An Active Approach to Solving the Stereo Matching Problem using Event-Based Sensors

Julien N.P. Martel*†, Jonathan Müller*, Jörg Conradt†, and Yulia Sandamirskaya*

*Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland.

†Neuroscientific Systems Theory, Technical University of Munich, Munich, Germany

Abstract—The problem of inferring distances from a visual sensor to objects in a scene – referred to as depth estimation – can be solved in various ways. Among those, stereo vision is a method in which two sensors observe the same scene from different viewpoints. To recover the three-dimensional coordinates of a point, its two projections – one in each view – can be used for triangulation. However, the pair of points in the two views that correspond to each other has to be found first. This is known as stereo-matching and is usually a computationally expensive operation. Traditionally, this is performed by describing a point in the first view with some information from its surrounding, e.g. in a feature vector, and then searching for a match with a point described in a similar way in the other view. In this work, we propose a simple idea that alleviates this stereo-matching problem using an active component: a mirror-galvanometer driven laser. The laser beam is deflected by actuating two mirrors, thus creating a sequence of "light spots" in the scene. At these spots, contrast changes quickly. We capture those contrast changes by two Dynamic Vision Sensors (DVS). The high time-resolution of these sensors enables the detection of the laser-induced events in time and their matching using lightweight computation. This method enables event-based depth estimation at a high speed, low computational cost, and without exact sensor synchronization.
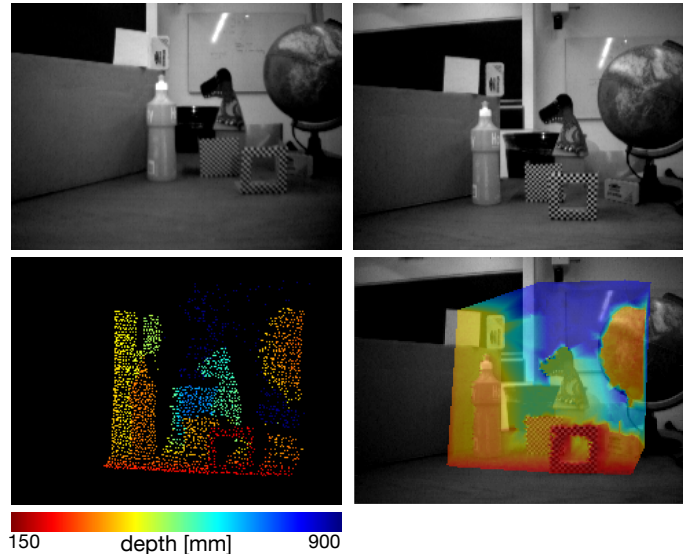
150   depth [mm]   900

Fig. 1. An exemplar depth estimation result. The top row shows the right and left grayscale images as captured by the DAVIS240C sensors. Note, these are not used for depth estimation (only events are used). The bottom row shows the sparse depth reconstruction obtained from scanning with the mirror-galvanometer driven laser and a bilinearly interpolated depth map overlaid on the rightw grayscale image.

I. INTRODUCTION

Humans, animals, and machines typically capture visual information with sensors in which light is projected onto a two-dimensional surface, e.g. a layer of photosensitive cells in biological retinas or arrays of photosensitive elements in silicon-based technologies. Inherent to this projection is the loss of the third dimension: depth. Biological systems evolved different strategies to infer this third dimension [1], [2]. Many of those strategies have also been reproduced in artificial vision systems. Those can estimate depth from binocular parallax using two cameras [2] or motion [3], [4], from focus [5], [6], from shading [7], from occlusions [8], from linear perspective [9], and many others. Animals do not solely use one of these strategies, but a combination of them to a different degree. Here, we focus on inferring depth from binocular parallax using sensors in a stereo-configuration.

