SCIENTIFIC REPORTS

OPEN

A Spiking Neural Network Model of Depth from Defocus for Event-based Neuromorphic Vision

Received: 30 July 2018
Accepted: 6 February 2019
Published online: 06 March 2019

Germain Haessig[1], Xavier Berthelon[1], Sio-Hoi Ieng[1] & Ryad Benosman[1,2,3]

Depth from defocus is an important mechanism that enables vision systems to perceive depth. While machine vision has developed several algorithms to estimate depth from the amount of defocus present at the focal plane, existing techniques are slow, energy demanding and mainly relying on numerous acquisitions and massive amounts of filtering operations on the pixels' absolute luminance value. Recent advances in neuromorphic engineering allow an alternative to this problem, with the use of event-based silicon retinas and neural processing devices inspired by the organizing principles of the brain. In this paper, we present a low power, compact and computationally inexpensive setup to estimate depth in a 3D scene in real time at high rates that can be directly implemented with massively parallel, compact, low-latency and low-power neuromorphic engineering devices. Exploiting the high temporal resolution of the event-based silicon retina, we are able to extract depth at 100 Hz for a power budget lower than a 200 mW (10 mW for the camera, 90 mW for the liquid lens and ~100 mW for the computation). We validate the model with experimental results, highlighting features that are consistent with both computational neuroscience and recent findings in the retina physiology. We demonstrate its efficiency with a prototype of a neuromorphic hardware system and provide testable predictions on the role of spike-based representations and temporal dynamics in biological depth from defocus experiments reported in the literature.

The complexity of eyes' inner structure implies that any visual stimuli from natural scenes contains a wide range of visual information, including defocus. Several studies have shown that defocus is essential in completing some tasks and more specifically for depth estimation[1,2]. Although a large body of research on Depth From Defocus (DFD) exists since the early 60's, there is currently a gap between the information output from biological retinas and the existing literature both in the vision science and computer vision that uses images as the sole source of their studies. Although images are perfect to display static information, their use in acquiring dynamic contents of scenes is far from being optimal. The use of images implies a stroboscopic acquisition of visual information (unknown to biological systems) at a low sampling frequency. They are thus unable to describe the full dynamics of observed scenes. On the other hand, retinal outputs are massively parallel and data-driven: ganglion cells of biological retinas fire asynchronously according to the information measured in the scene[3,4] at millisecond precision. Recent neuroscience findings show that this temporal precision can also be found in other subcortical areas, like the lateral geniculate nucleus (LGN)[5,6] and the visual cortex[7]. The last decade has seen a paradigm shift in neural coding. It is now widely accepted that precise timing of spikes open new profound implications on the nature of neural computation[8,9]. The information encoded in the precise timing of spikes allows neurons to perform computation with a single spike per neuron[10]. Initially supported by theoretical studies[11], this hypothesis has been later confirmed by experimental investigations[12,13].

Here, we present a novel approach to the depth from defocus, inspired by biological retina ouput, which is compatible with ultra low latency and low power neuromorphic hardware technologies[14]. In particular, we exploit advances made in both mixed signal Analog/Digital VLSI technology and computational neuroscience which enabled us to combine a new class of retina-like artificial vision sensors with brain-inspired spiking neural processing devices to build sophisticated real-time event-based visual processing systems[15–17]. We show how precise

[1]Sorbonne Universite, INSERM, CNRS, Institut de la Vision, 17 rue Moreau, 75012, Paris, France. [2]University of Pittsburgh Medical Center, Biomedical Science Tower 3, Fifth Avenue, Pittsburgh, PA, USA. [3]Carnegie Mellon University, Robotics Institute, 5000 Forbes Avenue, Pittsburgh, PA, 15213-3890, USA. Germain Haessig and Xavier Berthelon contributed equally. Correspondence and requests for materials should be addressed to R.B. (email: benosman@pitt.edu)

timing of spiking retinas allows the introduction of a novel, fast and reliable biologically plausible solution to the problem of estimating depth from defocus directly from the high temporal properties of spikes.

Silicon retinas located at the core of the hereby presented system are a novel piece of hardware which do not sense scenes as a serie of frames. Conventional cameras wastefully record entire images at fixed frame rates(30–60 Hz) that are too slow to match the temporal sub-millisecond resolution of human senses. Silicon retinas are asynchronous and clock-less, every pixel is independent from its neighbors and only reacts to changes caused by movements in a scene. Data are transmitted immediately and are scene driven, resulting in a stream of events with a microsecond time precision equivalent to conventional high-speed vision sensors, with the addition of being low power and sparse[18]. This type of acquisition increases the sensor dynamic range and reduces power computation.

Spiking Neural Networks (SNNs[19]) are computational models using neural stimulation. It has been shown that such networks are able to solve constraint satisfaction problems[20,21], depth extraction from stereovision[22,23] or flow computation[24,25]. As they are mimicking real neurons behavior, they allow a massively parallel, low power calculation, which is highly suitable for embedded computation. The use of a SNN in this work is a natural choice to build a complete neuromorphic event-based system, from the signal acquisition to the final output of the depth information. This is advantageous because of the resulting low-power system promised by the spiking/neuromorphic technology. The developed architecture is particularly adapted on a variety of existing neuromorphic spiking chips such as the SpiNNaker[26], TrueNorth[27] or LOIHI[28] neural chips. More specific neuromorphic hardware, such as the 256 neurons ROLLS chip[29], can also be used. When combined with an event-based camera, power as low as 100 mW is proven to be sufficient to achieve a realtime optical flow computation[25]. We are showing with this work that a low-power ($\leq$100 mW), computationally inexpensive and realtime DFD system can be similarly achieved.

Among the multitude of techniques developed by vision scientists to estimate depth, those called *depth from focus* (DFF) or *depth from defocus* (DFD) have the great advantage of requiring only a monocular camera[30]. The DFF method uses many images, and depth clues are obtained from the sharpness at each pixel. This method is computationally expensive and the amount of data to process is substantial. On the other hand, DFD estimates the variance of spatially varying blur spots based on a physical model. This technique requires less images but at the cost of a greater error in positioning. Current methods that use DFD or DFF generate depth maps for static scenes only[31] as they are limited by the frame rate of the camera driven at maximum of 25 fps. The computer vision and engineering community have described a number of algorithms for defocus computation[32–34]. However, they typically require multiple concurrent images[35–37], lightfield systems[38], specific lens apertures[35,39], correlations[40], specific hardware[41] or light with known patterns projected onto the environment[37]. The use of images and luminance implies high computational costs of around 17 ms to process a single frame[40].

