Supplementary Material: Event-based Shape from Polarization

Manasi Muglikar [1] Leonard Bauersfeld [1] Diederik Paul Moeys [2] Davide Scaramuzza [1]

[1]Robotics and Perception Group, University of Zurich, Switzerland
[2]Advanced Sensors and Modelling Group, SONY R&D Center Europe, SL1

1. Surface Normal from Events

In this section, we describe the details for surface normal estimation from polarizer images. We then extend this knowledge to estimate surface normals from events.

1.1. Basics of Shape-from-Polarization (SfP)

Intensity change at ϕ_{pol} can be expressed as:

$$I(\phi_{\text{pol}}) = \frac{I_{\text{max}} + I_{\text{min}}}{2} + \frac{I_{\text{max}} - I_{\text{min}}}{2} \cdot \cos(2(\phi_{\text{pol}} - \phi)), \quad (1)$$

where I_{min} and I_{max} represent the minimum and maximum magnitude seen through the polarizer respectively [4, 7]. This equation can be expressed in terms of the magnitude of the light I_{un} and the proportion of polarized component ρ (also known as degree of polarizer) as follows:

$$I = I_{max} + I_{min} \quad (2)$$

$$\rho = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} \quad (3)$$

Lastly, ϕ is the angle of the linearly polarized component which corresponds to the phase shift of the sinosoid. [7]. Estimating these three parameters forms the crux of shape-from-polarization techniques [9]. These quantities can be estimated from images captured at 4 different polarization angles as follows:

$$I_{un} = \frac{I[0] + I[\pi/4] + I[\pi/2] + I[3\pi/4]}{2} \quad (4)$$

$$\rho = \frac{\sqrt{(I[0] - I[\pi/2])^2 + (I[\pi/4] - I[3\pi/4])^2}}{I_{un}} \quad (5)$$

$$\phi = \frac{1}{2} \cdot \arctan \frac{I[\pi/4] - I[3\pi/4]}{(I[0] - I[\pi/2]} \quad (6)$$

To estimate these quantities, minimum 3observations of the intensity are required. However, increasing the observations, improves the accuracy of surface normals. To use 12 polarization angles the above quantities can be derived as follows:

$$I_{un} = \sum_{i=0}^{i=\pi} I[i] \quad (7)$$

$$Q1 = (I[0] - I[\pi/2]) \quad (8)$$
$$Q2 = (I[\pi/12] - I[7\pi/12]) \quad (9)$$
$$Q3 = (I[\pi/6] - I[3\pi/2]) \quad (10)$$
$$U1 = (I[\pi/4] - I[3\pi/4]) \quad (11)$$
$$U2 = (I[\pi/3] - I[5\pi/6]) \quad (12)$$
$$U3 = (I[5\pi/12] - I[11\pi/12]) \quad (13)$$
$$(14)$$

$$\rho = \frac{\sqrt{Q1^2 + U1^2 + Q2^2 + U2^2 + Q3^2 + U3^2}}{3 * I_{un}} \quad (15)$$

$$\phi = 1.5 * (\arctan(U1/Q1) \quad (16)$$
$$+ \arctan(U2/Q2) - \pi/6 \quad (17)$$
$$+ \arctan(U3/Q3) - \pi/3) \quad (18)$$

Estimating the surface normals from ρ and ϕ is a matter of estimating the zenith angle θ and azimuth angle α as shown in the equations below:

$$\rho^{diffuse} = \frac{(n - \frac{1}{n})^2 \sin^2\theta}{2 + 2n^2 - (n + \frac{1}{n})^2 \sin^2\theta + 4\cos\theta\sqrt{n^2 - sin^2\theta}} \quad (19)$$

$$\rho^{spec} = \frac{2n\tan\theta}{\tan^2\theta\sin^2\theta + n^{*2}} \quad (20)$$

where n denotes the refractive index and θ is the zenith angle. Depending on the type of reflection (diffuse or specular), the ρ is computed differently. Similarly depending the type of reflection, the azimuth angle α is ϕ if diffuse reflection dominates otherwise it is $\phi - \pi/2$:

1.2. SfP from Events

When estimating surface normals from events, we reconstruct event intensities (I_e) as explained in the main paper. Using the above equations, we estimate ρ and θ by first estimating event intensities at 12 polarizer angles, The use of event intensities enables us to use the traditional SfP algorithms to estimation surface normals. Depending on the type of polarization (specular or diffuse), this can result in multiple solutions. We observed using the specular solution results in the lowest angular error. We also used the Smith *et al*. [8] baseline with our event intensities. However, this results in a lower performance as shown in Table 1.

