Local Event Alignment for Monocular Distance Estimation

Nan Cai, Pia Bideau

Local Event Alignment for Monocular Distance Estimation

Nan Cai
Technical University of Berlin
Germany
nan.cai@campus.tu-berlin.de

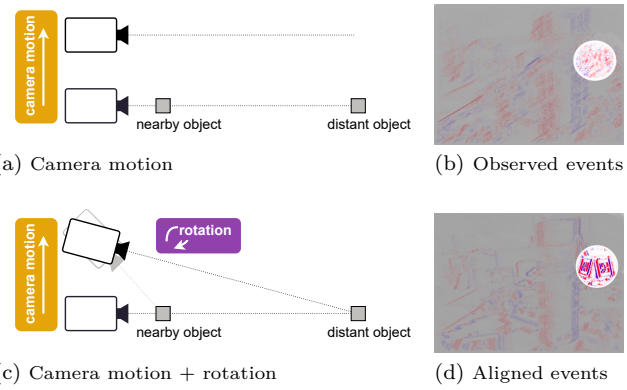
Pia Bideau
Univ. Grenoble Alpes, Inria,
CNRS, Grenoble INP, LJK
France
pia.bideau@inria.fr

Abstract

Event cameras provide a natural and data efficient representation of visual information, motivating novel computational strategies towards extracting visual information. Inspired by the biological vision system, we propose a behavior driven approach for object-wise distance estimation from event camera data. This behavior-driven method mimics how biological systems, like the human eye, stabilize their view based on object distance: distant objects require minimal compensatory rotation to stay in focus, while nearby objects demand greater adjustments to maintain alignment. This adaptive strategy leverages natural stabilization behaviors to estimate relative distances effectively. Unlike traditional vision algorithms that estimate depth across the entire image, our approach targets local depth estimation within a specific region of interest. By aligning events within a small region, we estimate the angular velocity required to stabilize the image motion. We demonstrate that, under certain assumptions, the compensatory rotational flow is inversely proportional to the object's distance. The proposed approach achieves new state-of-the-art accuracy in distance estimation - a performance gain of 16% on EVIMO2. EVIMO2 event sequences comprise complex camera motion and substantial variance in depth of static real world scenes.

1. Introduction

Event cameras mimic certain biological features of the human visual system. Instead of recording RGB frames at a fixed frequency, they record brightness changes as *events* asynchronously and at high temporal resolution. This offers great potential for high-speed automation [18], robotics [22] and microscopic analysis [9], where capturing motion is crucial. Current vision algorithms struggle to efficiently process and interpret asynchronous event data. In this work, we propose to combine the biologically inspired visual acquisition with natural behavioral strategies to enhance the interpretation of event data. Rather than passively analyzing incoming event data at each pixel location, we induce a movement that stabilizes image motion within a targeted region. Similar to eye movements, we introduce rotational adjustments to counterbalance translational camera motion.



(a) Camera motion (b) Observed events

(c) Camera motion + rotation (d) Aligned events

Figure 1. Akin to gaze stabilization, local event alignment stabilizes a local image region by applying a rotation that counteracts the camera's motion. This rotation leads to locally well aligned events as pictured in (d). The relative distance between two objects can then be inferred by comparing their compensatory rotations.

What are the benefits of stabilization? From an ecological perspective, stabilization reduces the number of brightness changes detected by the eye's photoreceptors. As photoreceptors ("light sensors" of the human eye) are slow and capture only a small part of the scene briefly, fast body movements can significantly degrade vision unless counteracted by eye movements [28]. By doing so stabilization conveys relevant scene-specific information, such as relative distances between objects. This paper proposes a novel method for distance esti-

mation by *local event alignment*. To stabilize incoming visual information within a specific area of the camera's sensor, we introduce rotational motion that compensates for the observer's movement. Under certain assumptions that are discussed in this paper, the compensatory rotational flow is inversely proportional to the object's distance. An example is shown in Fig. 1. Intuitively speaking, the farther away an object the less compensatory motion is needed for stabilization on the camera sensor. Thus the amount of compensatory motion needed to stabilize an object bears strong cues for 3D object localization.

We propose a novel alignment strategy, that unlike prior work *does not* aim at reconstructing the camera's motion. Instead, it determines the rotational velocity that leads to best event alignment across a set of small image regions, from which the object's relative distances are inferred. Without computing explicit correspondences between events and without knowledge about the camera's pose, we measure the degree to which a set of events are jointly aligned. The quality of the joint event alignment is assessed by calculating pixelwise entropy, leveraging redundant information to achieve a more effective representation [3, 4, 14].

Contributions. We propose a novel approach for relative distance estimation through object-wise event alignment. Our two step optimization strategy for event alignment robustly estimates a compensatory rotation that is directionally consistent but varies in velocity across the image. Relative depth is then computed by comparing the compensatory rotations of the object and the reference region. Notably, our method relies only on relative camera motion, not absolute camera poses. The approach is supported by new state-of-the-art results on a *de facto* benchmark - EVIMO2 for relative object-wise depth estimation. The approach aligns with object-centered visual representations, which are beneficial for various tasks [2, 29, 35]. **Code will be made publicly available.**

2. Related Work

Previous work on distance estimation using event cameras does not focus on behavioral strategies to extract information about the scene's structure - like relative distances among objects. We review general active perception methods (not necessarily relying on event data) and common depth estimation approaches using single or stereo event cameras.