These approaches cannot serve as conceivable models of defocus estimation in natural visual systems, as mammalian usually operate on a complete different data format and acquisition principles. Early studies[42,43] show that the border between blurred and sharp regions can be used to establish the depth-order of objects. For example, an out-of-focus target with a blurry textured region and a blurry border was perceived to be located proximal to the plane of focus, while an out-of-focus target with a blurry region and a sharp border was perceived to be located distant to the plane of focus. Recent findings in neuroscience show that blur perception in human is a dynamic process that allows depth assessment. In particular, the retinal defocus blur provides information regarding the relative and/or absolute distance of objects in the visual field[44]. Recently[45], it has been demonstrated that subjects were able to detect the relative distance of two vertical edges, justifying that the retinal blur allowed the subjects to judge target distance deferentially without any other depth cues. Other studies demonstrated that motor efference and/or sensory feedback related to the blur-driven accommodative response contain sufficient information to estimate the absolute distance of visual targets[46]. In addition, information derived from image blur can be integrated by the visual system with other visual cues (e.g., retinal disparity, size, interposition, etc.), which would assist in enabling one to judge the depth order of objects over a range of distances[43,47–50]. The addition of blur information can improve the speed and accuracy in such a depth-ordering task[51].

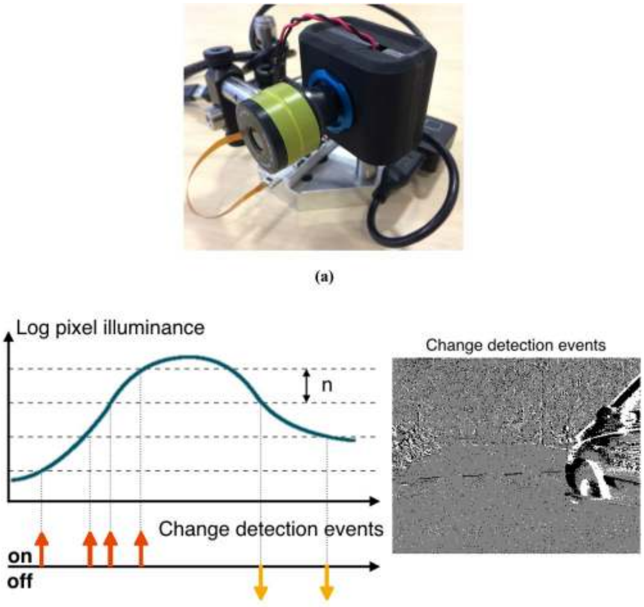
Materials and Methods

Event based cameras.

Biomimetic neuromorphic silicon event-based cameras are a novel type of vision sensor that are data driven. Unlike their frame-based counterparts, they are not controlled by artificially created timing and control signals (frame clock) with no relation to the source of the visual input. Events are generated when significant changes of the relative luminance occur at the pixel level as shown on Fig. 1. The visual output is in the form of an address event (AER) and encodes the visual information in the time dimension at the microsecond time precision. As soon as a change of luminance is detected, the process of communicating the event off-chip is initiated. The process executes with low latency, of the order of a microsecond, ensuring that the time at which an event is read out from the camera inherently represents the time at which a contrast change is detected. Let $e(x, y, p, t)$ be an event occurring at time t at the spatial location $(x, y)^T$. A positive change of contrast will result in an "ON" event ($p = +1$) and a negative change of contrast will result in an "OFF" event ($p = -1$). The threshold n beyond which a change of contrast is high enough to trigger an event is tuned according to the scene. Smaller intensity fluctuations do not generate any event and are not recorded. The camera used in our setup is issued from a new generation of asynchronous sensor based on[18] and developed by Prophesee. It has a 640 × 480 pixels resolution with a high temporal resolution of 1 μs. This array of fully autonomous pixels combines both a luminance relative change detector circuit and a conditional exposure measurement block (not used in the paper). When no change of luminance is detected, no events are generated and the static information is not recorded. This reduces the data load and allows high speed online processing at the native resolution of the sensor.

Depth estimation from the time of focus.

When a sweep of the focal length over its dynamic range is carried out, objects will successively appear out of focus, then in focus and out of focus again. The blurry


(a)

(b)

Figure 1. (**a**) The neuromorphic silicon event based camera with the variable motorized focal lens controlled at 100 Hz. (**b**) (left) Operating principle of event detection of an event-based camera: relative changes of the luminance greater than a predefined threshold *n* generate ON/OFF events when there is a positive/negative change of contrast. (right) Events output from the senors are shown as on the focal plane as a frame for purpose display, black dots represent OFF events while white dots represent ON events.

spot around the object will therefore shrink until the object is sharp and grow again as shown in Fig. 2(a) and in Fig. 2(b) for a cross section of the blur spot. The size of the blur spot increases in connection to the distance respectively to the depth of field (DoF) location. When the object is in focus, the image spot will have its minimum size and the contrast will be maximum (sharp edges). The DoF of the lens is increasing with the distance of focus (see *Supplemental data*). Beyond a certain distance, called the hyper-focal, the whole scene appears in focus and differences in depth can no longer be distinguished. Ideally a DFD sensor should have an infinitely thin DoF for each focusing distance and an infinite hyper-focal. In practice one needs to minimize the DoF and increase the hyper-focal to have the best spatial resolution in depth on the longest distance possible.

Let $s(t)$ be the size of the defocus blur at the focal plane. It will vary according the equation (see *Supplemental data* for details):

$$s(t) = \frac{f^2}{N} \times \frac{|z(t) - d|}{(d - f)z(t)}, \tag{1}$$

with f the focal value of the optical system, N the numerical aperture, d the position of the object when in focus and $z(t)$ the variable position of the object over time, or in other words the depth. Due to the aberrations, diffraction phenomenon and non-idealities of the lenses, a Gaussian point spread function (PSF) is commonly used to describe the defocus blur spot[52]. The spread parameter $\sigma(t)$ is proportional to the diameter $s(t)$ of the ideal blur circle, i.e. $\sigma(t) = \alpha s(t)$. The resulting intensity onto the sensor, at a pixel (x_i, y_i) is:

$$I_{i,j}(x, y, t) = A.\exp\left(-\frac{r_i^2}{2\sigma(t)^2}\right). \tag{2}$$

with $r_i^2 = (x - x_i)^2 + (y - y_i)^2$ and A the amplitude. At the pixel level the evolution of the intensity will depend on how close to the camera the object is. The Gaussian PSF is actually related to the classical formulation of the blur in the focal plane as a problem of 2D-heat diffusion. As such, the solution is the Green's function equivalent to Eq. (2). As a function of time, the standard deviation in I can be used to determine the time t at which an event is triggered by the pixel, assuming σ is invertible i.e.:

$$t = \sigma^{-1}\left(\sqrt{\frac{r_i^2}{2(\log A - \log I_{i,j}(x, y, t))}}\right) \tag{3}$$

We are dropping subscripts (i, j) for readability purpose as what we are describing is valid for any pixel. Hence, given the intensity at an arbitrary time t_0, if the variations of its log reach some threshold $\pm n$ (described in the previous section), then:

