Vergence Control with a Neuromorphic iCub

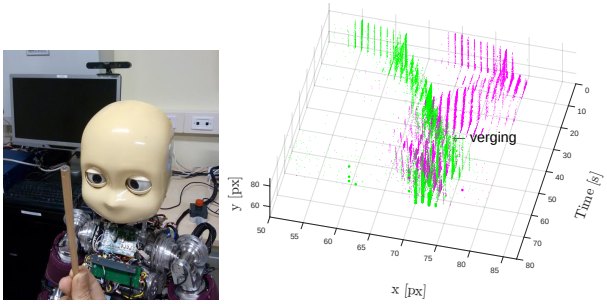
Valentina Vasco, *Student Member, IEEE*, Arren Glover, *Member, IEEE*, Yeshasvi Tirupachuri, Fabio Solari, Manuela Chessa, and Chiara Bartolozzi, *Member, IEEE*

***Abstract*— Vergence control and tracking allow a robot to maintain an accurate estimate of a dynamic object three dimensions, improving depth estimation at the fixation point. Brain-inspired implementations of vergence control are based on models of complex binocular cells of the visual cortex sensitive to disparity. The energy of cells activation provides a disparity-related signal that can be reliably used for vergence control. We implemented such a model on the neuromorphic iCub, equipped with a pair of brain inspired vision sensors. Such sensors provide low-latency, compressed and high temporal resolution visual information related to changes in the scene. We demonstrate the feasibility of a fully neuromorphic system for vergence control and show that this implementation works in real-time, providing fast and accurate control for a moving stimulus up to 2 Hz, sensibly decreasing the latency associated to frame-based cameras. Additionally, thanks to the high dynamic range of the sensor, the control shows the same accuracy under very different illumination.**

I. INTRODUCTION

Binocular, or stereo, vision is an anthropomorphic method to estimate the distance of objects, or depth, in the three dimensional space. Depth estimation is essential for successfully interacting with the environment, for example to avoid obstacles during navigation, or to plan a correct grasp for object manipulation.

Active vergence movements that put a target object in the fovea of both eyes, naturally performed by humans and primates, can improve depth accuracy by reducing visual ambiguity [1]. It has been suggested that stereo vergence might occur as a fast, reflexive action directly driven by the activity of disparity-sensitive cells of the visual cortex [2], [3]. In the corresponding model, vergence simply occurs as an inverted response to cells sensitive to stereo disparity rather than from high-level depth estimation, sensibly reducing complexity and the cost associated with the computation of the full depth map. Indeed, the processes of estimating disparity and of controlling vergence are carried out in parallel by two different neural mechanisms [4], and few works in literature address both tasks (e.g., see [5]). The vergence control can be approached by considering learning strategies: in [6], a reinforcement learning framework is used to learn a time-constrained closed-loop control rule; in [7], the authors propose a learned sensory representation that guides vergence movements by exploiting a biological and unsupervised reward; a binocular vergence control based on attention-gated reinforcement learning is proposed in [8], where the control policy is implemented by a neural network. More recently, a neural architecture, trained by using a particle swarm optimization, is described in [9]; and finally, a specific neural network is used for learning and performing sensory-sensory and sensory-motor transformations in [10]. A different approach is to impose a desired behavior of the vergence control [11]: in this way it is possible to exploit all the neural resources, but there is the need of completely knowing them. In this paper, we consider such an approach: indeed, this model was shown to work reliably on the humanoid robot iCub, that was able to continuously verge on a moving target producing an accurate depth estimation [12]. The performance of this model, based on image acquisition from traditional frame-based cameras, suffers from latency up to 1 s, that produces a lag in the vergence movements, that fails when tracking fast moving objects. To improve speed and performance, and to perform a neuromorphic model on neuromorphic hardware, we adapted this model to work with the biologically inspired dynamic vision sensors (DVS) on the humanoid iCub robot. DVS sensors [13] operate more similarly to a biological eye, in comparison to the standard frame-based cameras, producing a spiking response (events) only when the light falling on a pixel changes, e.g. when an object or the sensor itself moves. Contrary to frame-based sensors, where a full frame is acquired at given times, in these sensors only active pixels send their information as soon as it is sensed and, therefore, they can offer low-



(a) (b)

Fig. 1: Neuromorphic vergence control: (a) the iCub robot equipped with DVS sensors verging on a pencil; (b) Events from the DVS in response to the pen, in the three-dimensional (x, y, t) space. Events from left (green) and right (magenta) camera overlap when the vergence movement is performed.

V.Vasco, A.Glover and C.Bartolozzi are with the iCub Facility, Istituto Italiano di Tecnologia, Italy. {valentina.vasco, arren.glover, chiara.bartolozzi}@iit.it

Y.Tirupachuri is with RCBS, Istituto Italiano di Tecnologia, Italy. {yeshasvi.tirupachuri}@iit.it

F.Solari and M.Chessa are with the Universitá di Genova. {fabio.solari, manuela.chessa}@unige.it

latency, high temporal precision and low redundancy visual information together with the possibility of low computation and power requirements.