Active perception. Active perception assumes the observer to be active - "The purpose of the activity is to manipulate the constraints underlying the observed phenomena in order to improve the quality of the perceptual results" [1]. The constraints of fixation and tracking have been studied in [12,13]. Following similar principles, Burner et al. [8], for the first time, provides a closed-form solution for the estimation of the distance from a sliding-window of time-to-contact and inertial measurements (IMU). Their closed-form solution relies on non-constant acceleration during the time interval of computation. In contrast, our approach computes nearly instantaneous depth estimates over small time intervals, requiring only translational motion and no acceleration constraints. Battaje et al. [5] address distance estimation with RGB cameras by pre-selecting a region and introducing an action to enhance visual processing. Similarly, we exploit the dynamic nature of event-based vision sensors for distance estimation through active vision.

Event based distance estimation Distance estimation has been a long-standing challenge for event cameras. We review distance estimation methods using Multi-View Stereo, Deep Learning, Neuromorphic Processing, and Active Vision.

Multi-View Stereo. Depth estimation with stereo event cameras relies on the epipolar constraint and the assumption that events occur simultaneously on both cameras. Zhou et al. propose an event based SLAM approach [41]. In particular, they reformulate temporal coincidence of events using the compact representation of space-time provided by time surfaces [27]. Inspired by [34], Ghosh et al. [19] circumvent the challenge of accurate event association by leveraging the sparsity of events and by exploiting the continuity of camera viewpoints. Using Space Sweeping, it builds ray density Disparity Space Images (DSIs) from each camera data and fuses them into one DSI [10]. A small set of work exploits Multi-View with only a single event camera. Here, event correspondences across time are established via alignment [16, 17] and under the limiting assumptions that the camera pose is known and the scene is static. Rebecq et al. [34] similarly relies on knowledge about the camera's absolute motion trajectory. A depth value is associated to each event resulting in a semi-dense depth map.

Deep Learning. Hidalgo-Carrio et al. [23] introduced the first supervised method to learn dense monocular depth from event data, followed by various self-supervised methods for dense depth estimation. Zhu et al. [43] exploit cross-modal consistency between frames and aligned events, while Zhu et al. [42] predict egomotion and depth by minimizing motion blur when events are projected onto the image plane. The supervision signal in the latter comes from motion compensation

Method	Algorithmic Approach	Monocular Camera	Input Data	Absolute Cam. Pose	Relative Cam. Motion	Data Structure of Estimated Depth	Freq. [in Hz]
E2Depth [23]	Deep Learning	✓	Events	✗	✗	Pixel-wise Depth	20
EMVS [34]	Multi-View Stereo	✓	Events	✓	✓	Event-wise Depth	1
Ours	Active Vision	✓	Events	✗	✓	Object-wise Depth	20

Table 1. Literature review on event-based depth estimation approaches. We characterize each algorithm that was used for our evaluation according to its properties: algorithmic approach, monocular vs. stereo camera setup, input data, required additional sensor information such as absolute camera pose or relative camera motion, the format of the algorithm's depth estimates as well as its frequency it was evaluated on.

through event alignment. A different line of work by Rudnev et al. [36] and Hwang et al. [25] investigates how NeRFs could be reconstructed from event data.

3. Distance Estimation via Region-wise Event Alignment

We present our approach for distance estimation from event data. We start with revising key aspects of the probabilistic model *"the Spatio-Temporal Poisson Point Process for event alignment"* from Gu et al. [21] and propose small - yet effective changes that increase robustness with respect to variations in speed and scene structure. This framework is applied to object-wise event alignment, detailed in Section 3.1.1. Different from classical event alignment approaches [15, 17, 33, 38], our goal is not to recover the camera's motion. Instead, we introduce a novel compensatory rotational motion that stabilizes the motion within a region of interest. While other approaches for event alignment could be adopted for local alignment, we build on [21], as this algorithm is currently best performing and methodological consistent with ideas introduced in this work. More specifically, events are aligned by minimizing the pixel-wise entropy resulting in reduced brightness changes per pixel. This implements an ecological principle that aims at reducing perceived brightness changes on the human eye's retina though gaze stabilization [3, 14, 28]. In Section 3.2 we introduce a novel formulation, that relates a rotational velocity estimated by local event alignment to object distance. Intuitively speaking, the smaller the rotational motion is to align a specific object-region, the farther the object-region away (see Figure 1). This approach for the first time allows distance estimation, without computing explicit event-to-event correspondences or knowledge of the camera's absolute pose.

3.1. Event Alignment

Event alignment describes the process of finding a transformation $\mathcal{T}_\omega$ that maps events triggered by the same world point to the same pixel location of the camera sensor. We define

$$\mathcal{O} = [\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_N], \tag{1}$$

to be a set of $N = |\mathcal{O}|$ events, where each element $\mathbf{o}_i = (o_i^x, o_i^t, o_i^p)$ comprises a pixel location o_i^x, timestamp o_i^t at which the event occurs and a polarity o_p determining the direction in brightness change. Most commonly, event alignment is formulated as an optimization problem over rotational camera motion parameters ω and translational parameters v,

$$\hat{\omega}, \hat{v} = \arg\max_{\omega, v} p\big(\mathcal{T}_{\omega,v}(\mathcal{O})\big). \tag{2}$$