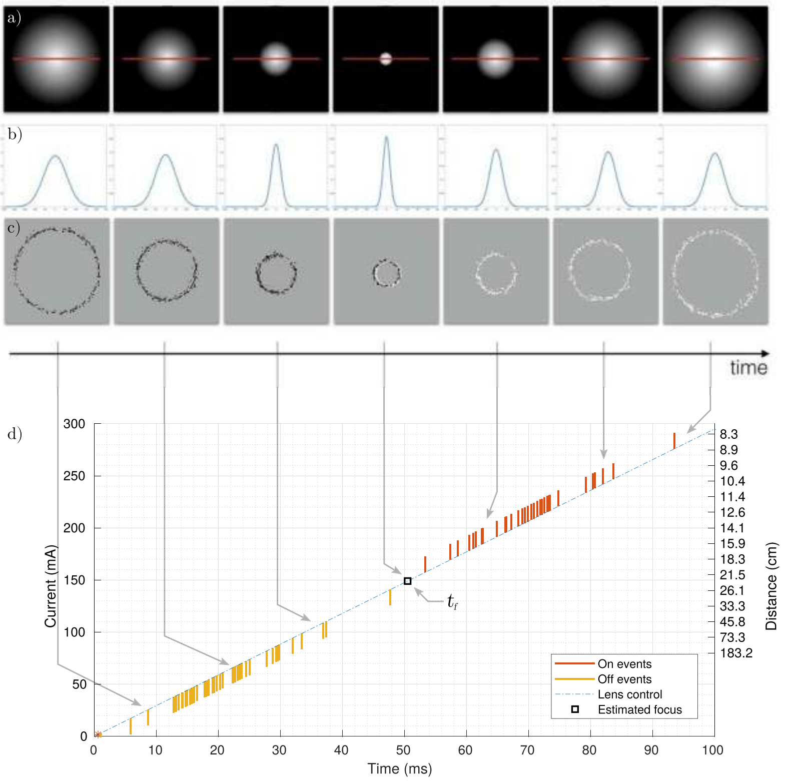


Figure 2. (**a**) Successive snapshots of a sphere when sweeping the focus range. The red line represents a line of pixels in the y direction. (**b**) Variations of the intensity profile along the red y-axis on the above snapshots. (**c**) Events corresponding to the sweeping of the focus range, in black are OFF events and in white ON events. (**d**) Representation of spikes among a single pixel, according to the driving current of the liquid lens. Here, the focus point is estimated to be at 22.6 cm from the sensor.

$$\log \frac{I(x, y, t)}{I(x, y, t_0)} = \pm n \text{ and } \log I(x, y, t) = \log I(x, y, t_0) \pm n. \tag{4}$$

This gives the time when an event is emitted according to (3):

$$t = \sigma^{-1}\left(\sqrt{\frac{r^2}{2(\log A - \log I_0 \mp n)}}\right) \tag{5}$$

The sign of n is chosen according to the polarity of the spiking event, itself related to the sign of the intensity's derivative:

$$\mathrm{SIGN}(n) = \mathrm{SIGN}(p) = \mathrm{SIGN}\left(\frac{dI}{dt}\right) \tag{6}$$

when the derivative is positive the polarity will be +1 (ON event) and −1 when negative (OFF event). Eq. (5) expresses when an event will be emitted w.r.t. n and to a reference event measured at t_0. As we reach focus, the value of σ will be constant for small duration of time, therefore the derivative of I, $\frac{dI}{dt}$ is equal to 0, followed by a polarity change as shown in Fig. 2(c) and expressed in the temporal domain in Fig. 2(d) around 50 ms. The detection of focus can then be determined by detecting the time t_f of the polarity change that can be estimated from the average timing between the consecutive ON and OFF events. We can then estimate the size of the defocus blur $s(t_f)$ according to (3) and deduce from (1), the depth information $z(t_f)$ as:

$$z(t_f) = \frac{\mp df^2/N}{S(t_f)(d - f) \mp f^2/N}. \tag{7}$$

The change of sign in z corresponds to the focal length that is the closest to the focus. Parameters d and f are controls of the liquid lens device.

Liquid lens control. The optical system shown in Fig. 1(a) is composed of three components:

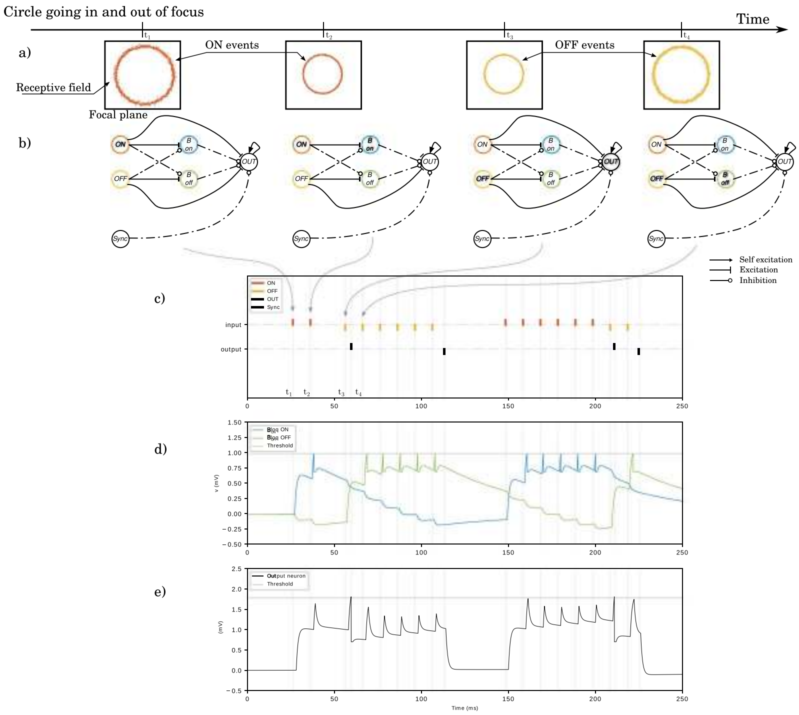


Figure 3. Spiking neural network. (**a**) Input data: a circle going in and out of focus, in front of a receptive field (a single pixel). (**b**) Neural network for focus detection composed of two input neurons, *ON* and *OFF*. They directly connect to the output neuron, and also to two blocker neurons B_{on} *and* B_{off} that are inserted to avoid parasite firings of the output neuron due to a sequence of only ON or OFF polarity events. A synchronization with the liquid lens via the *Sync* neuron is added, in order to encode the depth in the length of the spike train. (**c**–**e**) Simulation of the SNN with NEST. (**c**) The input spikes (ON and OFF events) and the output of the network (OUT and Sync). The point of focus is given by the OUT neuron, while the distance is encoded in the timing between the OUT and SYNC spikes. (**d**) Membrane potential for the two blockers neurons. After the first spike of its respective polarity, the blockers send a inhibition to the output neuron. (**e**) Membrane potential of the output neuron. Spikes from the same polarity do not allow the output neuron to reach its firing threshold, while a succession of ON and OFF events make the output neuron fire. As the output neuron is self-excitatory, the output spike train will be maintained until the strong inhibition from the synchronization comes.

