Entropy Minimisation Framework for Event-based Vision Model Estimation

Urbano Miguel Nunes and Yiannis Demiris

Personal Robotics Laboratory, Department of Electrical and Electronic Engineering,
Imperial College London, London, UK
{um.nunes, y.demiris}@imperial.ac.uk

Abstract. We propose a novel Entropy Minimisation (EMin) framework for event-based vision model estimation. The framework extends previous event-based motion compensation algorithms to handle models whose outputs have arbitrary dimensions. The main motivation comes from estimating motion from events directly in 3D space (e.g. events augmented with depth), without projecting them onto an image plane. This is achieved by modelling the event alignment according to candidate parameters and minimising the resultant dispersion. We provide a family of suitable entropy loss functions and an efficient approximation whose complexity is only linear with the number of events (e.g. the complexity does not depend on the number of image pixels). The framework is evaluated on several motion estimation problems, including optical flow and rotational motion. As proof of concept, we also test our framework on 6-DOF estimation by performing the optimisation directly in 3D space.

Keywords: Event-based vision · Optimisation framework · Model estimation · Entropy minimisation

1 Introduction

Event-based cameras asynchronously report pixel-wise brightness changes, denominated as *events*. This working principle allows them to have clear advantages over standard frame-based cameras, such as: very high dynamic ranges (> 120 dB vs. ≈ 60 dB), high bandwidth in the order of millions of events per second, low latency in the order of microseconds. Thus, event-based cameras offer suitable and appealing traits to tackle a wide range of computer vision problems [3,6,7,14,18]. However, due to the different encoding of visual information, new algorithms need to be developed to properly process event-based data.

Significant research has been driven by the benefits of event-based cameras over frame-based ones, including high-speed motion estimation [3,6,7,28], depth estimation [2,10,18,24,26] and high dynamic range tracking [5,14,16,21], with applications in robotics [8,9] and SLAM [11,22], among others. In particular, event-based motion compensation approaches [6,7,16,26] have been successful in tackling several estimation problems (*e.g.* optical flow, rotational motion and depth estimation). These methods seek to maximise the event alignment of point

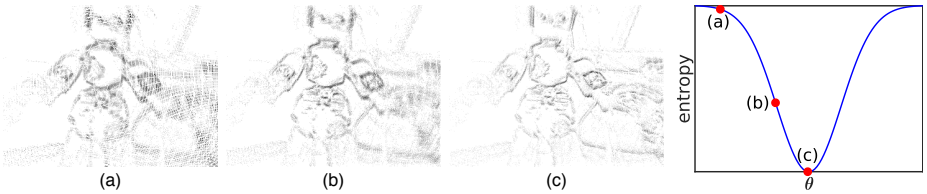

Fig. 1. Entropy measure examples of modelled events according to candidate parameters. (a)-(c) Projected events modelled according to candidate parameters (for visualisation purposes only) and respective entropy measures (right). As a particular case, our framework can also produce motion-corrected images, by minimising the events' entropy. Events generated from a moving DAVIS346B observing an iCub humanoid [15]

trajectories on the image plane, according to some loss function of the warped events. By solving this optimisation, the parameters that better compensate for the motion observed can be retrieved, whilst the resultant point trajectories produce sharp, motion-corrected and high contrast images (*e.g.* Fig. 1 (c)). All these event-based motion compensation methods require the warped events to be projected onto an image, where mature computer vision tools can then be applied. For instance, Gallego *et al.* [6] proposed to maximise the variance of the Image of Warped Events (IWE), which is a known measure of image contrast.

We propose a distinct design principle, where event alignment of point trajectories can still be achieved, but without projecting the events onto any particular sub-space. This can be useful in modelling point trajectories directly in 3D space, *e.g.* acquired by fusing event-based cameras with frame-based depth sensors [23]. Thus, instead of using a contrast measure of the IWE as the optimisation loss function, we propose to minimise a dispersion measure in the model's output space, where events are seen as nodes in a graph. Particularly, we will consider the entropy as a principled measure of dispersion. Fig. 1 presents an example of events modelled according to candidate parameters. As shown, modelled events that exhibit a lower dispersion also produce motion-corrected images. Although the main motivation for developing the proposed method comes from being able to model events directly in 3D space, in principle, there is no constraint on the number of dimensions the framework can handle and we will use the term *features* throughout the paper to emphasise this. We note that this work is not focused on how to get such features, which is application/model dependent.