While there have been proposed a number of different loss functions for solving this optimization problem, we implement the loss by Gu et al. In their method, Gu et al [21]. model an aligned event stream at a particular pixel location as a *Poisson Point Process*. Based on this model, a maximum likelihood approach is developed to register events that are initially unaligned. We find the transformations $\mathcal{T}$ of the observed events $\mathcal{O}$ that make them as likely as possible under the model:

$$p\big(\mathcal{T}_{\omega,v}(\mathcal{O})\big) = \prod_{\mathbf{x}\in\mathcal{X}} \mathcal{NB}\big(k_{\mathbf{x}}(\mathcal{T}_{\omega,v})\big). \tag{3}$$

Here, $k_{\mathbf{x}}$ denotes the number of events occurring at a location $\mathbf{x}$ and is modeled with a negative binomial distribution $\mathcal{NB}(\cdot)$[1]. In Gu et al., $\mathcal{O}$ is defined as a set of N discrete events. In contrast, we redefine $\mathcal{O}$ as a set of events occurring within a fixed time interval ΔT. This modification aligns with the classical Poisson Point Process definition, better capturing event dynamics. Its effectiveness is demonstrated in Sec. 4.

3.1.1 Object-wise Event Alignment

Events arise by the relative motion of the camera and the scene. The events recorded at a single sensor loca-

[1]The negative binomial distribution arises as a mixture of Poisson distributions where the Poisson rate parameter itself follows a gamma distribution. In other words, we can view the negative binomial as a Poisson(λ) distribution, where λ is itself a random variable, distributed as a gamma distribution.

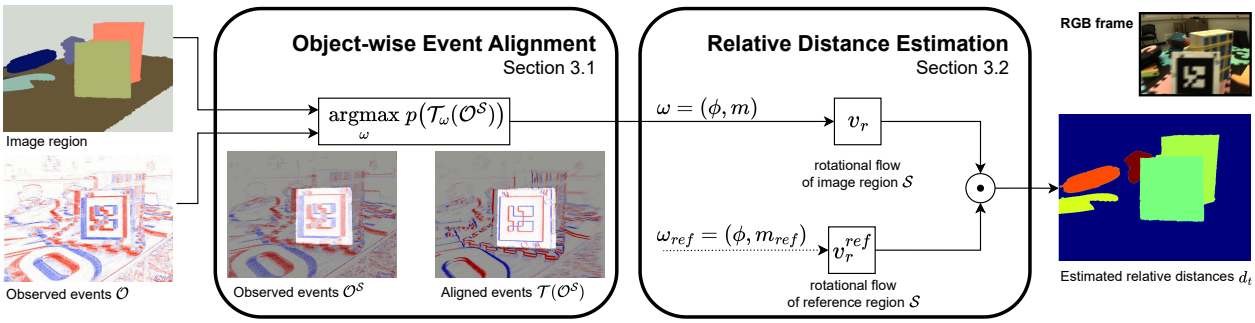


Figure 2. Overview of our algorithm. We estimate relative object-wise-distance from active event alignment. Given a set of events, we process alignment in a object-wise fashion. Object regions maybe determined by a provided object segmentation mask or a default segmentation mask (e.g honeycomb) without semantic information. The obtained relation between different angular velocities within the image plane determine the respective object-wise relative depth.

tion seldom correspond to the same world point. Classical event alignment algorithms aim to register events triggered by the same world point through estimating the camera's motion. Opposed to classical event alignment, we do not aim at recovering the camera's motion through event alignment, instead we wish to find a rotational velocity w that locally leads to accurate event alignment. Although the estimated rotational velocity does not relate to the true camera motion, that triggered the observed events, we will show in Sec. 3.2 that this rotation carries valuable information about relative distances across different image regions.

The rotational velocity w is estimated via region-wise event alignment, that is:

$$\hat{\omega} \;=\; \arg\max_{\omega} p\big(\mathcal{T}_\omega(\mathcal{O}^\mathcal{S})\big), \tag{4}$$

$\mathcal{S}$ denotes the local image region. This optimization leverages the concept that, for rigid camera motion and fronto-parallel planar scene regions, a camera translation[2] can be well approximated by a camera rotation, and vice versa [30]. Limiting the transformation $\mathcal{T}$ to rotation only motions, comes with the benefit of a significantly reduced amount of parameters to be estimated, on the other hand if the optimization is confined to a small image region only, this may lead in unstable rotation estimates, that do not serve as a surrogate of the true camera translation. We propose a strategy to perform object-wise alignment in two steps: (1) we determine a global *velocity direction* that performs alignment across all objects present in the scene and (2) we determine the *velocity's magnitude* that aligns the events of a specific object. This strategy takes the best of both worlds - it determines the global velocity direction across a large image region while assessing velocity's magnitude in specific, potentially smaller regions, thereby preserving robust rotation estimates.

[2]In case of additional camera rotation, IMU information is used to remove the rotational motion part.