- an electrically focus-tunable liquid lens with a 10 mm clear aperture and focus range f_{ll} ranging from 50 to 120 mm[53].
- an offset lens with a focal $f_o = -150$ mm. It acts as a relay imaging system between the focus-tunable lens and the objective and ensures a proper focus.
- an objective lens with focal length $f_{ol} = 35$ mm, $f_{ol}/2$ objective lens. This objective is a good compromise between large focal value, large clear aperture and low bulk (23.4 *mm* length). It is used to form an image directly on the camera pixel array.

The electrically focus-tunable lens used for this work is a shape-changing lens, consisting of an optical fluid whose deflection property changes w.r.t. the pressure applied on it via an electromagnetic actuator (coil). The focal distance is then controlled by the amount of current injected in the coil. More details can be found in Supplemental data[53].

The thin lens approximation is given as follows:

$$d = f_{eq} + \frac{f_{eq}^2}{D_{cam/obj} - f_{eq}}, \tag{8}$$

where d is the position of the point in focus, f_{eq} is the global optical system's equivalent focal value (liquid lens + offset lens + objective lens) and $D_{cam/obj}$ is the distance between the camera and the object. This technique is not specific to tunable liquid lens i.e. a mechanical focus controlled lens would also work as well if we can change the focus at a high enough frequency. However mechanical device has usually a reduced operational frequency and a shorter lifetime.

Spiking neural network. To estimate t_f for each pixel, we are looking for the smallest time interval between two consecutive events of opposite signs. We implement a Spiking Neural Network (Fig. 3a) based on Leaky Integrate and Fire neurons[54] to process the spikes from the output of the neuromorphic silicon retina. When

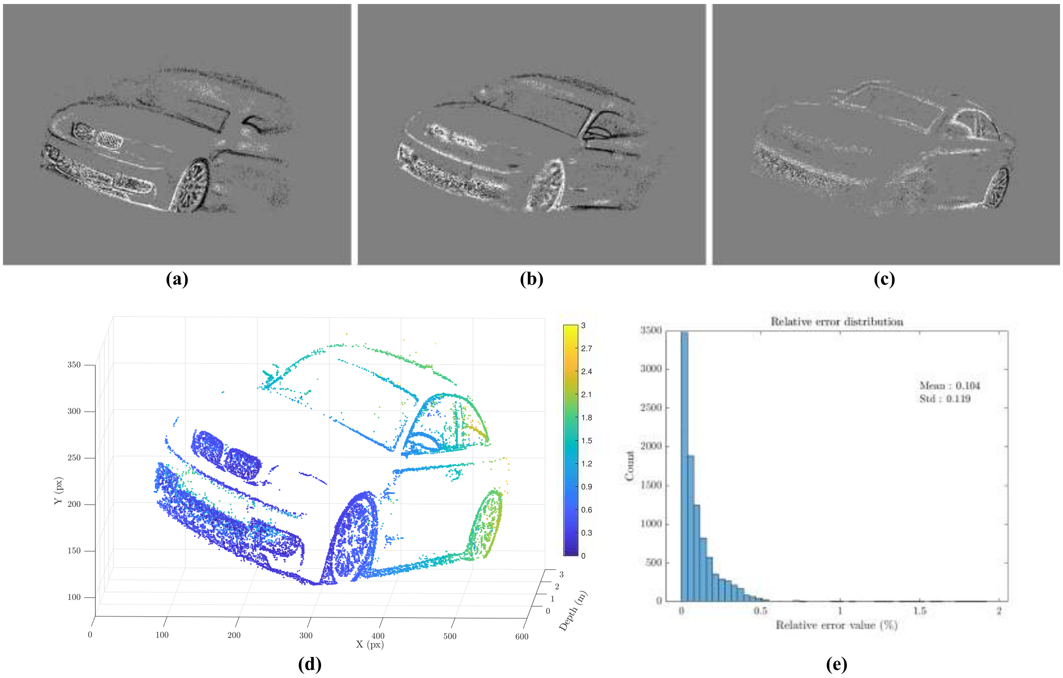


Figure 4. (**a**–**c**) Snapshots during a sweep of an object (**d**) Reconstructed depth scene for the car. The depth is also color-coded for clarity. (**e**) Distribution of the error. The mean relative error is at around 0.1% and a standard deviation of 0.12%.

the membrane potential of the neuron reaches a threshold (spiking threshold, as in Fig. 3(d) and (e)), a spike is generated and the membrane potential is reset to a rest value. For every pixel, five neurons are required. Figure 3a shows events generated by a circle going in and out of focus. At time t_1, the stimulus in front of the receptive field generates a ON event (orange - Fig. 3c). The synaptic weight between the ON and B_{on} neurons is not strong enough to trigger yet the B_{on} neuron (Fig. 3d). As a second spike is generated by the same neuron at time t_2, the B_{on} neuron reaches its threshold value and spikes (Fig. 3d). An inhibition link to the OUT neuron ensures that the OUT neuron won't fire now. After the focus, at time t_3, we have a polarity inversion: the OFF neuron fires, thus exciting the output neuron that fires (Fig. 3e). The next OFF spike, at time t_4, activates the B_{off} neuron, thus preventing the OUT neuron to fire again in response to the future OFF spikes. Finally, the Sync neuron is triggered by the liquid lens, warning that the sweep is over and resetting the OUT neuron to its initial state. The depth can then be extracted as the timing between the OUT and Sync spikes. The neural architecture shown in Fig. 3b is sharing some similarities with the one presented in[55] to measure contrast. However, they are both fundamentally different as one is focusing on measuring spatial contrast with no consideration to the temporal domain, while the other detects the maximum contrast in time and in space, by detecting the shortest duration between polarity changes.

Results

Results are obtained for a field of view of 15° and a depth that ranges from 0.12 to 5.5 m. The distance upper bound corresponds to the hyper-focal distance of the global optical setup. The sparse nature of the data allows the algorithm to operate in real time at the native resolution of the sensor (640 × 480 pixels).

The Spiking Neural Network previously described in section 3 was implemented using the PyNN framework[56] and simulated using the NEST neural simulator[57]. All neurons are modeled as Leaky Integrate-and-Fire (LIF) neurons. Results are presented on Fig. 3. We set the dimension of the network to fit a region of 447 × 447 pixels, the network then using 999045 neurons. This amount is compatible with existing neuromorphic hardware implementation on the TrueNorth platform (1 million neuron[27]) or SpiNNaker capability[26].

To better understand the possibilities and limits of the system, we performed a simulation on synthetic data generated with a controlled optical setup where all parameters can be tuned. The aim of this simulation is to study the algorithm without constraints from the physical setup. Figure 4 shows three snapshots of the events generated during a sweep of a car. Figure 4d shows the reconstructed depth computed by the system.