By observing a scene with two sensors separated by some baseline, the third dimension – depth – can be recovered by triangulation [10]. Triangulation amounts to intersecting two rays originating from the optic centers of each camera and going through the two-dimensional projection of the three-dimensional point in each image plane. This intersection can be found using the geometrical parameters of the stereo system: how the cameras are positioned and oriented relative to each other (extrinsic parameters) and what their optical properties are (intrinsic parameters). These parameters are captured in what is usually called the calibration of the system.

To be able to triangulate two points, at the core of stereo vision is the matching of a pair of two-dimensional projections coming from the same three-dimensional point. Typically this problem, known as stereo-matching, has been addressed with two different types of approaches that are referred to as area-based and feature-based [11]. In area-based approaches, matching of points is made according to the similarity of their surroundings (e.g., patches in images [12] or time surfaces in event-based systems [13]). Feature-based approaches first detect salient keypoints, such as corners, then describe them with a relatively low-dimensional vector ("feature vector"), and match those to conclude that two points should be paired [14]. The area-based approaches usually lead to dense reconstruction (for textured regions), while feature-based depth estimation can be extremely sparse. In both these approaches, matching of areas or feature vectors can be computationally expensive. Specifically, 1) a description of each keypoint or area is required in both views that has to be robust to appearance variations between the two views: illumination, projective transformations, time warping, occlusions, etc., then 2) a search is performed to know which area or feature might match in the other view. The difficulty of stereo-matching, whether it is area based or feature based, is illustrated by the numerous hardware accelerators that have been devised to perform these operations [15]–[19].

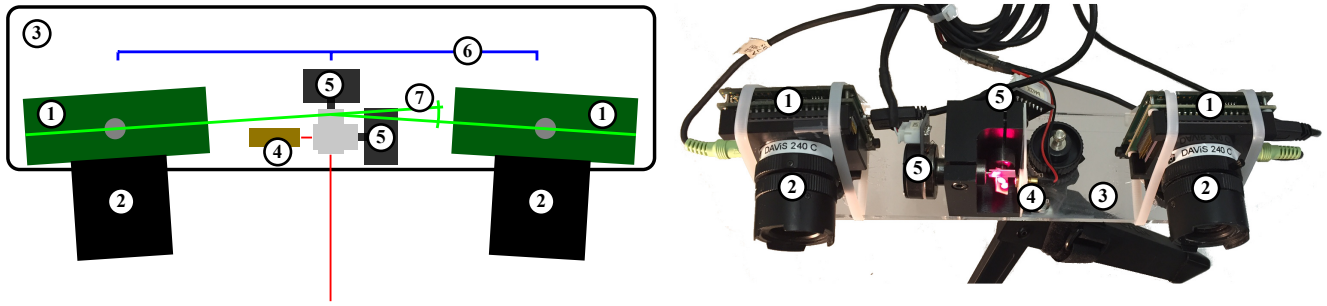We are interested in alleviating the problem of stereo-

Fig. 2. Top view of the system, both as a schematic and a photograph. Two DAVIS240C sensors (1) with optical lenses in front (2) are placed on a baseplate (3). A laser diode (4), whose beam can be deflected by two mirror-galvanometers (5) is mounted between the sensors. The sensors are separated by a baseline distance (6), and have a relative rotation angle (7). The sensors are connected to a PC via USB.

matching by introducing an active component: a mirror-galvanometer driven laser, whose beam can be deflected quickly within a scene. The laser is used to create artificial stereo-matches that can be easily detected. For instance, two conventional cameras could be used to detect a high-intensity red blob in their images: if it is unique, it is a match and it can be triangulated. Then, the red blob can be moved by deflecting the laser at another place, and we get another match and another triangulated point. Such a procedure assumes that finding the red blob in two images is fast, and due to the limited sampling rate of frame-based cameras, say about 30Hz, we would get at most 30 matches per second. Can we find a better solution?