Estimation of the global velocity's direction. According to physical rules of perspective geometry, a local motion direction is independent upon the scene depth in case of pure camera translation [6]. Thus for planar scenes the motion direction can be well approximated via a rotational camera motion [30]. In scenes with significant depth variations, objects that are closer tend to exhibit larger motion magnitudes, while those that are farther away show smaller motion magnitudes on the image plane. One option to deal with unknown depth is to estimate pixel-wise depth values through maximum likelihood estimation alongside ω. This would drastically increase the number of parameters to be estimated. Instead, we adopt a partially Bayesian approach by maximizing the marginal likelihood of $k_\mathbf{x}$ under the velocity's magnitude m. The angular velocity is expressed in polar form $\omega = (m, \phi)$. Consequently, the direction ϕ can be determined by integrating over the unknown magnitude m:

$$p\big(\mathcal{T}_\omega(\mathcal{O}^\mathbf{x})\big) = \int_m p(k_\mathbf{x}(m,\phi)|m) \cdot p(m)\, dm \tag{5}$$

$$= \int_m \mathcal{NB}_{r,q}\big(k_\mathbf{x}(m,\phi)\big) \cdot \mathcal{U}_{[0,m_{\max}]}\, dm \tag{6}$$

$$= \int_m \mathcal{NB}_{r,q}\big(k_\mathbf{x}(m,\phi)\big)\, dm \tag{7}$$

An estimate of the velocity's direction ϕ is obtained by maximizing the probability of aligned events:

$$\hat{\phi} \;=\; \arg\max_{\phi} p\big(\mathcal{T}_\omega(\mathcal{O})\big) \;=\; \arg\max_{\phi} \prod_{\mathbf{x}\in\mathcal{X}} p\big(\mathcal{T}_\omega(\mathcal{O}^\mathbf{x})\big). \tag{8}$$

The maximization is applied across all pixel locations $\mathbf{x} \in \mathcal{X}$ on the image plane. Afterwards, depth dis-

continuities are handled by calculating the velocity's magnitude (speed) for each object.

Estimation of the velocity's magnitude. Given the estimate of the velocity's direction $\hat{\phi}$, the object's magnitude is estimated in a similar fashion:

$$\hat{m} = \arg\max_{m} p\big(\mathcal{T}_{\omega}(\mathcal{O}^{\mathcal{S}})\big). \tag{9}$$

The maximization is applied across all pixel locations $\mathbf{x} \in \mathcal{S}$ on within an object region. The advantage of sequential alignment is two-fold. First, obtaining an estimate of the velocity's direction across the image introduces an additional geometric constraint. Second, this approach significantly reduces the number of parameters that need to be optimized.

3.2. Relative Distance Estimation

The estimated rotational velocity ω counteracts the camera's translational movement, stabilizing the image within a specific region $\mathcal{S}$. If the object region is distant, stabilization requires minimal compensatory rotational movement. In contrast, if the object region is nearby, greater compensatory rotation is needed for accurate alignment. Mathematically, we can express this stabilization in terms of optical flow[3]. To formalize this, we make the following model assumptions:

i) Fronto-parallel planar object regions
ii) Zero translational motion along depth axis

$$\frac{1}{z}\boldsymbol{v_t} \overset{\text{i),ii)}{=} -\boldsymbol{v_r} \tag{10}$$

$$z = -\boldsymbol{v_r^+}\boldsymbol{v_t}. \tag{11}$$

Numbers over equality signs give the assumption that is invoked. Object-wise alignment results in zero local flow. More specifically, the rotational flow $\boldsymbol{v_r}$ compensates the translational flow $\boldsymbol{v_t}$ leading to zero local flow, which is indicated by Eq (10). This concept enables a behavior-driven approach to estimating relative distances from event camera data. By relating the two rotational flow vectors, the relative distance d between two objects can be inferred:

$$d = \frac{z}{z^{\text{ref}}} \overset{\text{ii)}}{=} \boldsymbol{v_r^{\text{ref}}}\boldsymbol{v_r^+}, \tag{12}$$

where $\boldsymbol{v_r^+}$ is the pseudo-inverse of the estimated rotational flow vector, and $\boldsymbol{v_r^{ref}}$ is the rotational flow vector estimated for the reference object. The reference object is defined as the largest region in the scene. Assumption ii) renders the translational flow invariant to the object's position. Consequently, the division can be simplified by canceling out translational flow.

[3]Optical flow $\boldsymbol{v}$ represents the camera's projected motion onto the image plane. Given the camera's motion, Horn et al. [24] provide equations to determine the flow at a specific pixel location.

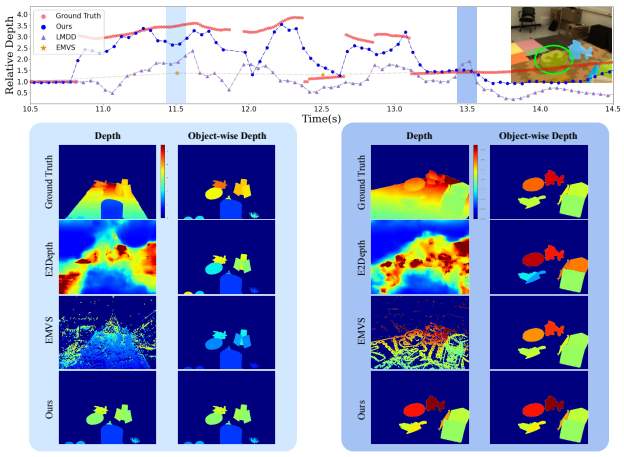


Figure 3. **Qualitative results** of object-wise relative depth estimation over time. Top: The line plot shows the relative depth estimation of the object with ID 24 (highlighted with a green circle) of the event sequence *scene_03_00_000000*.