We show that the neuromorphic model of vergence control implemented with neuromorphic sensors achieves robust vergence control in real-time, with only 200 *ms* latency, that allows to reliably track objects moving at up to 2 *Hz* in the direction of the robot, in a wide range of illumination.

A. Stereo Vision with Event-driven Cameras

Traditional computer vision methods for stereo matching and depth estimation have been adapted to event-driven camera data, improving the state-of-the-art over the last decade [14]–[18].

In the the "co-operative" approach proposed by [18], a dynamic cooperative neural network is designed to extract a global spatio-temporal correlation for input events: sets of inhibitory and excitatory units, created by the intersection of pixel pairs in the left and right images, interact with each other based on temporal correlation of stereo-events and physical constraints. All the other proposed algorithms attempt to exploit the high temporal precision of the DVS, performing one-to-one matching of events between the left and right camera that occur within a very short time window [14], [15]. However, due to jitter in event timing, as well as mismatch between the pair of sensors, temporal precision alone was found not to be accurate enough. Further constraints such as polarity, ordering and epipolar restrictions improved performance [16]. Finally, to further increase the discrimination of stereo event-matching, appearance features in the form of edge-orientations were implemented to ensure geometric consistency in a neighbourhood around the events [17]. The orientation was estimated using a bank of biologically plausible Gabor filters that respond to oriented edges. The resulting implementation was validated by measuring the depth of three objects [17], given the stationary stereo set-up was properly calibrated.

Other implementations of Gabor filters have been used with event-driven sensory input for various tasks such as feature tracking [19], velocity estimation [20] and feature matching [14], [15]. In the event-driven context, filter convolution can usually be performed cheaply, since it only needs to occur in a spatio-temporal window where events exist, and in terms of computational complexity, the biological models are well suited to the biologically inspired sensor data.

B. Vergence Control with Binocular Gabor Filters and Phase-Shift Model

Gabor filters were initially proposed as a model for the response of neurons to differently oriented stimuli in the visual cortex. Physiological studies [2], [3] have suggested that binocular neurons with a phase-shift across the left and right portions of their receptive field respond to stimuli at different disparities. This effect can be modelled using a bank of Gabor filters with a phase-shift between left and right components of each filter. Fig. 2 shows the gain of three different Gabor filters tuned to different disparities, obtained by shifting the phase of the right Gabor component with respect to the left, whose phase is set to 0.

The minimisation of the average energy level of such a bank of phase-shifted Gabor filters directly controls the vergence of stereo cameras, without the explicit calculation of the disparity of the scene. The depth estimation of the object can then be calculated through the geometry of the relative pose between cameras [21], as opposed to relying on exact matching of pixels in visual space. Vergence control was shown to operate with a 0.5 Hz stimulus and an image normalisation method was introduced to allow operation in a variety of lighting conditions.

C. Event-driven implementation of the neuromorphic vergence control

In this paper we present for the first time an event-driven implementation of the neuromorphic model of vergence control based on disparity tuned binocular Gabor filters driven by the events generated by the neuromorphic dynamic vision sensors. Our work is relevant because we demonstrate the feasibility of a fully neuromorphic system for vergence control. The implementation of vergence eye movements is then useful towards salient objects and depth estimation in a system where the stereo setup moves continuously. The advantages of compressive and low-latency data acquisition of event-driven cameras lead to a decrease in control latency, crucial for robots interacting in real-time with the environment and the high dynamic range of the sensor lead the control to be independent on illumination changes.

II. EVENT-DRIVEN PHASE SHIFT MODEL

Vergence control model based on Gabor filters with binocular phase-shift was originally used with standard frame-based cameras. We describe how the disparity models are applied to event-driven data and the vergence controller itself.

A. Binocular Gabor Filters with a Phase-shift

The Gabor filter is composed of a Gaussian kernel function modulated in the spatial domain with a sinusoidal wave and responds strongly to object edges. A rotated filter responds to an edge with a specific spatial orientation, θ. A binocular Gabor filter is applied to both left and right images simultaneously with a phase shift, ψ, added to the spatial frequency of the right image convolution. The maximum response therefore occurs if the left and right images have edges that are offset by an amount corresponding to the phase shift applied, i.e. the binocular Gabor filter responds to a given stereo disparity (Fig. 2).