2. Event Representation

When learning surface normals from events, the event representation have a significant effect on the performance of the network. In this section, we describe the performance of 3 kinds of input representations namely: event intensities (I_e), voxel grid [11], CVGR representation and CVGR-I representation on the ESfP-synthetic dataset. The event intensity representation concatenates I_e at polarizer angles of $15, 60, 105, 150$ as input the network. (note, we cannot use the intensity at 0 angle, since it will always be zero for all pixels). The CVGR representation builds on top of voxel grid representation as follows:

$$E(x,y,b) = \sum_{i=0}^{i=b} C \cdot V(x,y,i) = \sum_{i=0}^{i=b} C \left(\sum_{\substack{e_k \in \mathcal{E}_i: \\ x_k = x, y_k = y}} p_k \right). \tag{21}$$

Lastly, the CVGR-I representation combines a single image with events and is expressed as follows:

$$E(x,y,b) = I[0] + \sum_{i=0}^{i=b} C \cdot V(x,y,i) \tag{22}$$

As can be seen from Table 1, the best performing representation is CVGR-I. The main reason for improvement is because the image gives more context to the network to estimate surface normals in the areas where the event information is insufficient. Qualitative results on real dataset are shown in Fig. 1. As can be seen, the events are triggered prominently on the edge of the vase and are missing from the front-parallel surface of the vase. The network using only events has a difficult time to estimate normals on these front-parallel surfaces. On the other hand, using CVGR-I representation, the network performs better resulting in a lower MAE score. Additionally, we also evaluate the effect of number of bins on the performance of the network. For the same representation, increasing the number of bins from 4 to 8 improves the performance by 6% in terms of angular error. Higher number of bins preserves the temporal information of events better. However, further increasing the bins to 12 results in a a decrease in performance. This is because not all bins add new information due to limitation of contrast threshold.

Method	Dimension	Angular Error ↓	Accuracy ↑		
		Mean	AE<11.25	AE<22.5	AE<30
Events (P) [8]	$12 \times H \times W$	69.722	0.028	0.067	0.098
Events (P, specular)	$12 \times H \times W$	58.196	0.007	0.046	0.095
Event intensities	$4 \times H \times W$	39.316	0.147	0.321	0.402
VoxelGrid [11] - 8Bins	$8 \times H \times W$	34.232	0.230	0.465	0.556
CVGR - 4Bins	$4 \times H \times W$	34.053	0.220	0.494	0.579
CVGR - 8Bins	$8 \times H \times W$	32.010	0.248	0.515	0.594
CVGR - 12Bins	$12 \times H \times W$	34.655	0.227	0.510	0.596
CVGR-I	$8 \times H \times W$	**27.953**	**0.263**	**0.527**	**0.655**

Table 1. Comparison of event representations: The first two rows correspond to physics-based baseline. Rest of the rows correspond to learning-based approaches with different event representations.