3.2.1 Recursive Bayesian Filtering

To maintain temporal consistency despite varying camera motions and occlusions, we employ Bayesian filtering. The recursive nature of Bayes filters enables continuous processing of measurements as they arrive [40]. A belief over the relative distance is propagated over time in two steps: 1) prediction and 2) update using new observation o_t.

$$p(d_t|o_{t-1}) = \int p(d_t|d_{t-1})p(d_{t-1}|o_{t-1})\, dd_{t-1} \tag{13}$$

$$p(d_t|o_t) = \eta p(o_t|d_t)p(d_t|o_{t-1}) \tag{14}$$

This notation highlights the recursive nature: the posterior from the previous step $p(d_{t-1}|o_{t-1})$ is used to predict the current posterior at time t, which is then updated in Eq. (14) using the latest observation. The estimated relative distance is treated as a Gaussian distribution, and a 1D Kalman filter is used to track the distance and its variance over time. The variance is modeled as $\sigma^2 = \frac{1}{|\boldsymbol{v_r}|^2}$, depending on the magnitude of the compensatory rotational motion $\boldsymbol{v_r}$. Further implementation details are in suppl. material.

4. Experiments

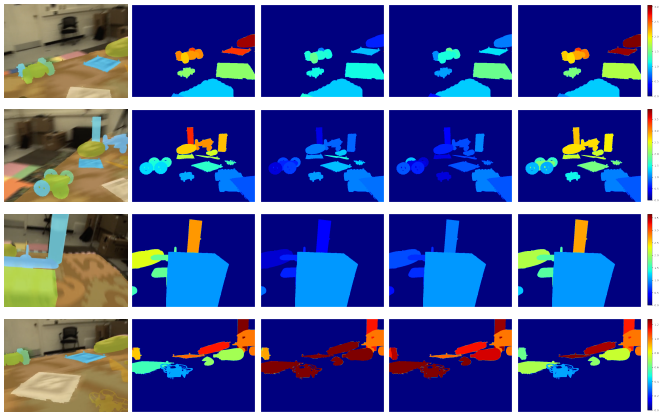
We present results for monocular distance estimation on the EVIMO2 [7], and improvements in camera velocity estimation on the DAVIS 240C [31]. We describe the datasets, experimental setup, evaluation metrics, and discuss our findings on event-based distance estimation via object-wise alignment.

EVIMO2 [7] is a widely used dataset for evaluating event-based algorithms on depth estimation and object segmentation. The sequences feature the objects on a board, under different lighting conditions. The recording setup of EVIMO2 includes three VGA-resolution event-based cameras, namely, a Samsung DVS Gen3 in the middle, two Prophesee cameras on the left and right side. Additionally, a Flea3 frame-based RGB camera is also mounted. In this work, we use the data from the Samsung DVS Gen3 event camera and from the IMU embedded in the Prophesee camera. EVIMO2 has various types of sequences for different event-based vision tasks under different conditions. In our experiments, we adopt the ones for structure from motion, where the GT camera poses (not used), object segmentation and scene depth are available. To showcase event cameras' capabilities in varying lighting, we use ten sequences recorded in normal light and two out of five under low-light conditions, totaling 147s and 82s, respectively. Only two low-light sequences include IMU sensor data and object masks.

The DAVIS 240C dataset [32] includes four 60-second sequences that showcase rotational motion: shapes, poster, boxes, and dynamic. Initially, the camera rotates around each axis at increasing speeds, followed by free rotation in 3 degrees of freedom (3-DOF). The shapes, poster, boxes sequence capture static indoor scenes with varying levels of texture complexity, generating approximately 20 to 200 million events. The dynamic sequence, features a dynamic scene and produces around 70 million events.

Evaluation metrics. Following the commonly used evaluation protocols of event-based and frame-based depth estimation in literature [20, 26, 37], four error metrics are employed: RMSE (linear), RMSE (log), squared relative distance (SRD) and absolute relative distance. The explicit mathematical formulation of each utilized measurement can be found in [11]. Accuracy is measured using three thresholds that determine the percentage of estimates with relative accuracy δ below each threshold. If $\delta = 1$, the estimate matches the ground truth perfectly. The higher the threshold the more values fall into the bin of relative accuracy lower than the provided threshold.

Implementation details. We use a fixed ΔT of 0.05 seconds (20Hz) for all our experiments. All model parameters required for alignment remain consistent with those from Gu et al. [21], but we reduce the maximum iterations from 250 to 50 to speed up event alignment. The process noise's standard deviation σ of the Kalman filter is set to 0.1.



(a) (b) (c) (d) (e)

Figure 4. Qualitative results of depth estimation on the EVIMO2 dataset, with framewise results on four exemplary video sequences: (a) Events and segmentation mask. (b) Ground truth. (c) E2Depth [23]. (d) EMVS [34]. (e) Ours.

4.1. Results

We evaluate our approach using the publicly available EVIMO2 dataset. We present the results and an ablation study to demonstrate our method for distance estimation from event camera data. Our approach utilizes event data, rotational velocity from the IMU to compensate for camera rotation, and object segmentation masks to identify regions of local alignment. Ablation studies show that while object segmentation masks can be beneficial, the proposed algorithm effectively estimates scene depth even with arbitrary masks.

Object-wise distance estimation. All monocular depth estimation methods, including ours, rely solely on visual information from events without using frame-based data. Unlike traditional multi-view stereo methods requiring known camera viewpoints, EMVS [34] estimates semi-dense 3D structures from event cameras using known motion trajectories.Traditional methods require absolute positional information, such as trajectories or camera poses. In contrast, our approach only needs relative camera motion, specifically angular velocity, which is readily available from on-board IMU sensors, unlike absolute camera pose that needs external systems like Motion Capture. Hidalgo-Carrió et. al use only events to learn a dense depth map. Unlike earlier learning-based methods, they propose a recurrent network architecture to maintain temporal consistency.