We follow [12] in the implementation of the complex-valued binocular Gabor filters, g, for each visual stimulus in pixel position, (x,y):

$$g(x,y,\theta,\psi) = e^{-\frac{{x_\theta}^2+{y_\theta}^2}{2\sigma^2}} e^{j(2\pi f_s x_\theta+\psi)}, \qquad (1)$$

where (x_θ, y_θ) is the pixel location rotated around the filter centre by θ, σ is standard deviation of the Gaussian kernel,

and f_s is the spatial frequency of the sinusoidal component. We chose to shift the right component with respect to the left one, which therefore is associated to a phase $\psi = 0$. Convolving the Gabor filter with visual data results in a complex number with a real and imaginary component, which represents a quadrature pair of binocular simple cells. Each event contributes to the energy level of the single filter with orientation θ_i and phase shift ψ_j, depending whether it comes from left or right camera, as follows:

$$\begin{cases} r_{ij} = r_{ij} + g(x_L, y_L, \theta_i, 0) & \textbf{if} \quad e_L = (x_L, y_L, t) \\ r_{ij} = r_{ij} + g(x_R, y_R, \theta_i, \psi_j) & \textbf{if} \quad e_R = (x_R, y_R, t). \end{cases} \tag{2}$$

The energy level e_{ij} of the single filter is then computed by summing up the squared responses of the quadrature pair [22], as follows:

$$e_{ij} = \mathrm{Re}[r_{ij}]^2 + \mathrm{Im}[r_{ij}]^2. \tag{3}$$

A filter's response will be maximum when the cell's binocular phase difference matches the disparity of the stimulus, δ, according to:

$$\delta = \frac{\Delta\psi}{2\pi f_s}. \tag{4}$$

The filters were applied to stereo visual data in [12] by instantiating a bank of filters in the centre of a region-of-interest (ROI) of the image with a variety of 5 orientations and 7 disparities. Each filter was convolved with all pixels in the ROI for both the left and right (with phase shift) camera images taken from a stereo camera pair and resulted in a scalar response from each of the instantiated filters. The response of the filter can be used to estimate the disparity in the images produced by the stereo pair [22], [23], however vergence was achieved by directly controlling the desired value of the filter responses.

B. Frame-based v.s. Event-based Cameras

A frame-based camera produces a two-dimensional image corresponding to the amount of light falling onto each photo-sensitive sensor, and does so at a set frame-rate, typically 10-30 Hz for robotic applications. The images are convolved with the Gabor filter by multiplying the pixel intensity by the filter gain associated with the pixel position and summing the value for each pixel in the image (or sub-region in the image).

An event-based camera produces an asynchronous, continuous stream of single pixel 'events' that occur when the change in light falling on the photo-sensitive sensor changes beyond a threshold. The event ($e_{x,y,p,t}$) is defined by its visual position (x, y), the direction/polarity of light change (p), and is time-stamped to microsecond resolution (t). Fig. 1b shows an event stream in three-dimensional (x, y, t) space.

The convolution of the event-based visual information with the binocular Gabor filters cannot be performed as with an image and must be modified to accommodate the event-based camera. However, the event-based camera offers fast

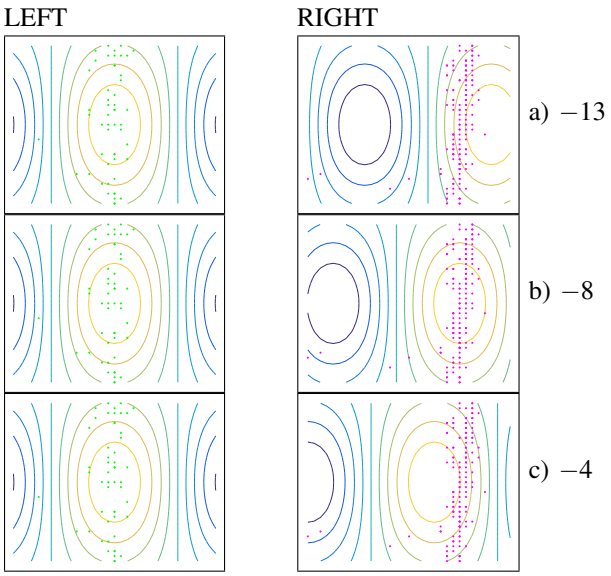


Fig. 2: Left (first column) and right (second column) receptive field for filters tuned at -13 (a), -8 (b) and -4 *px* (c). Left and right events are shown by green (first column) and magenta dots (second column) respectively. The central pair of filters (b) matches the stimulus disparity and produces the highest response. The real part of left and right receptive field is shown.

convolution operations by performing incremental updates to filter responses as the event-steam occurs rather than performing the full convolution at each pixel location as is required with an image. Typically an event-based algorithm can respond with lower latency as output occurs after a single pixel event, and does not require a full 'frame' of pixels to be extracted from the camera.