All the parameters being known we can estimate the relative error to the ground truth. We notice that most of the error is located at the front of the car on the grating where close to one another straight lines patterns are located. This is a known limitation of several vision algorithms such as stereo matching, which will be further discussed in section 3.1. Figure 4(e) displays the error repartition with a mean relative error of 10.4%. An example video on a car is available online[58].

The second experiment, the depth estimated from the DFD is assessed with our setup on a monitored scene where the ground truth is provided by a Microsoft Kinect sensor. The Kinect is taken as the reference similarly to previous studies[59,60], reporting reconstruction precision of few mm at 50 *cm* to 3 *cm* at 3 *m*. Figure 5 shows the setup and the depth map computed for the presented neuromorphic technique with a comparison with the

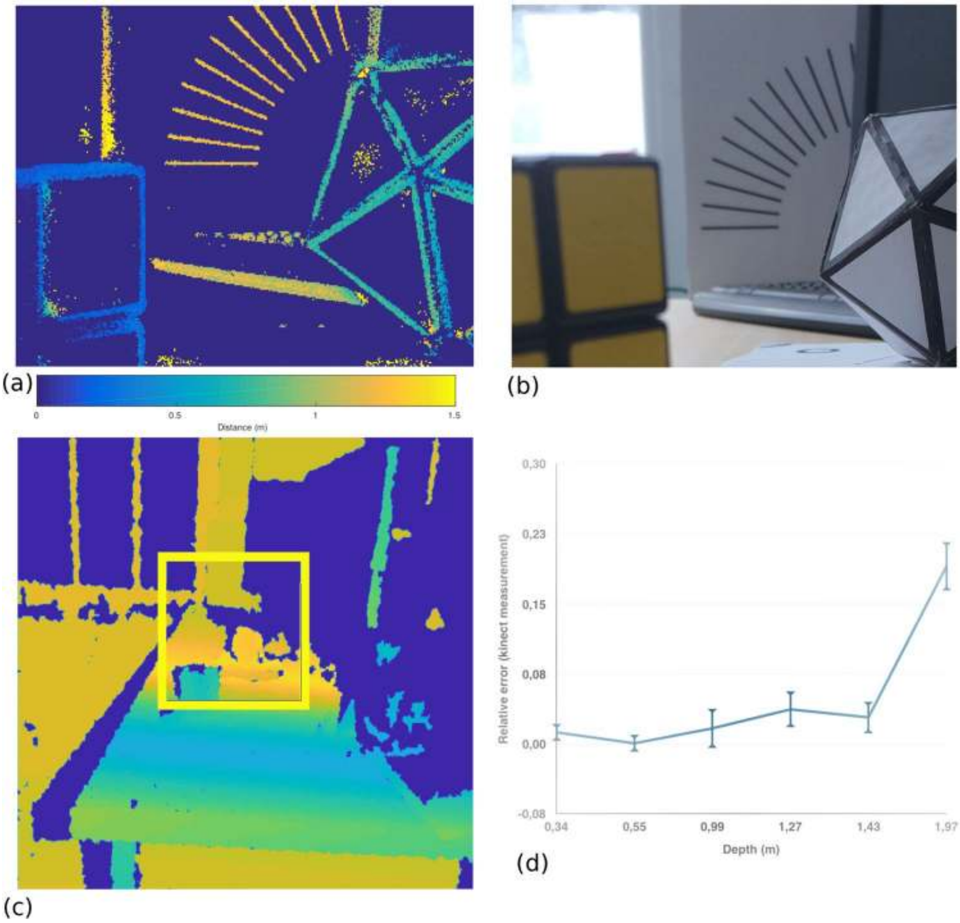


Figure 5. (**a**) Depth map from the developed setup (raw data, no post-processing). (**b**) Conventional Image of scene for display purposes. (**c**) Depth map from the Kinect used as reference. The yellow square corresponds to the field of view. (**d**) Relative error for this scene, with the output of a Microsoft Kinect as ground truth. A sparse set of handpicked points were selected in the ground truth and then compared to depth estimations from our network.

groundtruth depth map: the error is increasing relative to depth. Up to 1.5 m, the relative error is upper-bounded at 4% and increased up to 23% at 2 m. This is however an expected result as the optical system's focal length is reaching the hyper-focal.

The third experiment shows reconstruction for several objects with different textures and sizes. Figure 6 shows for each object its corresponding depth map while the lens is sweeping through the object.

Remarks and limitations. The algorithm is inexpensive in computational power, in the presented experiments it is able to deal with around 15 million events per second. A shaky scene viewed by the event-based camera will generate at most 2 million events per second. In the worst case, thus the current approach is 7 times faster than real time. However for most objects used it is more around 20 times faster than real time using an off-the-shelf laptop. The computational load of the proposed method is lower than any other existing method because it relies on detecting changes of polarities from a sparse output while existing techniques such as[41] require to compute the local gradient on entire frames. This algorithm can be easily embedded on portable devices such as smartphones or autonomous vehicles as an ideal method for low power solutions to obstacle side-stepping or 3D scanners. The low-cost liquid lens used in this paper consumes ~300 mA. New consumer ready products using electrowetting[61] and more advanced research prototypes[62,63] allow a low power budget of less than 1 mW at the cost of losing temporal accuracy.

As pointed out during experiments, repetitive patterns can lead to incorrect depth estimation. Figure 7 shows this situation for simulated data. If we consider two objects that are well separated (Fig. 7f), the sweep of the liquid lens will produce an event stream (Fig. 7l) with non overlapping spikes. Figure 7j is a snapshot of the sweeps' beginning. The four OFF edges are distinct. As the focus evolves, we reach the focus point for object 1 (Fig. 7i). The two edges O_{1L} and O_{1R} of object 1 now generate ON events. After the focus point for object 2 (Fig. 7h), the two other edges O_{2L} and O_{2R} now generate ON events. As the objects are in a sufficient relative distance, the edges O_{1R} and O_{2L} are not overlapping.