Qualitative results for object-wise relative depth estimation are presented in Fig. 3. We present relative depth estimation results at two distinct timestamps: 11.5s and 13.5s, respectively. The GT relative depth is shown and compared to three different approaches that

	Method	Freq. [Hz]	Error ↓				Accuracy ↑		
			RMSE (linear)	RMSE (log)	ARD	SRD	$\delta < 1.25$ [%]	$\delta < 1.25^2$ [%]	$\delta < 1.25^3$ [%]
sfm (10 Seq.)	E2Depth [23]	20	1.042	0.667	0.428	0.606	35.791	54.284	71.035
	EMVS [34]	1	0.871	0.621	0.401	0.496	40.775	59.460	70.369
	Ours	20	**0.725**	**0.448**	**0.273**	**0.293**	**56.308**	**80.786**	**90.540**
sfm (low light) (2 Seq.)	E2Depth [23]	20	1.034	**0.567**	0.414	0.609	**47.276**	**64.790**	**79.701**
	EMVS [34]	1	0.928	0.630	**0.404**	0.424	39.860	61.111	71.018
	Ours	20	**0.883**	0.930	0.414	**0.420**	41.077	56.728	68.168

Table 2. Relative object-wise depth estimation of static scenes with multiple objects. Accuracy comparison on event sequences from EVIMO2 [7] - sfm and sfm (low light).

perform monocular depth estimation from an event stream: E2Depth [23], EMVS [34] and Ours. While each method yields depth estimates in various formats - dense depth, sparse depth, and object-wise depth - we unify results into a common format for consistent evaluation. In addition, more comparisons are shown in Fig. 4, where our method (column d) reports the closest results to GT (column e) while E2depth and EMVS deviate a lot (column b-c).

Table 2 shows the quantitative comparison of accuracy on all test sequences of the structure-from-motion (sfm) and the structure-from-motion split in low light conditions (sfm low light) of EVIMO2. On *smf*, our approach improves RMSE (linear) by over 16%. On *sfm low light*, it achieves a 5% improvement. In terms of RMSE (log), E2Depth outperforms all other methods. The performance gain on the *smf* split is less pronounced in low light conditions. We identified two reasons for this lower performance. First, low light conditions cause significantly higher event sparsity. Second, the ratio of hot pixels to informative events deteriorates. Hot pixels are sensor failures that consistently "fire" regardless of camera or scene motion. Compared to EMVS and E2Depth, our approach, which relies on *local event alignment*, can be affected by pixel failures in the dynamic vision sensor. Although hot pixels, being locally stable, can hinder event alignment, our remarkable 16% performance gain highlights the efficacy of combining dynamic vision systems like event cameras with active vision approaches.

Object-wise event alignment utilises a modified version of the original Spatio-Temporal Poisson Point Process for distance estimation. While the original version performs event alignment given a fixed number of events - typically 30K events [16, 16, 21, 33], we perform event alignment of all events within a fixed time interval ΔT. Table 3 shows, that defining the Poisson Point Process for a fixed time interval not only improves performance, also it is consistent with the original definition of the Poisson Point Process [39].

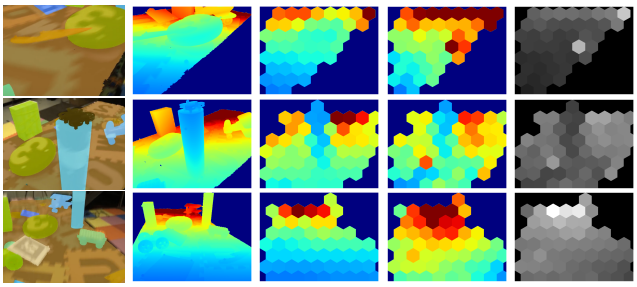
4.2. Ablation Study

In Sec. 4.1 we discussed our results on object-wise relative distance estimation. These results are based on two key assumptions: (1) the presence of fronto-parallel planar object regions, and (2) zero translational motion along the depth axis (refer to Sec. 3.2). Here we ask, *How sensitive is our algorithm to the quality of predefined object regions?* and *How does z-motion affect the accuracy of relative depth estimates?* To address these questions, we first present qualitative results of relative depth estimation using segmentation masks that are entirely object-independent. Then, we examine our algorithm's performance, with respect to its sensitivity to camera motion, with a particular focus on z-motion.

Object regions. Fig. 5 qualitatively shows region-wise depth estimation results of our algorithm. To avoid relying on informative object masks, that may