C. Event-based Phase-shift Model

An incoming event can be convolved with the filter at its visual position in the same way that an individual pixel from an image is convolved by considering the intensity of the event to be a binary value. A single event does not hold enough information for the convolution to be relevant and therefore an accumulation of the convolutions from a set of events is required. A commonly used biological model for accumulating and decaying energy is the leaky-integrate-and-fire model [19], where the activity of the filter is updated at each incoming event, following an exponential decay function. However a model dependent on a fixed temporal parameter [24], [25] is highly dependent on the speed of the object's motion, since it puts a lower bound on the measurable velocity. In addition, the disparity signal becomes weak when the object stops moving as the event-driven camera stops producing a signal.

Instead, for each incoming event, we consider a fixed window of the most recent N events such that of the pixel locations in a central region-of-interest, only a fixed percentage can be active at any given time (green and magenta dots in Fig. 2). Within the region-of-interest (ROI) we force both

high and low responses to produce a signal with a strong contrast for vergence.

Given such a set of events, we use an incremental approach in which as an event occurs in the ROI, the small convolution operation is performed and incremented to the current energy of the filter. Simultaneously we remove the oldest event from the fixed window, and remove the associated convolution energy. The energy of the filter is therefore constantly updated given the N events in the filter without having to perform the full convolution operation required by a frame-based counter-part.

A fixed window of events is sensitive to the amount of visual contrast presented by the stimulus. However, as we use a ROI (and not the full scene) and we assume we need a minimum number of events for vergence to occur, a fixed number of events is suited to the task and is reasonably robust despite the stimulus used. A fixed number of events also provides a constant verging signal for a stationary stimulus.

The incremental filter response is continuously calculated driven by incoming events according to Algorithm 1. At any point in time, the response state of the filter can be observed by applying Eq. 3.

Algorithm 1 Event-based Incremental Gabor Convolution

Input: $e_{in}\{x, y, p, ts\}$
Add each event to the window and pop the oldest
$W \leftarrow e_{in}$
if $W.size() > N$ **then**
  $e_{out} \leftarrow W.pop()$
end if
For each filter do a single addition and subtraction convolution
for $i \in$ orientations **do**
  for $j \in$ phases **do**
    if e_L **then**
      $r_{ij} \leftarrow r_{ij} + g(x_L, y_L, \theta_i, 0)$
    end if
    if e_R **then**
      $r_{ij} \leftarrow r_{ij} + g(x_L, y_L, \theta_i, \psi_j)$
    end if
    if e_{out_L} **then**
      $r_{ij} \leftarrow r_{ij} - g(x_L, y_L, \theta_i, 0)$
    end if
    if e_{out_R} **then**
      $r_{ij} \leftarrow r_{ij} - g(x_R, y_R, \theta_i, \psi_j)$
    end if
    $e_{ij} = \mathrm{Re}[r_{ij}]^2 + \mathrm{Im}[r_{ij}]^2$
  end for
end for

D. Vergence Controller

Following [12], the vergence controller uses the energy of the filters directly to produce the controller error signal. Although a decoding phase has been proposed for the extraction of the disparity from the population responses [22], [23], the explicit disparity of the scene is not needed [11], [12], as also proposed in humans [26].

The controller is implemented using the weighting methodology proposed in [12], in which each filter response contributes to the final control signal according to an individual weighting value. A proportional velocity controller is implemented, in which the control signal $\dot{v}$ is obtained by normalising for the total energies of the filter as:

$$\dot{v} = k_p \frac{\sum_{i=1}^{N_\theta} \sum_{j=1}^{N_\psi} w_{ij} e_{ij}}{N_\theta N_\psi \sum_{i=1}^{N_\theta} \sum_{j=1}^{N_\psi} e_{ij}}, \tag{5}$$

where w_{ij} is the set of relative filter weights, k_p is the overall gain of the controller, N_θ is the number of orientations and N_ψ the number of phase-shifts. The weights w_{ij} are set to be positive, negative or zero according to the disparity of the tuned filters. Such configuration allows the filter responses of positively and negatively tuned filters to negate each other during correct vergence, resulting in zero (or very small) vergence velocity. Vergence therefore occurs when the response of positively (or negatively) tuned filters outweighs others. An example is shown in Fig. 3, with only one positively (red line) and one negatively (blue line) tuned filter, along with the resulting control signal.

III. EXPERIMENTS AND RESULTS

In comparison to a frame-based camera, the event-based camera produced data in the temporal domain differently. We thus are interested in evaluating the temporal response of the filters and vergence controller when using the event-based approach. The cameras also respond to visual stimulus in a spatially different manner (i.e. only responding to edges) and will therefore produce a different filter response to a highly textured object. Finally the cameras have a much higher dynamic range than a standard camera and, while [11] implemented a specific computational layer to normalise for lighting conditions, we evaluate the inherent hardware-based invariance to illumination.

The presented algorithm was implemented and tested on the neuromorphic iCub head [27], equipped with a pair of dynamic vision sensors (DVS) acting as stereo visual system, with a total of 6 Degrees of Freedom (DOFs). In order to quantitatively validate the experimental results, the ground truth depth was measured using a 3D sensor device, placed behind the iCub head, as shown in Fig. 1a, to capture the depth of the stimulus relative to the head.