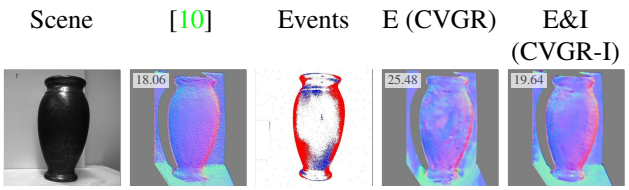

Figure 1. Qualitative comparison of event representation

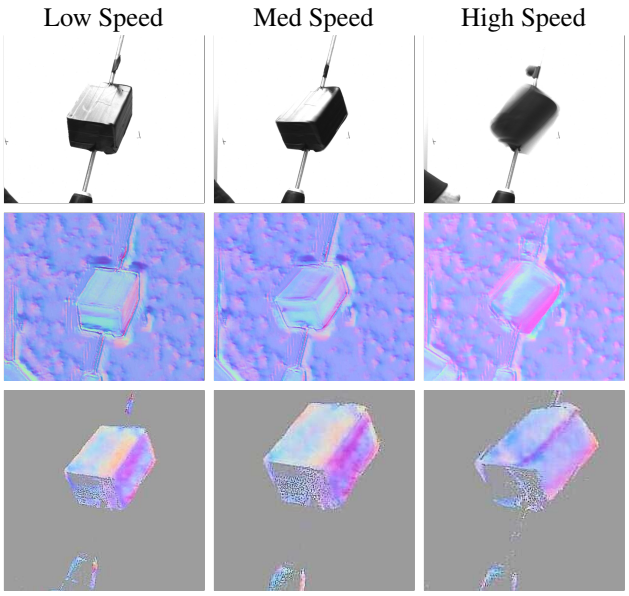

Figure 2. Comparison on dynamic scenes: The object is rotating with increasing speeds from left to right. The top row shows the images captured by the camera. The second row shows the surface normals estimated by image-based SfP baseline Ba *et al*. [10] and last row shows the surface normals estimation by our learning-based SfP baseline.

2.1. Dynamic scenes

An advantage of using event camera is the high temporal resolution as compared to the frame-based sensor. To highlight this, we record a dynamic scene. The scene consists of a rectangular block which is rotating about it's diagonal axis with a drill. The speed of the drill could be adjusted to three increasing levels. As seen in Fig. 2, the images corresponding to three different speeds are shown in the first row. Increasing the speed introduces motion blur for the standard camera, which is also reflected in the surface normals (second row, high speed). In contrast, event-based SfP methods (last row) are better than the image-based counterpart, as can be seen by the sharpness of the edge of the rectangular block. This is primarily due the high rotation speeds of the polarizer enabled by the high temporal resolution of the camera.

3. Dataset

ESfP Synthetic Dataset In this section, we provide details on the ESfP-Synthetic dataset which we use for evaluation. The dataset was generated using publicly available meshes [2] which consists of over 1000 3D scanned common household objects. These meshes were textured using the 25 textures available in this dataset [1]. These textures provide polarimetric BRDF of real-world materials which provide accurate polarimetric state information when used with physically-based simulation such as Mitsuba.

4. Limitations

Our real-world dataset only considers specular objects such as reflective metallic surfaces. We specifically chose specular objects as they are the most challenging to obtain surface normals for. The geometry of objects with diffuse reflection can be captured easily by methods such as structured light (SL). Additionally, the intensity changes of diffuse materials when observed with a rotating polarizer is low compared to specular objects, which the real event camera cannot capture due to a high contrast threshold. Therefore, capturing events for diffuse materials was not possible with the current version of event cameras.

5. Effect of speed

Conducting experiments at different rotation speeds of the polarizer, we observed a slight increase in the performance for our method and linked this to the decreasing relevance of nonidealities in the event-camera pixel circuits. In this section, we provide more details on why an increase in rotational speed improves the performance. Our analysis is based on additional experiments, general considerations on event-camera circuitry [6] and the technical details of the Prophesee Gen 4 event camera [3]. The additional

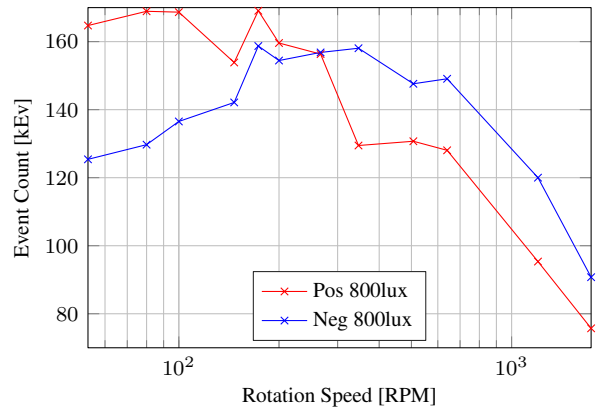

Figure 3. At low speeds more positive than negative events are triggered. Towards higher rotation speeds this trend reverses but the overall number of events per filter rotation decreases visibly.

data we recorded cover rotation speeds between 53 RPM up to 1,734 RPM and two illumination conditions, 200 lux and 800 lux. In this section, we will use the number of events triggered on the object per revolution of the polarizer as a proxy-measure for the quality of the resulting surface normals.