If we consider two objects that are close each other (Fig. 7a), the sweep of the liquid lens will now produce the event stream shown in Fig. 7k. As the sweep starts (Fig. 7e), only the external edges of objects 1 and 2 (O_{1L} and

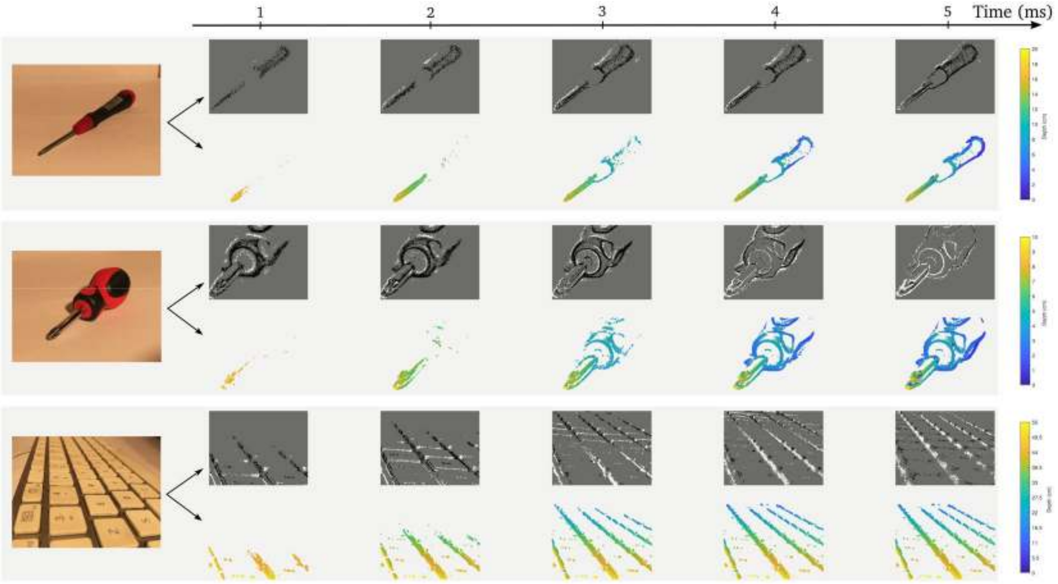


Figure 6. Snapshots of the event stream, and associated depth maps during a sweep (5 ms) for multiple objects. Black and white dots are the OFF and ON events from the event-based silicon retina, as described in Section 2.1. Distance is color-coded.

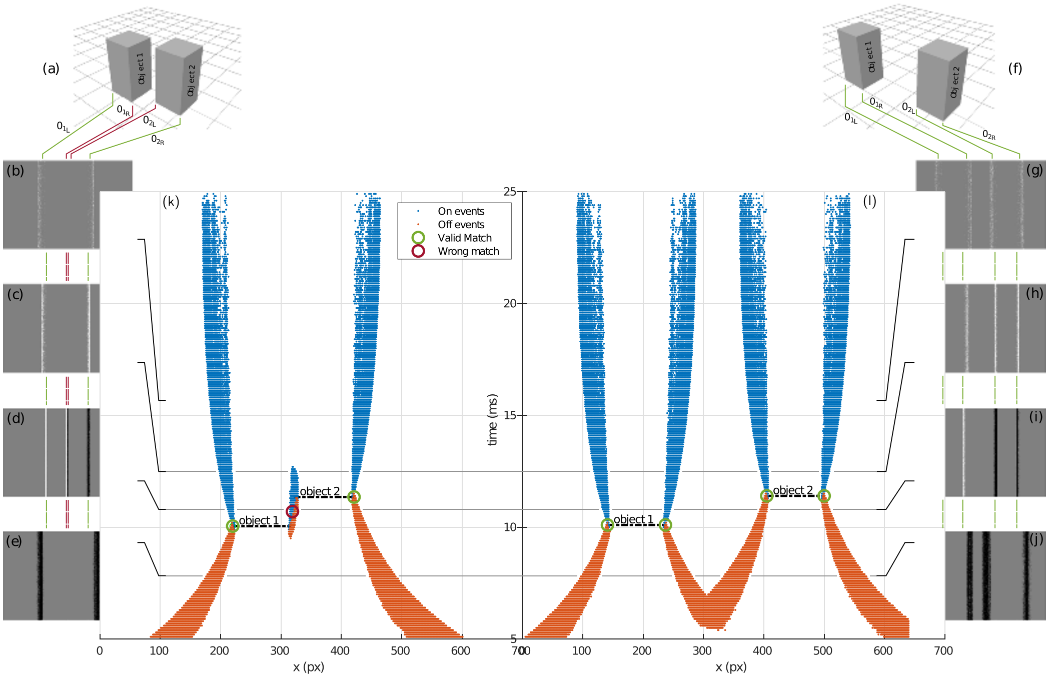


Figure 7. Highlighting of the wrong depth measurements for two closeby edges. The two central plots show events in the x-time plane, smashing the y-dimension. Events are color coded with their polarity (red for OFF events, blue for ON events). The right one is a valid case, with no overlap. The left one contains an overlap in the event stream, leading to wrong depth deductions in this case. 4 snapshots of events are presented for every case.

O_{2R}) generate OFF spikes. As the focus reaches object 1, object 1 generates ON spikes and object 2 OFF spikes. The two middle edges (O_{1R} and O_{2L}) are now superimposed, with two different polarities, causing the failure of the algorithm (Fig. 7d). Decreasing the size of the pixels is equivalent to increase the spatial resolution of the sensor. This will allow to estimate depth as long as we manage to separate the two edges, however the same ambiguity problem will occur once we reached the limit of the sensor. In principle as we are stimulating the same pixel a possible solution to solve this issue is to change the point of view of the camera to disambiguate depth estimation at critical locations.

Conclusions and Discussions

In this paper we proposed a spiking neural network model that solves the depth from focus efficiently by exploiting an event-based representation amenable to neuromorphic hardware implementations. The network operates on visual data in the form of asynchronous events produced by a neuromorphic silicon retina. It processes these address-events in a data-driven manner using artificial spiking neurons computation units. This work introduces a valid explanation and a robust solution to depth estimation from defocus that has not been reported in the literature. The overall system matches recent existing literature of neuroscience, biological retinas and psychophysics studies on the role of defocus in the visual system. This network is nonetheless an abstract simplification of the depth estimation problem that must surely combine more complex information in biological systems. More importantly, this study should be coined depth from focus rather than from defocus as the neural structure developed aims at detecting the exact time of focus during a sweep.

During the last five decades of research, DFD has remained an unsolved issue. The fundamental difference and novelty of this work is that the proposed network operates using exclusively precisely-timed contrast events. These events are measured directly from the neuromorphic silicon retina, which models only the transient responses of retinal cells (i.e., of the Y-ganglion cells), without including the sustained ones, yet present in the system. While the sustained information is present in the silicon retina used, we show that this information is not necessary to provide depth estimation from defocus. Silicon retina transient responses produce single events. Their precise timing plays a crucial role in the estimation of blur and more importantly in determining when the observed object is in focus.

In contrast, the vast majority of computational models of depth from defocus are based on images that are known to be absent from the visual system and only rely on luminance information. Additionally, none of them use the precise timing of spikes. In these models, convolutions techniques are used to determine the level of blur. These methods are computationally expensive and meaningfully slower as several acquisitions are often needed to provide an accurate result. By contrast, the model we presented does not incorporate any notion of filtering or convolutions. These choices are based on the perception of spatial contrast, whereas the presented model solely responds to temporal contrast.