The laser beam produces a brightness change when blinking or moving. Thus, a natural sensor to use to detect such a change is a Dynamic Vision Sensor (DVS) [20]–[22]. A DVS is a vision chip that reports events indicating changes in brightness for each pixel asynchronously. As soon as a pixel (x, y) incurs a change in brightness, a timestamped packet is sent with the pixel's position and the nature of the change: whether the brightness has increased or decreased. Hence, whenever the laser blinks or moves to a position, it creates events that are captured in the two DVS (as seen in Figure 3). In a static scene (nothing moves, and the illumination is unchanged), this would be the only source creating events and the stereo-matching problem is reduced to the match of two blobs in the event streams of the cameras. In a dynamic scene, we can benefit from using a laser blinking at high frequency, thus producing many more events than moving objects can realistically generate and consider only stereo-matches at high frequency. These scenarios are discussed in this work.

In this paper, we use a mirror-galvanometer driven laser to create a point in the scene that generates a blob in space-time that can be easily detected in the two event streams. Our approach is to create a sequence of these blobs. A sequence can be trivially matched across the two sensors benefiting from their high temporal resolution. We demonstrate a proof of concept of the system in a number of table-top scenes.

II. PRINCIPLES OF THE ACTIVE STEREO VISION SYSTEM

A. Overview of the system

The system we propose in this work consists of two DAVIS sensors [21] mounted on a common base plate (Fig. 2). DAVIS sensors contain both a Dynamic Vision Sensor (DVS) and an Active Pixel Sensor (APS) circuit in each pixel. Thus,

conventional APS frames producing grayscale images can be triggered along event readouts. In this work, we use a DAVIS240C with an array of 240×180 pixels.

The sensors are separated by a baseline of 18cm and their relative orientation of $\pm 5°$ tries to maximize the overlap between their field of view – in our setup, depth can only be inferred for 3D points in the scene that are seen by both sensors and whose projection can be matched.

Between the two sensors, we place a two-axis mirror-galvanometer equipped with a laser. The laser shines onto two magnetically actuated mirrors. The rotation of the first mirror controls how the laser beam is deflected on the vertical axis, while the second mirror controls its horizontal deflection. The current drivers of the galvanometer are on a separate board, they can be controlled by a microcontroller that is connected together with the two sensors to a computer. The computer is used for reading out the data of the two sensors, sending commands to the mirror-galvanometer and implementing our detection, matching, and triangulation routines.

In our system, we perform calibrated stereo vision. A calibrated stereo rig enables the reconstruction of metric three-dimensional geometry, where the geometric parameters of the system –both of the projection in each sensor and of their relative positioning– are known, as they are estimated during calibration. Then, a pair of points matched from the first and second sensors' views can be used to triangulate the position of their corresponding point in three-dimension.

Processing in our system unfolds in the following steps:

- We first calibrate the sensors individually (intrinsics) and then calibrate them together as a stereo rig (extrinsics).
- We start scanning the scene with the laser beam. The way scanning is performed influences how points will be matched and indeed which points are triangulated.
- We detect events induced by the laser in each sensors while scanning and matching them across sensors, yielding stereo-matching pairs.
- We triangulate the two-dimensional pairs into three-dimensional points.

In the following, we explain those steps in greater detail.

B. Calibration

We consider a pinhole camera model. This assumption is a reasonable first order approximation for the projection of a point in the scene onto each sensor's plane, given that the

combination of our sensors and lenses features a relatively narrow field of view and a low aperture. Calibration consists of two stages:

Intrinsics calibration: The first stage aims at estimating the optical properties of each of the sensors together with their lenses, such as the focal length of the combined lens system, the optical centering with respect to the sensor, and eventual distortions that are introduced by the lenses. These characteristics are captured in the intrinsic parameters of the cameras.

Extrinsics calibration: In the second stage, we aim at calibrating the stereo rig by finding the relative positioning of the two sensors –their baseline– and the relative orientation of the two cameras. These characteristics are summarized in what is referred to as the extrinsic parameters of the setup.