Ideal Event Camera A key observation stated in the main paper is that the illumination intensity cancels out in an idealized event camera model. For a given surface on the object degree of polarization ρ and polarization angle ϕ, the event camera observes a sinusoidal intensity profile of the form

$$I(t) = I_{\text{un}}(1 + \rho\cos(2\omega t - \phi)) \tag{23}$$

The ideal event-sensor triggers and event when the temporal contrast T (logarithmic intensity change) exceeds some threshold C [6]. Combining this with (23) yields an expression which, for an object with illumination independent polarization characteristics, only depends on the rotational speed.

$$T = \frac{\mathrm{d}\left(\ln I(t)\right)}{\mathrm{d}t} = \frac{-2\omega\rho\sin(2\omega t - \phi)}{1 + \rho\cos(2\omega t - \phi)} \tag{24}$$

Based on (24), we can see that the number of events triggered per unit time linearly depends on the rotational speed. By considering only the number of events per rotation, this dependency is also cancelled out and *an ideal event camera* would not show any dependency on the illumination condition or rotational speed.

Real Event Camera In practice however, we observe that illumination and rotation speed have an effect on the quality of the surface normal estimation. To better understand this, we look at the number of events triggered per rotation of the polarizer and observe that

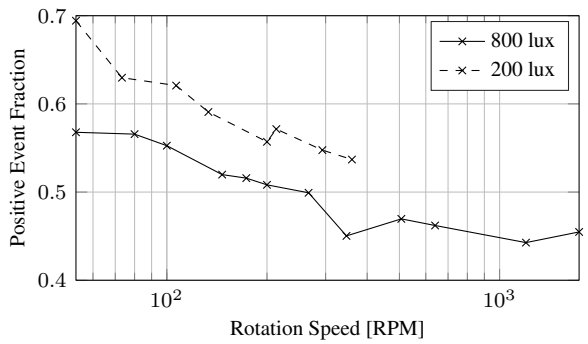

Figure 4. At low rotational speeds and low light conditions the number of positive events drastically exceeds 0.5 due to the background rate [5]

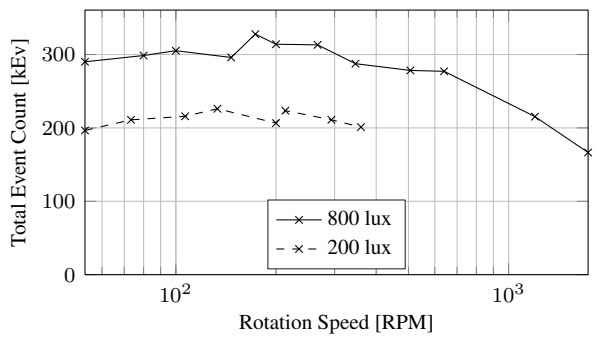

Figure 5. The total number of events (given in thousands) per rotation of the polarizer decreases at higher speeds and at lower illumination levels.

1. at low rotation speeds, more positive events are triggered (Fig. 3, Fig. 4),
2. the difference in fraction of positive and negative events at low RPM becomes more pronounced at lower illumination conditions (Fig. 4),
3. the number of events per rotation decreases at high rotation speeds (Fig. 5), and
4. at low illumnation conditions, less events are triggered at a set rotation speed (Fig. 5).

While the ideal event camera model fails to explain those observations, a more realistic model takes the non-idealities of the circuitry into account. In [5] the leakage of the reset-transistor is described as a major source of non-ideality as it leads to the spurious positive events, thus increasing the fraction of positive events. Because we consider the number of events per rotation, slower rotation speeds correspond to a longer accumulation times and the BG (background rate) rate corrupts such low-speed measurements stronger as shown in Fig. 3. This so-called *BG rate* (background rate) is illumination dependent [3]. Together with the increase in BG rate at lower light levels [3] (for bright scenes) this explains observations 1 and 2.

At high rotation speeds the BG rate has negligible in-

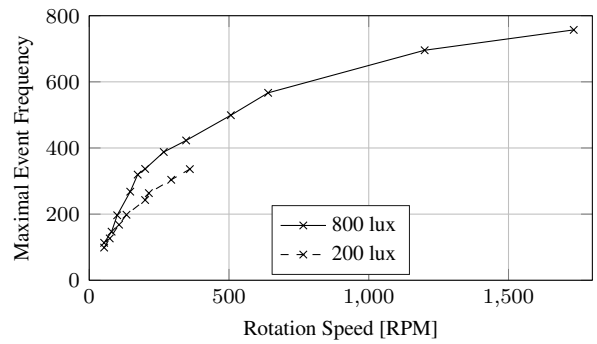

Figure 6. The frequency at which events are triggered on the object shows a saturation effect due to pixel dead time [5]. This explains the decrease in event count at high polarzier speeds.