For the described experiments, we only controlled the vergence angle, which can be set within a range of [0° - 50°]. We empirically selected the following parameters: a set of filters with $N_\theta = 5$ orientations, $N_\psi = 7$ phase-shifts, $\sigma = 6\ px$ and $f_s = 0.02\ 1/px$; a window W of $N = 300$ events, a ROI of $37 \times 37\ px$ and $k_p = 5000$.

A. Experiment 1: Step input

In the first experiment, the stimulus was placed at a depth of 300 *mm* and the fixation point at 400 *mm*, corresponding to a vergence angle of 20°. As shown in Fig. 4, when the trigger

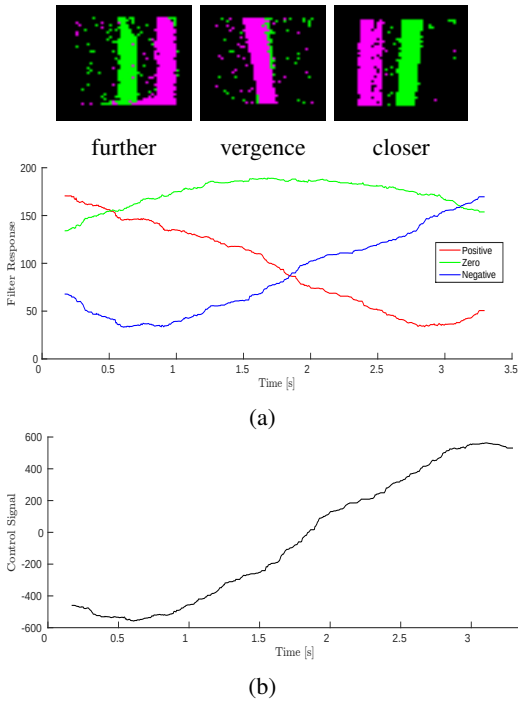


Fig. 3: Example disparities as a stimulus is at the correct vergence point (middle), or further (left) or closer (right) to the robot (the left camera is shown in green and the right camera is shown in magenta). The corresponding filter responses during a smooth transition towards the robot (a) and the resulting relative control signal (b). The control signal (b) is negative / positive for objects further / closer than the fixation point, leading the eyes to diverge / converge respectively.

signal activates the algorithm (black dashed line), the fixation point depth converges to the ground truth within ~ 200 *ms*, decreasing considerably the 1 *s* frame-based latency.

B. Experiment 2: Sinusoidal input

In the second experiment, the stimulus oscillates at three different speeds, with a frequency of approximately $0.5, 1.25$ and 2 Hz and an amplitude that varies between 250 and 500 *mm*, corresponding to a change in the vergence angle from 23.7° to 12.4°.

In the related work [12], the control yields an effective tracking in depth of the stimulus, but several limitations can be observed from the reported results:

- there is a temporal delay of ~ 0.5 *s* between the computed depth and the ground truth;
- the highest frequency reached is 0.5 Hz;
- the tracking accuracy deteriorates with increasing frequency.

The event-driven algorithm takes advantage of the sensors' microsecond temporal resolution to achieve a faster and more accurate tracking: when the stimulus moves slowly, at a frequency comparable to [12] (0.5 Hz in Fig. 5b), the temporal delay is removed and the tracking results in a more precise movement; the same accuracy is kept at higher

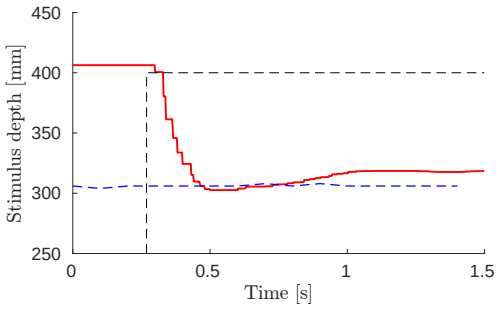


Fig. 4: The response of the depth estimate (solid red) given a stimulus step input (dotted blue) of approximately 100 mm, in which vergence began at the dashed grey line. The rise time is less than 250 ms.