In the DAVIS240C sensor we use, each pixel contains both an APS and DVS circuit sharing the same photodiode. Thus, we can calibrate our system using the grayscale images produced by the APS part and use the estimated parameters for further operations involving events produced by the DVS part. We use a known calibration pattern in a standard calibration procedure based on Zhang's algorithm [23] to find the intrinsic parameters of our two sensors. Using the intrinsics found for each of them, we find the extrinsics of our stereo rig using another standard procedure minimizing the total reprojection error (using the known position of the dots in our known calibration pattern).

C. Scanning the scene with the laser: two sampling strategies

Our goal is to solve the stereo-matching problem by inducing events in our two sensors with a laser. Thus, we want to induce events in patterns that can be easily segregated from events generated by other sources in the scene. We have implemented two different strategies to scan the scene with our mirror-galvanometer driven laser beam.

Random scanning: The first strategy we implemented consists in scanning the scene by deflecting the laser beam at random locations. Before moving the laser to a new position, it is switched off –so that no event is created in the sensors when the laser is on the way to the target position– and switched on again when it has reached the target position. There, it is kept fixed for a very short time (a few milliseconds or less) before starting this procedure again moving it to a new position. At the target position, the laser can also be blinked at a specific frequency (for instance, a kilohertz) so that it can be tracked. This allows us to uniquely identify it and segregate the events it produces from other events in the scene, created, for instance, by moving objects or other blinking sources. Note that in this setting, multiple lasers, uniquely identified by their frequencies could be used simultaneously. A space-time diagram illustrating this strategy is shown on the top of Fig. 3.

Continuous "line" scanning: The second strategy we considered moves the laser continuously in lines, that is, the two mirrors of the galvanometers are rotated at constant, tunable, angular velocities. In this mode, the laser is always switched "on" and creates events in the form of polygonal chains (the lines are deformed depending on the depth of the scene, e.g. reducing to simple lines if the laser is projected
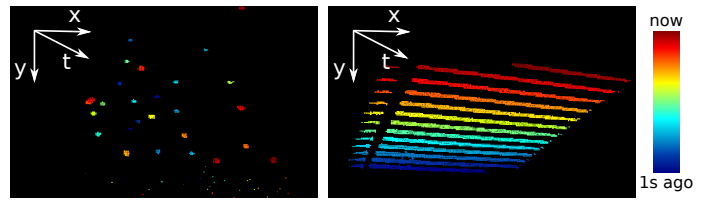

Fig. 3. Two different strategies for scanning the scene with the mirror-galvanometer driven laser beam are illustrated with a space-time diagram of events, showing a time window of 1s. In the first figure, the laser illuminates random spots for the duration of 10ms each. In the second, the laser scans the scene linearly in negative-x and negative-y direction.

on a simple plane). This strategy is shown in the space-time diagram Fig. 3.

D. Detection and matching of the laser-induced events

We investigated three different techniques to detect and then match event blobs induced by the laser beam in each of the sensors. They are: 1) clustering events by space-density, 2) tracking the frequency of the laser-induced events, and 3) filtering events by time-density. In the following we briefly discuss their implementation and uses depending on the scanning strategy that we employ.

Event clustering: In our continuous-scanning strategy, event-blobs produced by the laser move smoothly across the field of view, suggesting that a method using spatial density could detect those. We use a simple mean-shift algorithm that creates a cluster when a high event density is detected and moves it around as new events are detected close to the cluster.

Frequency based detection: With our random scanning strategy, if the laser is blinked at a known frequency, a way to detect laser induced events is to search for events spaced in time by the laser blinking period. Various methods to detect a blinking pattern in an event-stream have been described, such as in [24].

Event-density filter: Again, in our random-scanning strategy, if the blinking frequency of the laser is about a kilohertz, one should note that in most real-world scenes, no other source naturally creates as many events. Thus, a variation on the frequency-based detection of laser-induced events is to search for pixels at which the event-density is particularly high. A simple implementation of such an event-density filter is to measure, at each pixel, the time it takes to collect N events. This gives a rough estimate of instantaneous event density when N is small (e.g., $N = 25$). Then this density is thresholded such that only the pixels where the laser is blinking trigger the detection of an event-blob.