	Method	e_{wx}	e_{wy}	e_{wz}	σ_{ew}	RMS	RMS%
boxes	CMax [16]	7.38	6.66	6.03	9.04	9.08	0.66
	AEMin [33]	6.75	5.19	5.78	7.77	7.81	0.56
	EMin [33]	6.55	4.40	5.00	7.00	7.06	0.51
	Ours ($Ne = 30K$)	6.72	3.93	4.55	6.64	6.73	0.49
	Ours ($\Delta T = 0.015$)	**5.68**	**3.81**	**3.92**	**6.32**	**6.34**	**0.46**
poster	CMax [16]	13.45	9.87	5.56	13.39	13.45	0.74
	AEMin [33]	12.57	7.89	5.63	12.35	12.36	0.68
	EMin [33]	11.83	7.31	4.37	10.85	10.86	0.60
	Ours ($Ne = 30K$)	11.78	6.33	3.67	10.30	10.37	0.57
	Ours ($\Delta T = 0.015$)	**9.37**	**5.77**	**3.49**	**9.15**	**9.21**	**0.51**
dynamic	CMax [16]	4.93	4.82	4.95	7.11	7.13	0.71
	AEMin [33]	5.02	3.88	4.55	6.16	6.19	0.62
	EMin [33]	4.78	3.72	3.73	5.33	5.39	0.54
	Ours ($Ne = 30K$)	4.42	3.61	**3.49**	**5.15**	**5.19**	**0.52**
	Ours ($\Delta T = 0.015$)	**4.29**	**3.60**	3.97	5.27	5.33	0.53
shapes	CMax [16]	31.19	26.83	38.98	55.86	55.87	3.94
	AEMin [33]	22.22	18.78	35.41	55.43	55.44	3.91
	EMin [33]	21.22	15.87	25.57	42.22	42.22	2.98
	Ours ($Ne = 30K$)	20.73	13.95	17.69	25.88	25.89	1.83
	Ours ($\Delta T = 0.015$)	**10.32**	**5.61**	**4.68**	**10.15**	**10.16**	**0.69**

Table 3. Comparison of angular velocity accuracy on the rotation sequences from DAVIS 240C [31].



(a) (b) (c) (d) (e)

Figure 5. Qualitative evaluation of region-wise distance estimation w/o object masks. We use a honeycomb grid to define pixel regions for depth estimation. Relative distances and confidence maps are shown in grayscale (white = low confidence, black = high confidence). (a) Events and segmentation masks. (b) Original ground truth. (c) Ground truth using honeycomb regions. (d) Our method using honeycomb regions. (e) $\pm 3\sigma$ confidence interval.

provide strong priors in terms of depth consistency within a particular region, we employed a honeycomb grid to define pixel regions independent of any object or scene information. We show that even without properly defined image regions we can determine the relative depth reasonably well. To the best of our knowledge our algorithm for the first time models the estimate's confidence plus tracks the confidence over time. The $\pm 3\sigma$ confidence interval is shown in Fig. 5e. It nicely captures deviating estimates from the GT. The resolution of the acquired depth estimates is undoubtedly influenced by the grid size of the masks. However, prior work emphasizes that region-wise estimates are crucial for distance estimation, while precise object boundaries are not essential [2, 29].

Zero translational motion along depth axis. The assumption of zero linear velocity along the z-axis (no forward/backward camera motion) allows the approximation of a camera translation with a rotation. However, the EVIMO2 dataset includes event sequences with arbitrary camera movement, meaning our assumption of $W = 0$ holds only in very few cases. Fig. 6 evaluates distance estimation via local alignment dependent upon the amount of linear velocity along the z-axis. As expected, error decreases with lower z-axis velocity, aligning better with our assumption.

Failure case analysis. Relative distance is estimated between two static objects. If one object is moving, the distance cannot be accurately inferred. Interestingly, relative depth estimates between neighboring static objects remain unaffected. This suggests potential new research directions, such as detecting

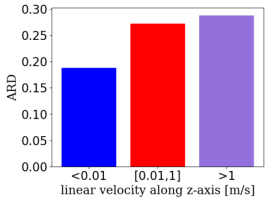


Figure 6. Evaluation of distance estimates based on the camera's z-motion: low, normal, and high speed.

object motion by analyzing discontinuities in relative distances between multiple objects. Specifically, given three objects, we can compute their relative distances. If these distances remain constant, none of the object is moving. Conversely, if the relative distances change, at least one of the objects is in motion. Fig. 7 shows that, in the presence of object motion, the relative depth estimates of neighboring static objects remain unaffected, while the estimated relative distance of the moving object diverges from the ground truth.

(a) RGB frame (b) Estimated relative depth d_t (c) Ground truth relative depth d_t

Figure 7. Failure case: Relative distance estimation fails for moving objects, as illustrated by the flying drone.

5. Conclusion

Inspired by visual ecology, we propose the first event-based approach for relative distance estimation that combines dynamic vision sensors with a behavioral strategy to infer relative distances between objects. Firstly, we introduce a novel optimization pipeline that estimates a rotational motion aimed at achieving *object-wise event alignment*. This rotation does not recover the actual camera motion but is rather a virtual adjustment designed to align events locally. Secondly, object-wise *relative distance* is determined by comparing the corresponding rotational flow vectors.

Compared to frame-based cameras event cameras capture visual information efficiently by reducing redundant data. However, processing methods are still developing. Local alignment, which computes compensatory rotational movements to extract visual data like depth, bear a high potential for novel, efficient vision algorithms. Our approach reduces computational load by using behavioral strategies, such as gaze stabilization, to streamline sensory input processing.