Whether the brain is using such a technique to estimate depth from defocus is an open question. However due to the nature of precisely timed information output by biological retinas[64] convolutions algorithms cannot provide a viable explanation as the stroboscopic nature of image acquisition and luminance use is incompatible with neural systems. Instead, we show that the change of polarity at the pixel level contains sufficient information to estimate depth from defocus. Recent findings in physiology show that several mechanisms used by our methodology exist in Nature. Biological retinas contain several types of ganglion cells, each informing the brain about a particular content of the visual scene, such as motion, edges or chromatic composition. In a recent paper, a newly discovered ganglion cell type 'On-delayed' is described[65]. This cell has been shown to respond vigorously to increasing blur. Its degree of firing directly encodes the amount of high spatial frequencies contained in its receptive field. More importantly, this cell gets input from both ON and OFF polarities. While it is currently unknown how this defocus information is used by the brain, it is most likely that this information projects to the visual thalamus and cortex and also to midbrain structures where accommodation is controlled[66].

We expect the most significant impact of our model to be in the field of artificial vision. Today's machine vision processing systems face severe limitations imposed both by the conventional sensors front-ends (which produce very large amounts of data with fixed sampled frame-rates), and the classical Von Neumann computing architectures (which are affected by the memory bottleneck and require high power and high bandwidths to process continuous streams of images). The emerging field of neuromorphic engineering has produced efficient event-based sensors, that produce low-bandwidth data in continuous time, and powerful parallel computing architectures, that have co-localized memory and computation and can carry out low-latency event-based processing. This technology promises to solve many of the problems associated with conventional computer vision systems. However, the progress so far has been chiefly technological, whereas related development of event-based models and signal processing algorithms has been comparatively lacking (with a few notable exceptions). This work elaborates on an innovative model that can fully exploit the features of event-based visual sensors. In addition, the model can be directly mapped onto existing neuromorphic processing architectures. Results show that the full potential is leveraged when single neurons from the neural network are individually emulated in parallel. In order to emulate the full-scale network, however, efficient neuromorphic hardware device capable of emulating large-scale neural networks are required. The developed architecture requires few neurons per pixel and is implementable on a variety of existing neuromorphic spiking chips such as the SpiNNaker[26], TrueNorth[27] or LOIHI[28] neural chips.

References

1. Held, R. T., Cooper, E. A., O'Brien, J. F. & Banks, M. S. Using blur to affect perceived distance and size. *ACM Transactions on Graphics* **29**, 19:1–16, http://graphics.berkeley.edu/papers/Held-UBA-2010-03/, https://doi.org/10.1145/1731047.1731057 (2010).
2. Vishwanath, D. & Blaser, E. Retinal blur and the perception of egocentric distance. *Journal of Vision* **10**, 26–26 (2010).
3. Gollisch, T. & Meister, M. Rapid neural coding in the retina with relative spike latencies. *Science* **319**, 1108–1111, https://doi.org/10.1126/science.1149639, http://science.sciencemag.org/content/319/5866/1108.full.pdf (2008).
4. Berry, M. J., Warland, D. K. & Meister, M. The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences* **94**, 5411–5416, https://doi.org/10.1073/pnas.94.10.5411, http://www.pnas.org/content/94/10/5411.full.pdf (1997).
5. Liu, R. C., Tzonev, S., Rebrik, S. P. & Miller, K. D. Variability and information in a neural code of the cat lateral geniculate nucleus. *Journal of neurophysiology* **86**(6), 2789–806 (2001).
6. Reinagel, P. & Reid, R. C. Temporal coding of visual information in the thalamus. *Journal of Neuroscience* **20**, 5392–400 (2000).
7. Mainen, Z. & Sejnowski, T. Reliability of spike timing in neocortical neurons. *Science* **268**, 1503–1506 (1995).
8. Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. *Spikes: Exploring the Neural Code*. (MIT Press, Cambridge, MA, USA, 1999).