These three techniques can be very efficiently implemented in an event-driven framework. In such a framework, computation takes place only when and where events occur. As a concrete example: instantaneous density in the case of the event-density filter can be updated solely for the pixels receiving events; in addition, the last update time of this pixel can be kept, so that we can detect when the instantaneous density estimate at that pixel is outdated.

Once event blobs are detected in the two cameras, they are trivially matched by their order of appearance in time, which allows us to be robust against inexact synchronization of our two sensors.
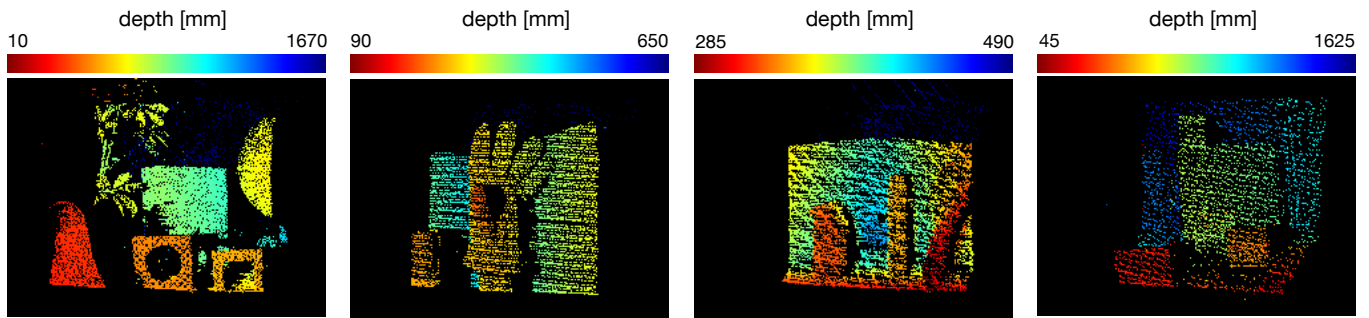
Fig. 4. Sparse depth maps, output by the system for four different scenes.

E. Triangulating matches

A pair of corresponding points in each sensor can be triangulated in our calibrated stereo vision setting to obtain their three-dimensional coordinates. Triangulation is performed using a Direct Linear Transform [10] after refining the two-dimensional coordinates of our points using constraints imposed by the geometry of our setup (epipolar constraints).

III. RESULTS

A. Depth maps of different scenes

Figures 4 and 1 show the output of our system for five different tabletop scenes. In these experiments, the scenes were scanned by a 20mW, 650nm (red) laser diode. The baseline of our system is 18.0cm and the relative orientation of the cameras is 3.9 and 5.0° in the vertical and horizontal directions. With such a setup we get 5mm of depth resolution one meter in front of the sensors (assuming the blobs are tracked with accuracy of one pixel). The galvanometer-driven laser covers a field of view of 35° horizontally as well as vertically. And our DAVIS240C mounted with the same lens model, have a $55° \times 42°$ ($W \times H$) angular field of view.

In our experiments, we used the high-density filter to track events using continuous line-scanning. We get depth estimates, only for the region in space where the field of views of the two sensors overlap, and that is reachable by the laser beam, as seen in Figure 1. The scenes on which we tested our system range from 30cm to about 5m. Indeed, depth resolution decreases for larger distances, whereas at short distances regions seen in each sensor are not likely to overlap.

In these proof of concept experiments, we took up to a minute to scan the scenes. This allows us to produce images that look dense: We have not yet worked on inpainting and denoising these pseudo-sparse depth maps to produce dense ones for this system. However we present in the bottom right plot of Figure 1 a bilinearly interpolated (with the closest depth measurement) depth map overlaid on one of the grayscale images of the scene.