Acknoledgements

The project was supported through the X-Student Research Group in cooperation with the Berlin University Alliance. In addition, this work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

References

[1] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *IJCV*, 1:333–356, 1988. 2

[2] Dana H Ballard. Animate vision uses object-centered reference frames. In *Advanced neural computers*, pages 229–236. Elsevier, 1990. 2, 8

[3] H.B. Barlow, T.P. Kaushal, and G.J. Mitchison. Finding Minimum Entropy Codes. *Neural Computation*, 1(3):412–423, 09 1989. 2, 3

[4] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989. 2

[5] Aravind Battaje and Oliver Brock. One object at a time: Accurate and robust structure from motion for robots. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 3598–3603. IEEE, 2022. 2

[6] Pia Bideau, Erik Learned-Miller, Cordelia Schmid, and Karteek Alahari. The right spin: Learning object motion from rotation-compensated flow fields. *International Journal of Computer Vision*, 132(1):40–55, Jan 2024. 4

[7] L. Burner, A. Mitrokhin, C. Ye, C. Fermüller, Y. Aloimonos, and T. Delbruck. Evimo2: An event camera dataset for motion segmentation, optical flow, structure from motion, and visual inertial odometry in indoor scenes with monocular or stereo algorithms. In *arXiv e-prints*, May 2022. 5, 6, 7

[8] Levi Burner, Nitin J Sanket, Cornelia Fermüller, and Yiannis Aloimonos. Ttcdist: Fast distance estimation from an active monocular camera using time-to-contact. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 4909–4915. IEEE, 2023. 2

[9] Clément Cabriel, Tual Monfort, Christian G. Specht, and Ignacio Izeddin. Event-based vision sensor for fast and dense single-molecule localization microscopy. *Nature Photonics*, 17(12):1105–1113, Dec 2023. 1

[10] R.T. Collins. A space-sweep approach to true multi-image matching. In *CVPR*, pages 358–363, 1996. 2

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014. 6

[12] Cornelia Fermüller and Yiannis Aloimonos. Tracking facilitates 3-d motion estimation. *Biological Cybernetics*, 67(3):259–268, 1992. 2

[13] Cornelia Fermüller and Yiannis Aloimonos. The role of fixation in visual motion analysis. *International Journal of Computer Vision*, 11(2):165–186, 1993. 2

[14] David J Field. What is the goal of sensory coding? *Neural computation*, 6(4):559–601, 1994. 2, 3

[15] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *CVPR*, pages 12280–12289, 2019. 3

[16] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *CVPR*, pages 3867–3876, 2018. 2, 7

[17] Guillermo Gallego and Davide Scaramuzza. Accurate angular velocity estimation with an event camera. *IEEE Robotics and Automation Letters*, 2(2):632–639, 2017. 2, 3

[18] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, May 2024. 1

[19] Suman Ghosh and Guillermo Gallego. Multi-event-camera depth estimation and outlier rejection by refocused events fusion. *Advanced Intelligent Systems*, page 2200221, sep 2022. 2

[20] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019. 6

[21] Cheng Gu, Erik Learned-Miller, Daniel Sheldon, Guillermo Gallego, and Pia Bideau. The spatio-temporal poisson point process: A simple model for the alignment of event camera data. In *ICCV*. IEEE, oct 2021. 3, 6, 7

[22] Eman Hassan, Zhuowen Zou, Hanning Chen, Mohsen Imani, Yahya Zweiri, Hani Saleh, and Baker Mohammad. Efficient event-based robotic grasping perception using hyperdimensional computing. *Internet of Things*, 26:101207, 2024. 1

[23] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *Int. Conf. 3D Vision (3DV)*, pages 534–542. IEEE, 2020. 2, 3, 6, 7

[24] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981. 5

[25] Inwoo Hwang, Junho Kim, and Young Min Kim. Evnerf: Event based neural radiance field. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 837–847, 2023. 3

[26] Huaizu Jiang, Gustav Larsson, Michael Maire Greg Shakhnarovich, and Erik Learned-Miller. Self-supervised relative depth learning for urban scene understanding. In *ECCV*, pages 19–35, 2018. 6

[27] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad B. Benosman. Hots: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1346–1359, 2017. 2

[28] Michael F Land and Dan-Eric Nilsson. *Animal eyes*. OUP Oxford, 2012. 1, 3

[29] Ajay Mishra, Yiannis Aloimonos, and Cornelia Fermuller. Active segmentation for robotics. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 3133–3139. IEEE, 2009. 2, 8

[30] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 1–9, 2018. 4

[31] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. *Int. J. Robot. Research*, 36(2):142–149, 2017. 5, 7

[32] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 6

[33] Urbano Miguel Nunes and Yiannis Demiris. Entropy minimisation framework for event-based vision model estimation. In *ECCV*, 2020. 3, 7

[34] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *IJCV*, 126(12):1394–1414, 2018. 2, 3, 6, 7

[35] Nicolas Roth, Martin Rolfs, Olaf Hellwich, and Klaus Obermayer. Objects guide human gaze behavior in dynamic real-world scenes. *bioRxiv*, pages 2023–03, 2023. 2

[36] Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. Eventnerf: Neural radiance fields from a single colour event camera. In *CVPR*, pages 4992–5002, 2023. 3

[37] Peilun Shi, Jiachuan Peng, Jianing Qiu, Xinwei Ju, Frank Po Wen Lo, and Benny Lo. Even: An event-based framework for monocular depth estimation at adverse night conditions. *arXiv preprint arXiv:2302.03860*, 2023. 6

[38] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *ICCV*, pages 7244–7253, 2019. 3

[39] Roy L Streit and Roy L Streit. *The poisson point process*. Springer, 2010. 7

[40] Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002. 5

[41] Yi Zhou, Guillermo Gallego, and Shaojie Shen. Event-based stereo visual odometry. *IEEE Transactions on Robotics*, 37(5):1433–1450, 2021. 2

[42] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, pages 989–997, 2019. 2

[43] Junyu Zhu, Lina Liu, Bofeng Jiang, Feng Wen, Hongbo Zhang, Wanlong Li, and Yong Liu. Self-supervised event-based monocular depth estimation using cross-modal consistency. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 7704–7710. IEEE, 2023. 2