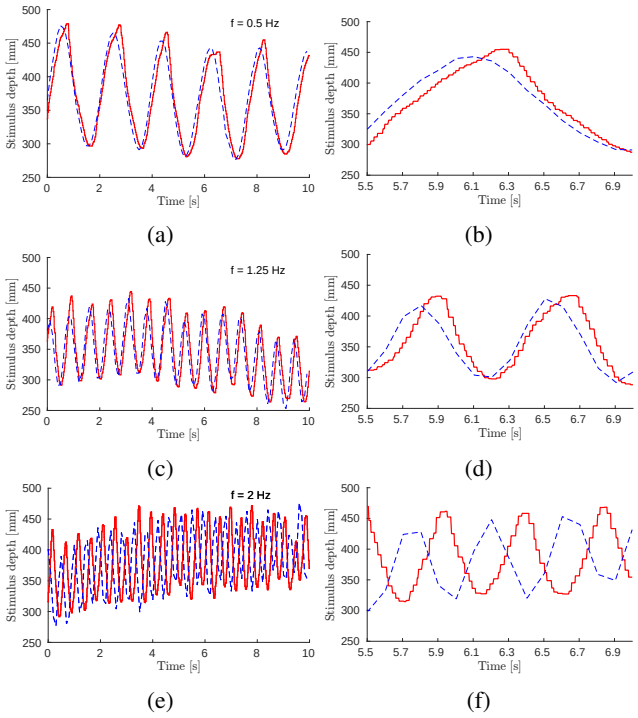


Fig. 5: The response of the depth estimate (red line) compared to the ground truth depth (dashed blue line) at various frequencies: 0.5 (a), 1.25 (c) and 2 Hz (e), with a smaller time scale of the same data shown in (b), (d) and (f) for the respective frequencies.

frequency (Fig. 5d), up to 1.25 Hz. The overshoot that is observable in Figs. 5b and 5d is within the characterised latency of 200 *ms*. The precision of the control decreases at higher speeds (Fig. 5f), still achieving the correct movement, but with a approximate mean latency of ~ 200 *ms*.

C. Experiment 3: Illumination change

The experiment was repeated by setting three luminous powers, $13410, 3450$ and 48 *lm*, the first two corresponding to normal office lighting conditions and the last to dark environment. In the frame-based approach, the luminous power strongly affected the control gain, producing wide

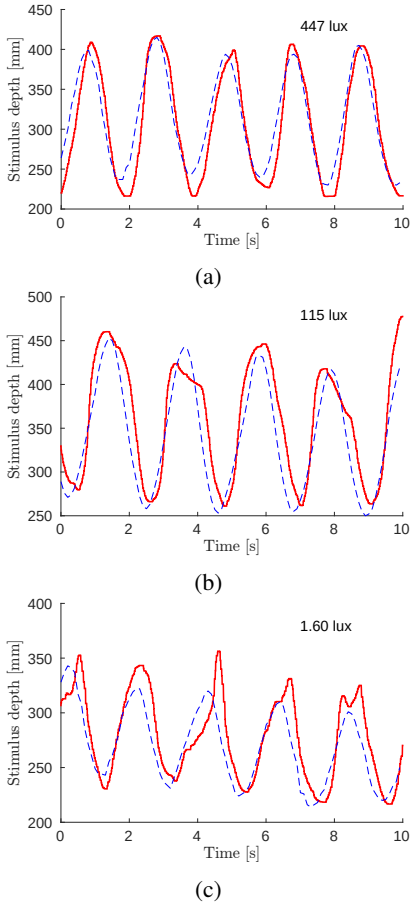


Fig. 6: The response of the depth estimate (red line) compared to the ground truth depth (dashed blue line) in various lighting conditions: high (a), medium (b) and no light (c).

oscillations around the depth of the stimulus with increasing luminosity. An extra normalization stage was therefore necessary to achieve an effective and stable control in various lighting conditions. Event-driven sensors have instead a wide dynamic range (120 dB) and reliably provide events from up to 1 *klx* down to less than 0.1 *lux*. Therefore, as shown in Fig. 6, the performance of the event-driven algorithm is still reliable when the light is changing, removing the need for the extra layer. The vergence control performs the correct movement even in a completely dark environment (Fig. 6c), reaching a much smaller level than the minimum value of 4400 *lm* achieved with frame-based cameras.

D. Experiment 4: Textured object

In the last experiment, a soft toy object (Fig.7a) was moved in front of the iCub. The control was able to perform the correct movement for a slow motion, but the performance deteriorated with velocity, as shown in Figs.7b and 7c. A textured object, indeed, generally produces more events than a single edge stimulus moving at the same speed. Therefore the textured object moving at higher speed produced many events in a variety of locations in the ROI. The resulting control signal is less clean, and has an overall lower magnitude than with the previous stimulus. Moreover, the left receptive field was centered in the middle of the region of interest, therefore the algorithm implicitly assumed the object to be in the center. While this was easy to guarantee for an edge stimulus, there was no certainty that a textured object produced a response in the middle of left camera, which caused the filter response to be lower.