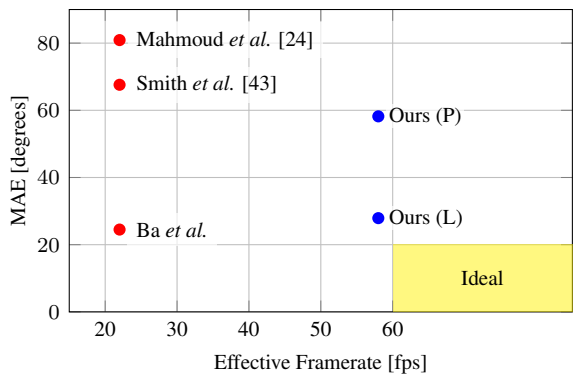

Figure 7. Acquisition time versus Mean Angular Error (MAE).

fluence. However, a second non-ideality becomes visible: after triggering an event the pixel needs a certain *dead time* until it can trigger the next event. This is typically done to avoid bus-saturation by a small group of pixels [5]. This effect is clearly visible when looking at the highest event frequency of pixels on the object (Fig. 6). For an ideal camera the event-frequency would depend linearly on the rotation speed as derived in (24). However, the data clearly shows a saturation effect because pixels can only be triggered with a limited frequency, around 1 kHz at 800 lux illumination. In accordance with literature, this maximum frequency decreases with decreasing illumination [5], explaining the remaining two observations.

In contrast to the ideal event camera model, the output of a real event-camera is sensitive to illumination and rotational speed of the polarizer. Low speeds increase the BG rate noise significantly and only medium polarizer speeds lead to a more even distribution of positive and negative events. If the speed is increased greatly, the pixel dead time may start to degrade the result again. This is in accordance with the results shown in the main paper.

6. Advantage of event camera

Fig. 7 illustrates the advantage of using an event-based SfP (blue) against frame-based SfP methods (red) using as metrics the framerate and the Mean Angular Error (MAE). Image-based approaches focus on maximizing the performance; however, they are restricted by the camera's framerate to 22 fps while reducing the effective resolution from 4MP to 1MP (DoFP approach). On the other hand, our event-based approach is 3 times faster and pushes the SfP methods toward higher framerates, without sacrificing the resolution. This enables the capture of surface normals of high-speed motion. Unlike high-framerate cameras, event cameras present a fundamentally new approach to visual information processing. While a high framerate camera would capture redundant information resulting in data bus saturation, an event camera only triggers events when there is contrast change, resulting in lower bandwidth.

References

[1] Seung-Hwan Baek, Tizian Zeltner, Hyun Jin Ku, Inseung Hwang, Xin Tong, Wenzel Jakob, and Min H. Kim. Image-based acquisition and modeling of polarimetric reflectance. *ACM Transactions on Graphics (Proc. SIGGRAPH 2020)*, 2020. 3

[2] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items, 2022. 3

[3] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch. A 1280x720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86μm pixels, 1.066geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *IEEE Intl. Solid-State Circuits Conf. (ISSCC)*, 2020. 3, 4

[4] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Polarized 3d: High-quality depth sensing with polarization cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3370–3378, 2015. 1

[5] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128x128 120dB 30mW asynchronous vision sensor that responds to relative intensity change. In *IEEE Intl. Solid-State Circuits Conf. (ISSCC)*, pages 2060–2069, 2006. 4

[6] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128$\times$128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, 2008. 3

[7] Olivier Morel, Fabrice Meriaudeau, Christophe Stolz, and Patrick Gorria. Polarization imaging applied to 3D reconstruction of specular metallic surfaces. In *Machine Vision Applications in Industrial Inspection XIII*, volume 5679, pages 178 – 186. International Society for Optics and Photonics, SPIE, 2005. 1

[8] William A. P. Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-polarisation with unknown lighting or albedo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2875–2888, 2019. 2

[9] Lawrence B Wolff. Polarization vision: a new sensory approach to image understanding. volume 15, pages 81–93. Elsevier, 1997. 1

[10] Ba Yunhao, Gilbert Alex, Wang Franklin, Yang Jinfa, Chen Rui, Wang Yiqin, Yan Lei, Shi Boxin, and Kadambi Achuta. Deep shape from polarization. 2020. 2

[11] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2018. 2