor and the horizontal cell are actually both logarithmic, and therefore one might express the equation of potential for the bipolar cell as follows

$$V_{bp} \propto \log(V_{pr}) - \log(V_h) \rightarrow V_{bp} \propto \log\left(\frac{V_{pr}}{V_h}\right) \tag{5}$$

The output of the bipolar cell is therefore proportional to the ratio of local light intensity to the average intensity in the ”neighborhood”. An approximation is shown in figure 3. [4] [1]

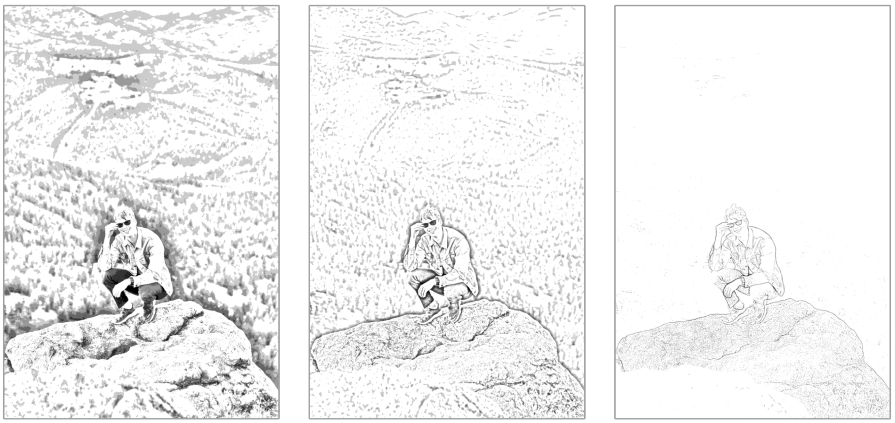

Figure 3: The original image (taken in black and white) is divided by a copy of it self but with Gaussian blur applied. The original image represents the local light intensity and the Gaussian blur image is an approximate to the Laplacian of the Gaussian but produces similar results to what the bipolar cell computes. From left to right the radius (size of neighborhood) of the Gaussian kernel decreases from about 200 pixels to 25 pixels.

2.3 *Depth Perception*

Depth perception poses a mathematically impossible problem: Extract three-dimensional information from a two-dimensional projection. Generalized it states that no information of N+1 dimensions can be obtained from its projection of N-dimensions. There exists many modules or cues involved in modeling a three-dimensional perception of our surroundings. Some of these are deriving shape from shadow, visual texture, past stimuli, motion and stereopsis, which can all be categorized into either monocular or binocular vision.

Monocular vision constructs three-dimensional geometry from only one eye together with knowledge of the object, and that it might be occluded by other known objects or by haze (the further away an object is, the more desaturated it appears due to particles in the air between the perceiver and the object of interest). Additionally, if there is an relative motion between the perceiver and the object, parallax will come into effect. Parallax is simply the illusion (but a helpful one non-the-less) that if several objects are located at different distances from the perceiver, they will appear to move at different velocities even though they are in fact moving with the same parallel velocity. Imagine sitting in a moving train, and the relative velocities of the trees going by outside. The trees further away seems to be moving slower to the ones closer to the train. One can argue that parallax is essential a time delayed stereoscopic cue, where the perspective is shifted by movement during a small timestep, as opposed to having two eyes perceiving two perspectives of a scene simultaneously. This is called stereoscopy.

Stereoscopic vision occurs due to the fact that we view the world through two eyes from different angles, with approximately a 6 cm lateral separation of the eyes. In other words, when the eyes focus at a point in space, the line of vision of both eyes converges towards that point, which means both eyeballs are mechanically at a relative angle from each other, and thus the distance to the object will not be equal for both eyes, giving rise to this binocular disparity. The process of figuring out what parts of the two images represents the same object is called the correspondence problem. Similarly, for computers, the algorithm tries to match pixels from each image (of slightly different lateral perspective) in order to compute the depth of the scene. The variable to solve for is called the disparity and is presented in equation 6. As complexity and noise increases in image, computational expense increases significantly in traditional vision systems. Thus, real time stereoscopic vision is slow. The so-called event-based visual sensors, dosent have this problem, because it transmits information asynchronously, based on if there is a significant shift of intensity in the pixel, instead of capturing the whole scene frame by frame as conventional imaging sensors do. This allows the neuromorphic system to tackle the previous problems of information loss and redundancy.[1] [5]

$$D = u - u' = \frac{B \cdot f}{L} \tag{6}$$

D is the disparity, and all other variables can be found in figure 4

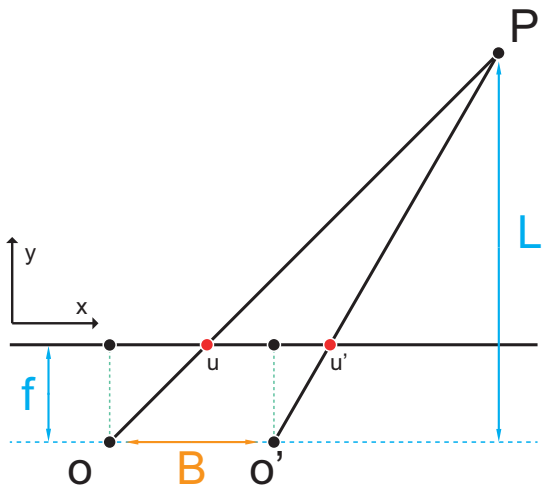

Figure 4: The eyes or the visual sensors are located at points u and u'. Equation 6 above shows that the disparity D is inversely proportional to the distance to the object L. F is the focal length and B the baseline length