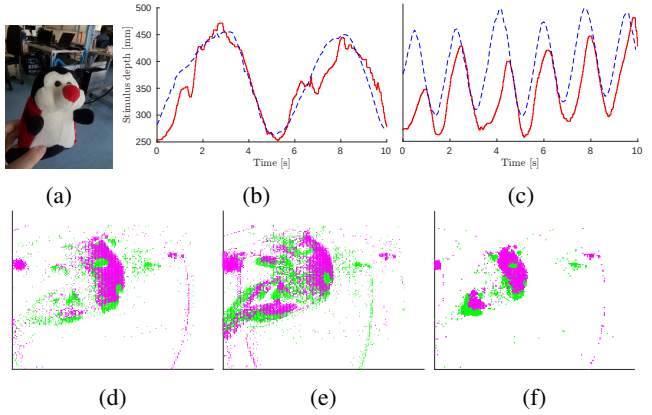


Fig. 7: Vergence on the toy (a) for 0.2 Hz (b) and 0.6 Hz (c) motions (red) compared to ground truth (dotted blue). Three examples are shown after vergence occurs (d), (e) and (f), respectively.

The effectiveness of the slow motion control was qualitatively evaluated by looking at instances of left and right events that clearly overlapped after that vergence was performed (Fig. 7d, 7e and 7f).

IV. DISCUSSION

The fixation range of the neuromorphic iCub was approximately 250 *mm* to 450 *mm*, which was closer than [12] at 600 *mm* to 1200 *mm*. At smaller depths the vergence angle is much larger and the change in vergence angle required by the controller is much larger for a given depth error, compared to a stimulus further away. The lower-latency and faster response of the event-based vergence controller makes it more suited to larger errors, and therefore closer depths. However, the low resolution of the DVS camera (128×128 pixels) results in a larger pixel quantisation error and therefore a larger disparity error proportional to the depth of the stimulus. The vergence range of the event-based controller is limited by this resolution; however, higher resolution event-based cameras are becoming increasingly common.

Experiments were conducted with a plain background that removed extraneous textures and distractions from the task of verging on the exact stimulus that was also measured by the ground-truth. The background used in [12] was, instead, somewhat textured, however it was difficult to assess how the texture influenced the algorithm. A plain background was used as we separate the problems of attention (i.e. what stimulus to verge upon) and the performance of verging itself. It should be noted that when removing the back-drop, the controller successfully verged on the wall behind (approx.

4 *m* away), seen as the left and right event visualisations were aligned. However, the disparity error due to pixel quantisation is too large for analysis at these distances.

V. CONCLUSIONS AND FUTURE WORK

In this work, we have described a method for applying neurmorphic models for real-time vergence control to neuromorphic sensors and tested it on the neuromorphic iCub, equipped with two stereo DVS sensors. As opposed to frame-based approach, where the filter bank is applied to the entire frame at a set frame-rate regardless of any position change, in the event-based algorithm the filter convolution can be performed cheaply only when events occur.

We showed that the vergence response is faster using the event-driven cameras, allowing the tracking in depth of the stimulus up to 2 *Hz*, with a 200 *ms* latency, as opposed to 1 *s* frame-based. Moreover the DVS wide dynamic range guaranteed the independence on the luminance, allowing the tracking in both normal and dark environment and eliminating the need of an extra computational layer.

The control was effective on a slow textured object, but the performance deteriorated for faster motion, due to the increased number of events that slowed down the control and the assumption of the object being in the middle of the region of interest that was not held. Possibly phase shifts can be added in the left camera events to produce a better vergence signal, or in combination with a horizontal gaze controller.

Vergence and tracking of an object of interest allows a robot to maintain an accurate estimate of a dynamic object in three dimensions. We plan to use such 3D position to allow the event-driven iCub to grasp objects in the environment and dynamically interact with it.

REFERENCES

[1] M. Mon-Williams, J. R. Tresilian, and A. Roberts, "Vergence provides veridical depth perception from horizontal retinal image disparities," *Experimental brain research*, vol. 133, no. 3, pp. 407–413, 2000.

[2] R. L. De Valois, N. P. Cottaris, L. E. Mahon, S. D. Elfar, and J. A. Wilson, "Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity." *Vision research*, vol. 40, no. 27, pp. 3685–3702, 2000.

[3] D. J. Fleet, H. Wagner, and D. J. Heeger, "Neural encoding of binocular disparity: energy models, positionshifts and phase shifts," *Vision Res.*, vol. 36, no. 12, pp. 1839–1857, 1996.

[4] L. M. Wilcox and R. S. Allison, "Coarse-fine dichotomies in human stereopsis," *Vision Research*, vol. 49, no. 22, pp. 2653 – 2665, 2009.

[5] Y. Zhao, C. A. Rothkopf, J. Triesch, and B. E. Shi, "A unified model of the joint development of disparity selectivity and vergence control," in *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. IEEE, 2012, pp. 1–6.

[6] J. H. Piater, R. A. Grupen, and K. Ramamritham, "Learning real-time stereo vergence control," in *Intelligent Control/Intelligent Systems and Semiotics, 1999. Proceedings of the 1999 IEEE International Symposium on*, 1999, pp. 272–277.