Note that we can reliably track the laser with our three methods (event clustering, high event density filter, and frequency tracking) at up to $70° \, \mathrm{s}^{-1}$ on a plane about 1m away from the system.

IV. DISCUSSION AND OUTLOOK

In this paper, we demonstrated how combining a laser that can quickly scan a scene with a pair of DVS allows us to create a stereo vision system, in which the stereo-matching problem is alleviated. Event blobs are induced by the laser and retrieved by computationally lightweight methods before being matched by their ordering in time.

To help placing our system in the vast landscape of event-based stereo vision approaches that have been devised, note that 1) it does not rely on the precise spatio-temporal matching of event time surfaces such as in [13], [25]–[28], and that 2) even though it is an active approach using the emission of light in the scene, it is not using the idea of analyzing the deformations of an a-priori known pattern of light such as in structured light (that often make use of a single sensor) [29]–[31], nor using any time-of-flight information [32].

Of course, one could think one of the main limitations of our system is the speed at which the laser pointer can be moved. This is limited by the actuation of the mirrors of the galvanometer to $35\mathrm{kPts}.\mathrm{s}^{-1}$, a point being a 12-bits command (uniformly covering the angular field view) sent to the mirror-galvanometer laser. In fact, the system is mostly limited by the number of events our sensors capture given the limited power of the laser diode we use and the velocity at which we move it. Beyond an angular velocity of $70° \, \mathrm{s}^{-1}$ at 1m, our laser produces so few events in that our detection algorithms fail. Higher power laser diodes can be used, but might not be safe in many applications.

Also, if the laser is blinked, a trade-off needs to be made between its blinking frequency and the velocity it is moving at. If it is blinked with a low frequency, we can only move the laser slowly as we need to detect one or two blinking periods to detect the laser. On the contrary, if the laser is blinked fast, we can move it quickly, but will only detect very few events in our sensors.

With our approach, depth in a large scene can be sampled at a few frames per second. However, if we want to probe the distance of a few points only, e.g. to update uncertain regions on a mobile robotic platform or in a scene where only a few objects are moving, this active system offers fast and precise depth estimation at a low computational cost.

REFERENCES

[1] M. F. Land and D.-E. Nilsson, *Animal eyes*. Oxford University Press, 2012.

[2] I. P. Howard, *Perceiving in depth, volume 1: basic mechanisms*. Oxford University Press, 2012.

[3] S. Ullman, "The interpretation of structure from motion," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 203, no. 1153, pp. 405–426, 1979.

[4] O. D. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 2, no. 03, pp. 485–508, 1988.

[5] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 16, no. 8, pp. 824–831, 1994.

[6] J. N. P. Martel, L. K. Mller, S. J. Carey, J. Mller, Y. Sandamirskaya, and P. Dudek, "Real-time depth from focus on a programmable focal plane processor," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. PP, no. 99, pp. 1–10, 2017.

[7] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1940–1948.

[8] I. Tsirlin, L. M. Wilcox, and R. S. Allison, "The effect of crosstalk on the perceived depth from disparity and monocular occlusions," *IEEE Transactions on broadcasting*, vol. 57, no. 2, pp. 445–453, 2011.

[9] J. A. Saunders and B. T. Backus, "The accuracy and reliability of perceived depth from linear perspective as a function of image size," *Journal of Vision*, vol. 6, no. 9, pp. 7–7, 2006.

[10] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[11] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Proceedings - IEEE Workshop on Stereo and Multi-Baseline Vision, SMBV 2001*, vol. 47, no. 1, pp. 131–140, 2001.

[12] T. Tao, J. C. Koo, and H. R. Choi, "A fast block matching algorthim for stereo correspondence," in *Cybernetics and Intelligent Systems, 2008 IEEE Conference on*. IEEE, 2008, pp. 38–41.