Another cue for depth perception is the ability for humans to remember past stimuli. Essentially, if subject sees a tree in a forest at time t and evaluates its distance through several cues (stereoscopy, knowledge about the size of the object and its surroundings, obscurity, haze, etc.), it can be assumed at a later time$t + \Delta t$ the tree must be at the same distance as it was before(given that the perceiver hasn't moved during the time Δt). This can be extended to a multitude of features in the scene, so that three-dimensional perception can arise even though received information at that instant can't be used to create a three-dimensional perception. It is clear that memories and knowledge about objects which are located in the brain (not the eyes)

will help in evaluating distance. But it is not intuitive where the information of past stimuli is stored. It may well be stored in the eye, as a short-term memory, or it is also stored in the brain. Speaking of memory, at fairly short distances, say at an arm's length and less, the closer the object is, more tension is exerted on the eyes lens in order to bring the object into focus. This tension can be felt and can act as an additional cue to assess distances at short range. At even shorter distances the movement of the eyes as they converge towards the object can also be felt and serves as an additional cue. These mechanical cues are appropriately called oculomotor stimuli. [1]

3 DISCUSSION

I believe there is a place for neuromorphic engineered sensors and computers in the world of growing automation. Automation is the ability for a system to solve and execute the tasks, without the previous dependency of a manual operator. The concept is widely applied in the manufacturing and automotive industry (along with many others), with the goal of replacing a cognitive and mechanical task (essentially moving things) carried out by a human. It could be as the driver of a car, as fully autonomous driving which many car companies are racing towards, or as a manufacturing robot arm's which is some cases utilizes algorithms for object recognition. But currently both cars and manufacturing robots are using conventional cameras with a traditional computational architecture. The features that need to be automated in vision is mainly reaction to change, object recognition and depth perception. The areas of application above has the apparent common denominator: replacing certain cognitive tasks of the human brain, so how is the information processed and represented in the brain? Well, in order to replicate what would mimic the results we see in humans, researchers has to dug down to the fundamental structure of the sensory organs and the brain, where the sophisticated and energy efficient neural network comes into appearance. This was realized in the 1980s but has been more widely adopted starting in the early 2000s. The human retina, which is the source of inspiration for neuromorphic vision and event-based cameras, does not exclusively act as a information gatherer like a conventional camera, but it also carries out calculations in the retinal layer and gives rises to edge detection-like properties. But many other cues for interpreting the outside world is processed in the brain. Therefore, one cannot rely upon just an neuromorphic sensor, but a complementary neuromorphic computational processing unit is required. It might not be necessary for the computer to have a neural architecture, a conventional computer with neural net software implementation might be sufficient, just as the majority of all deep learning is based upon today. What cues requires this additional computer then? It would be one of the key modules in depth perception - stereoscopy, and its moving counterpart parallax. As we have seen, three-dimensional Reconstruction of a scene is dependent on many cues, but binocular vision stands out the most as being the most reliable, since other cues often give only relative positions, stereoscopic vision produces absolute values, which is far more superior. This is why stereoscopic event-based sensors most likely will utilize this feature. But while the latency of EBS are very low, when disturbances are introduced into the scene, efficiency drops noticeably, which calls for some sort of additional filtering. Another module would be object recognition which could be supported by a neural net. The main benefit of deploying an event-based camera is its efficiency in these types of areas: where attention to

changes are important, and where the responses needs to happen with very fast. Cars comes to mind of course, and industrial environments. But there is surely a wider area of application for this technology. The direction this technology needs to go now is further power and memory efficiency in terms of lowering the bandwidth which can be realized by applying a filter so that only relevant events are passed through the address event communication system. [2]

4 CONCLUSION

Neuromorphic sensors can adapt many cues from the nature of the human retina for the purpose of replacing certain cognitive tasks of the human brain as vision. The neuro-retinal architecture and its asynchronous event stream fundamentally differentiates itself from a conventional imaging sensor and offers substantial advantages. The event-based sensor, as it is named, offers very high temporal resolution and a low bandwidth and thus also low memory dependence. Its however only responsible for a small part of the necessary processing and depends on an external processor, preferably, but not necessarily to a neuron inspired architecture. The main properties in which an event-based camera differs from a conventional camera can be split up into four categories: Amplification, pre-processing, detection and quantization. In conventional cameras the gain control is affecting the whole sensor, while an EBS is designed to have a local gain principle, enabling for higher dynamic range. Secondly information processing occurs by the artificial neurons in event-based sensor processes information within the sensor it self neuron a s bandpass filter, while a normal camera sends raw information to the processing unit. Furthermore, such a camera has a fixed information acquisition frequency (framerate) while a neuromorphic EBS can adapt its frequency on an individual neuron level, making the information stream asynchronous. Conclusively, the primary advantages a neuromorphic event-based camera has is asynchronous data stream, low latency and processing power and its high dynamic range. All of which makes it superior to the conventional camera when considering high speed changes in the environment and action must be taken rather quickly.

REFERENCES

[1] L. Steffen, D. Reichard, J. Weinland, J. Kaiser, A. Roennau, and R. Dillmann, "Neuromorphic stereo vision: A survey of bio-inspired sensors and algorithms," *Frontiers in Neurorobotics*, vol. 13, p. 28, 2019.

[2] J. Barrios-Avilés, A. Rosado-Muñoz, L. D. Medus, M. Bataller-Mompeán, and J. F. Guerrero-Martínez, "Less data same information for event-based sensors: A bioinspired filtering and data reduction algorithm," *Sensors (Basel, Switzerland)*, vol. 18, p. 4122, 11 2018.

[3] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," 2019.

[4] M. A. Mahowald and C. Mead, "The silicon retina," *Scientific American*, vol. 264, no. 5, pp. 76–83, 1991.

[5] T. Poggio, "Vision by man and machine," *Scientific American*, vol. 250, pp. 81–96, 04 1984.