[7] A. Franz and J. Triesch, "Emergence of disparity tuning during the development of vergence eye movements," in *2007 IEEE 6th International Conference on Development and Learning*, 2007, pp. 31–36.

[8] Y. Wang and B. Shi, "Improved binocular vergence control via a neural network that maximizes an internally defined reward," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 3, pp. 247–256, 2011.

[9] A. Gibaldi, A. Canessa, M. Chessa, F. Solari, and S. P. Sabatini, "How a population-based representation of binocular visual signal can intrinsically mediate autonomous learning of vergence control," *Procedia Computer Science*, vol. 13, pp. 212 – 221, 2012.

[10] W. Muhammad and M. W. Spratling, "A neural model of binocular saccade planning and vergence control," *Adaptive Behavior*, vol. 23, no. 5, pp. 265–282, 2015.

[11] A. Gibaldi, M. Chessa, A. Canessa, S. P. Sabatini, and F. Solari, "A cortical model for binocular vergence control without explicit calculation of disparity," *Neurocomputing*, vol. 73, no. 7-9, pp. 1065–1073, 2010.

[12] A. Gibaldi, A. Canessa, M. Chessa, S. P. Sabatini, and F. Solari, "A neuromorphic control module for real-time vergence eye movements on the iCub robot head," *IEEE-RAS International Conference on Humanoid Robots*, pp. 543–550, 2011.

[13] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 x 128 120 dB 15 us latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[14] J. Kogler, C. Sulzbachner, and W. Kubinger, "Bio-inspired stereo vision system with silicon retina imagers," *Computer Vision Systems*, pp. 174–183, 2009.

[15] P. Rogister, R. Benosman, S.-h. Ieng, and P. Lichtsteiner, "Asynchronous Event-Based Binocular Stereo Matching," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 2, pp. 347–353, 2012.

[16] J. Carneiro, S. H. Ieng, C. Posch, and R. Benosman, "Event-based 3D reconstruction from neuromorphic retinas," *Neural Networks*, vol. 45, no. March, pp. 27–38, 2013.

[17] L. A. Camunas-Mesa, T. Serrano-Gotarredona, S. H. Ieng, R. B. Benosman, and B. Linares-Barranco, "On the use of orientation filters for 3D reconstruction in event-driven stereo vision," *Frontiers in Neuroscience*, vol. 8, no. 8 MAR, pp. 1–17, 2014.

[18] M. Firouzi and J. Conradt, "Asynchronous Event-based Cooperative Stereo Matching Using Neuromorphic Silicon Retinas," *Neural Processing Letters*, vol. 43, no. 2, pp. 311–326, 2016.

[19] X. Lagorce, C. Meyer, S.-H. Ieng, D. Filliat, and R. Benosman, "Asynchronous Event-Based Multikernel Algorithm for High-Speed Visual Features Tracking." *IEEE transactions on neural networks and learning systems*, pp. 1–11, sep 2014.

[20] S. Tschechne, R. Sailer, and H. Neumann, "Bio-Inspired Optic Flow from Event-Based Neuromorphic Sensor Input," *Artificial Neural Networks in Pattern ...*, pp. 1–12, 2014.

[21] A. Roncone, U. Pattacini, G. Metta, and L. Natale, "A Cartesian 6-DoF Gaze Controller for Humanoid Robots," *Robotics: Science and Systems (RSS)*, 2016.

[22] M. Chessa and G. Pasquale, "Graphics processing unit-accelerated techniques for bio-inspired computation in the primary visual cortex," *Concurrency Computation Practice and Experience*, vol. 22, no. 6, pp. 685–701, 2010.

[23] M. Chessa, V. Bianchi, M. Zampetti, S. P. Sabatini, and F. Solari, "Real-time simulation of large-scale neural architectures for visual features computation based on GPU." *Network (Bristol, England)*, vol. 23, no. 4, pp. 272–91, 2012.

[24] R. Benosman, S. H. Ieng, C. Clercq, C. Bartolozzi, and M. Srinivasan, "Asynchronous frameless event-based optical flow," *Neural Networks*, vol. 27, pp. 32–37, mar 2012.

[25] M. B. Milde, O. J. N. Bertrand, R. Benosman, M. Egelhaaf, and E. Chicca, "Bioinspired event-driven collision avoidance algorithm based on optic flow," in *Event-Based Control, Communication, and Signal Processing*, Krakow, Poland, 2015.

[26] H. A. Mallot, A. Roll, and P. A. Arndt, "Disparity-evoked vergence is driven by interocular correlation," *Vision Research*, vol. 36, no. 18, pp. 2925–2937, 1996.

[27] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano, "The iCub humanoid robot: An open-systems platform for research in cognitive development," *Neural Networks*, vol. 23, no. 8-9, pp. 1125–1134, 2010.