[13] J. Carneiro, S. H. Ieng, C. Posch, and R. Benosman, "Event-based 3D reconstruction from neuromorphic retinas," *Neural Networks*, vol. 45, pp. 27–38, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2013.03.006

[14] B. Tang, D. Ait-Boudaoud, B. J. Matuszewski, and L.-k. Shark, "An Efficient Feature Based Matching Algorithm for Stereo Images," *Geometric Modeling and ImagingNew Trends GMAI06*, vol. 293, pp. 195–202, 2006. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1648766

[15] F.-C. Huang, S.-Y. Huang, J.-W. Ker, and Y.-C. Chen, "High-performance sift hardware accelerator for real-time image feature extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 3, pp. 340–351, 2012.

[16] K. Sugita, T. Naemura, and H. Harashima, "Performance evaluation of programmable graphics hardware for image filtering and stereo matching," in *Proceedings of the ACM symposium on Virtual reality software and technology*. ACM, 2003, pp. 176–183.

[17] O. Reiche, K. Häublein, M. Reichenbach, F. Hannig, J. Teich, and D. Fey, "Automatic optimization of hardware accelerators for image processing," *arXiv preprint arXiv:1502.07448*, 2015.

[18] C. Kim, K. Bong, S. Choi, and H.-J. Yoo, "A 43.7 mw 94 fps cmos image sensor-based stereo matching accelerator with focal-plane rectification and analog census transformation," in *Circuits and Systems (ISCAS), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1418–1421.

[19] C. Kim, K. Bong, S. Choi, K. J. Lee, and H.-J. Yoo, "A cmos image sensor-based stereo matching accelerator with focal-plane sparse rectification and analog census transform," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 12, pp. 2180–2188, 2016.

[20] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 X 128 120db 30mw asynchronous vision sensor that responds to relative intensity change," *2006 IEEE International Solid State Circuits Conference - Digest of Technical Papers*, pp. 2004–2006, 2006. [Online]. Available: http://siliconretina.ini.uzh.ch/wiki/lib/exe/fetch.php?media=lichtsteiner{_}isscc2006{_}d27{_}09.pdf

[21] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.

[22] C. Posch, D. Matolin, and R. Wohlgenannt, "A qvga 143db dynamic range asynchronous address-event pwm dynamic image sensor with lossless pixel-level video compression," pp. 400 – 401, 03 2010.

[23] Z. Zhang, "A Flexible New Technique for Camera Calibration," *Technical Report*, 1998.

[24] A. Censi, J. Strubel, C. Brandli, T. Delbruck, and D. Scaramuzza, "Low-latency localization by active LED markers tracking using a dynamic vision sensor," *IEEE International Conference on Intelligent Robots and Systems*, pp. 891–898, 2013.

[25] P. Hess, "Low-Level Stereo Matching using Event-based Silicon Retinas," *Semesterarbeit am Institut für Neuroinformatik, ETH . . .*, pp. 1–19, 2006. [Online]. Available: http://www.ini.unizh.ch/{~}tobi/studentProjectReports/hessAERStereo2006.pdf

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

[27] D. R. Valeiras, G. Orchard, S. H. Ieng, and R. B. Benosman, "Neuromorphic Event-Based 3D Pose Estimation," vol. 9, no. January, pp. 1–15, 2015.

[28] M. Osswald, S.-H. Ieng, R. Benosman, and G. Indiveri, "A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems," *Scientific Reports*, vol. 7, no. December 2016, p. 40703, 2017. [Online]. Available: http://www.nature.com/articles/srep40703

[29] O. S. Cossairt, N. Matsuda, and M. Gupta, "Motion contrast 3d scanning," in *Computational Optical Sensing and Imaging*. Optical Society of America, 2015, pp. CT2E–1.

[30] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[31] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 195–202, 2003. [Online]. Available: http://dasan.sejong.ac.kr/{~}dihan/cv/S07{_}StereoUsingStructuredLight.pdf

[32] I. Vornicu, R. Carmona-Galán, and Á. Rodríguez-Vázquez, "Arrayable voltage-controlled ring-oscillator for direct time-of-flight image sensors," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2017.