9. Maass, W. Pulsed neural networks. In Maass, W. & Bishop, C. M. (eds) *Pulsed Neural Networks*, chap. Computing with Spiking Neurons, 55–85 (MIT Press, Cambridge, MA, USA, 1999).
10. Thorpe, S. Spike arrival times: A highly efficient coding scheme for neural networks. *Parallel processing in neural systems* (1990).
11. Thorpe, S. J., Delorme, A. & VanRullen, R. Spike-based strategies for rapid processing. *Neural Networks* **14**, 715–725 (2001).
12. Johansson, R. & Birznieks, I. First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nat Neurosci* **7**, 170–177 (2004).
13. Petersen, R. S., Panzeri, S. & Diamond, M. Population coding of stimulus location in rat somatosensory cortex. *Neuron* **32**, 503–414 (2001).
14. Chicca, E., Stefanini, F., Bartolozzi, C. & Indiveri, G. Neuromorphic electronic circuits for building autonomous cognitive systems. *Proceedings of the IEEE* **102**, 1367–1388 (2014).
15. Neftci, E. e. a. Synthesizing cognition in neuromorphic electronic systems. *Proceedings of the National Academy of Sciences* 3468–3476 (2013).
16. Indiveri, G., Corradi, F. & Qiao, N. Neuromorphic architectures for spiking deep neural networks. *IEEE Electron Devices Meeting (IEDM)* 1–4 (2015).
17. Serrano-Gotarredona, R. E. A. Caviar: A 45 k neuron, 5 m synapse, 12 g connects aer hardware sensory-processing- learning-actuating system for high-speed visual object recognition and tracking. *IEEE Transactions on Neural Networks* 1417–1438 (2009).
18. Posch, C., Matolin, D. & Wohlgenannt, R. High-dr frame-free pwm imaging with asynchronous aer intensity encoding and focal-plane temporal redundancy suppression. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 2430–2433 (IEEE, 2010).
19. Ghosh-Dastidar, S. & Adeli, H. Spiking neural networks. *International journal of neural systems* **19**, 295–308 (2009).
20. Binas, J., Indiveri, G. & Pfeiffer, M. Spiking analog vlsi neuron assemblies as constraint satisfaction problem solvers. In *Circuits and Systems (ISCAS), 2016 IEEE International Symposium on*, 2094–2097 (IEEE, 2016).
21. Mostafa, H., Müller, L. K. & Indiveri, G. An event-based architecture for solving constraint satisfaction problems. *Nature communications* **6**, 8941 (2015).
22. Osswald, M., Ieng, S.-H., Benosman, R. & Indiveri, G. A spiking neural network model of 3d perception for event-based neuromorphic stereo vision systems. *Scientific reports* **7**, 40703 (2017).
23. Dikov, G., Firouzi, M., Röhrbein, F., Conradt, J. & Richter, C. Spiking cooperative stereo-matching at 2 ms latency with neuromorphic hardware. In *Conference on Biomimetic and Biohybrid Systems*, 119–137 (Springer, 2017).
24. Giulioni, M., Lagorce, X., Galluppi, F. & Benosman, R. B. Event-based computation of motion flow on a neuromorphic analog neural platform. *Frontiers in neuroscience* **10** (2016).
25. Haessig, G., Cassidy, A., Alvarez, R., Benosman, R. & Orchard, G. Spiking optical flow for event-based sensors using ibm's truenorth neurosynaptic system. *IEEE Transactions on Biomedical Circuits and Systems* 1–11 (2018).
26. Furber, S., Galluppi, F., Temple, S. & Plana, L. The spinnaker project. *Proceedings of the IEEE* **102**, 652–665 (2014).
27. Merolla, P. A. *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
28. Davies, M. *et al.* Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**, 82–99, https://doi.org/10.1109/MM.2018.112130359. (2018).
29. Qiao, N. *et al.* A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128 k synapses. *Frontiers in neuroscience* **9** (2015).
30. Gaganov, V. & Ignatenko, A. Robust shape from focus via markov random fields. In *Proceedings of Graphicon Conference*, 74–80 (2009).
31. Suwajanakorn, S., Hernandez, C. & Seitz, S. M. Depth from focus with your mobile phone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3497–3506 (2015).
32. Wandell, B. A., El Gamal, A. & Girod, B. Common principles of image acquisition systems and biological vision. *Proceedings of the IEEE* **90**, 5–17 (2002).
33. Pentland, A. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-9**, 523–531 (1985).
34. Moeller, M., Benning, M., Schönlieb, C. & Cremers, D. Variational depth from focus reconstruction. *IEEE Transactions on Image Processing* **24**, 5369–5378 (2015).
35. Zhou, C., Lin, S. & Nayar, S. K. Coded aperture pairs for depth from defocus and defocus deblurring. *International Journal of Computer Vision* **93**, 53–72 (2011).
36. Watanabe, M. & Nayar, S. K. Rational filters for passive depth from defocus. *International Journal of Computer Vision* **27**, 203–225 (1998).
37. Pentland, A., Scherock, S., Darrell, T. & Girod, B. Simple range cameras based on focal error. *JOSA A* **11**, 2925–2934 (1994).
38. Tao, M. W., Hadap, S., Malik, J. & Ramamoorthi, R. Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, 673–680 (2013).
39. Levin, A., Fergus, R., Durand, F. & Freeman, W. T. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)* **26**, 70 (2007).
40. Mateos-Pérez, J. M. *et al.* Comparative evaluation of autofocus algorithms for a real-time system for automatic detection of mycobacterium tuberculosis. *Cytometry Part A* **81**, 213–221 (2012).
41. Martel, J. N., Müller, L. K., Carey, S. J. & Dudek, P. High-speed depth from focus on a programmable vision chip using a focus tunable lens. In *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*, 1150–1153 (IEEE, 2017).
42. Mather, G. Image blur as a pictorial depth cue. *Proc. R. Soc. Lond. B* **263**, 169–172 (1996).
43. Mather, G. & Smith, D. R. Blur discrimination and its relation to blur-mediated depth perception. *Perception* **31**, 1211–1219 (2002).
44. Grant, V. W. Accommodation and convergence in visual space perception. *Journal of Experimental Psychology* **31**, 89 (1942).
45. Nguyen, V. A., Howard, I. P. & Allison, R. S. Detection of the depth order of defocused images. *Vision Research* **45**, 1003–1011 (2005).
46. Fisher, S. K. & Ciuffreda, K. J. Accommodation and apparent distance. *Perception* **17**, 609–621 (1988).
47. Ciuffreda, K. J. Why two eyes. *Journal of Behavioral Optometry* **13**, 35–7 (2002).
48. Ciuffreda, K. J. & Engber, K. Is one eye better than two when viewing pictorial art? *Leonardo* **35**, 37–40 (2002).
49. Mather, G. The use of image blur as a depth cue. *Perception* **26**, 1147–1158 (1997).
50. Mather, G. & Smith, D. R. Depth cue integration: stereopsis and image blur. *Vision research* **40**, 3501–3506 (2000).
51. Mather, G. & Smith, D. R. Combining depth cues: effects upon accuracy and speed of performance in a depth-ordering task. *Vision research* **44**, 557–562 (2004).
52. Lin, H.-Y. & Chang, C.-H. Depth recovery from motion and defocus blur. *Image Analysis and Recognition* 122–133 (2006).
53. Blum, M., Büeler, M., Grätzel, C. & Aschwanden, M. Compact optical design solutions using focus tunable lenses. In *SPIE Optical Systems Design*, 81670W–81670W (International Society for Optics and Photonics, 2011).
54. Lapicque, L. Recherches quatitatives sur l'excitation electrique des nerfs traitee comme polarisation. *J. Physiol. Pathol. Gen.* **9**, 620–635 (1907).
55. Neumann, H., Pessoa, L. & Hanse, T. Interaction of on and off pathways for visual contrast measurement. *Biological cybernetics* **81**, 515–532 (1999).

56. Davison, A. P. *et al*. Pynn: a common interface for neuronal network simulators. *Frontiers in neuroinformatics* **2** (2008).
57. Gewaltig, M.-O. & Diesmann, M. Nest (neural simulation tool). *Scholarpedia* **2**, 1430 (2007).
58. Haessig, G. & Berthelon, X. https://youtu.be/ia5gfvln0ay (2017).
59. Khoshelham, K. Accuracy analysis of kinect depth data. In *ISPRS workshop laser scanning*, 133–138 (2011).
60. Macknojia, R., Chávez-Aragón, A., Payeur, P. & Laganière, R. Experimental characterization of two generations of kinect's depth sensors. In *Robotic and Sensors Environments (ROSE), 2012 IEEE International Symposium on*, 150–155 (IEEE, 2012).
61. Berge, B. Liquid lens technology: principle of electrowetting based lenses and applications to imaging. In *Micro Electro Mechanical Systems, 2005. MEMS 2005. 18th IEEE International Conference on*, 227–230 (IEEE, 2005).
62. Hendriks, B., Kuiper, S., As, M. V., Renders, C. & Tukker, T. Electrowetting-based variable-focus lens for miniature systems. *Optical review* **12**, 255–259 (2005).
63. Wei, X., Kawamura, G., Muto, H. & Matsuda, A. Fabrication on low voltage driven electrowetting liquid lens by dip coating processes. *Thin Solid Films* **608**, 16–20 (2016).
64. Gollisch, T. & Meister, M. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron* **65**, 150–164 (2010).
65. Mani, A. & Schwartz, G. Circuit mechanisms of a retinal ganglion cell with stimulus-dependent response latency and activation beyond its dendrites. *Curr. Biol.* **27**, 471–482 (2017).
66. Baden, T., Schaeffel, F. & Berens, P. Visual neuroscience: A retinal ganglion cell to report image focus? *Curr. Biol.* **27**, 138–141 (2017).

Author Contributions

G.H. and X.B. wrote the main manuscript text and created all the figures in the manuscript. They also contributed to the conception of the experiments and the data analysis. All authors contributed to manuscript revision, read and approved the submitted version.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-40064-0.

Competing Interests: The authors declare no competing interests.