Main contributions: We propose an EMin framework for arbitrary event-based model estimation that only requires event-based data, which can be augmented by other sensors (*e.g.* depth information). This is achieved by modelling the event alignment according to candidate parameters and minimising the resultant model outputs' dispersion according to an entropy measure, without explicitly establishing correspondences between events. In contrast to previous methods, our framework does not need to project the modelled events onto a

particular subspace (*e.g.* image plane) and thus can handle model outputs of arbitrary dimensions. We present a family of entropy functions that are suitable for the optimisation framework, as well as efficient approximations whose complexity is only linear with the number of events, offering a valid trade-off between accuracy and computational complexity. We evaluate our framework on several estimation problems, including 6-DOF parameters estimation directly from the 3D spatial coordinates of augmented events.

2 Related Work

Event-based motion compensation algorithms have been proposed to tackle several motion estimation problems, by solving an optimisation procedure, whereby the event alignment of point trajectories is maximised. These estimation problems include rotation [7], depth [18,26], similarity transformations [16].

Gallego *et al.* [6] proposed a Contrast Maximisation (CMax) framework that unifies previous model estimation methods. The framework maximises a contrast measure, by warping a batch of events according to candidate model parameters and projecting them onto the IWE. It is unifying in the sense that the optimisation procedure is independent of the geometric model to be estimated (*e.g.* optical flow, rotation). Several new loss functions were recently added to this framework in [4]. However, the general working principle was not modified and, for brevity reasons, hereinafter, we will only refer to the CMax framework [6], which we review next.

Contrast Maximisation Framework: An event $e = (\mathbf{x}, t, p)$ is characterised by its coordinates $\mathbf{x} = (x, y)^\mathsf{T}$, the time-stamp it occurred t and its polarity $p \in \{-1, +1\}$, *i.e.* brightness change. Given a set of events $\mathcal{E} = \{e_k\}_{k=1}^{N_e}$, where N_e is the number of events considered, the CMax framework estimates the model parameters $\boldsymbol{\theta}^*$ that best fit the observed set of events along point trajectories. This is achieved by first warping all events to a reference time-stamp t_{ref} according to the candidate parameters $\boldsymbol{\theta}$

$$e_k \rightarrow e'_k: \quad \mathbf{x}'_k = \mathcal{W}(\mathbf{x}_k, t_k; \boldsymbol{\theta}), \tag{1}$$

where $\mathcal{W}$ is a function that warps the events according to the candidate parameters. Then, the warped events are projected onto the IWE and accumulated according to the Kronecker delta function $\delta_{\mathbf{x}_k}(\mathbf{x})$, where $\delta_{\mathbf{x}_k}(\mathbf{x}) = 1$ when $\mathbf{x} = \mathbf{x}_k$ and 0 otherwise:

$$\mathbf{I}'(\mathbf{x}; \boldsymbol{\theta}) = \sum_k^{N_e} b_k \delta_{\mathbf{x}'_k}(\mathbf{x}), \tag{2}$$

such that if $b_k = p_k$ the polarities are summed, whereas if $b_k = 1$ the number of events are summed instead. In [6,7], the variance of the IWE

$$f(\boldsymbol{\theta}) = \sigma^2\left(\mathbf{I}'(\mathbf{x}; \boldsymbol{\theta})\right) = \frac{1}{N_p} \sum_{i,j}^{N_p} \left(i'_{ij} - \mu_{\mathbf{I}'}\right)^2 \tag{3}$$

was proposed to be maximised, where N_p corresponds to the number of pixels of $\mathbf{I}' = (i'_{ij})$ and $\mu_{\mathbf{I}'} = \frac{1}{N_p} \sum_{i,j} i'_{ij}$ is the mean of the IWE. The variance of an image is a well known suitable contrast measure. Thus, the model parameters $\boldsymbol{\theta}^*$ that maximise the contrast of the IWE correspond to the parameters that best fit the event data $\mathcal{E}$, by compensating for the observed motion

$$\boldsymbol{\theta^*} = \arg\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta}). \tag{4}$$

The proposed EMin framework can also be used to estimate arbitrary models, by minimising the entropy of the modelled events. As opposed to the CMax framework, whose optimisation is constrained to the image space, our framework can estimate models whose outputs have arbitrary dimensions, since the entropy measure can be computed on a space of arbitrary dimensions.

Related similarity measures were proposed by Zhu *et al.* [25] for feature tracking, where events were explicitly grouped into features based on the optical flow predicted as an expectation maximisation over all events' associations. Distinctively, we avoid the data association problem entirely, since the entropy already implicitly measures the similarity between event trajectories.

3 Entropy Minimisation Framework

The CMax framework [6] requires the warped events to be projected onto a sub-space, *i.e.* the IWE. This means that the optimisation framework is constrained to geometric modelling and does not handle more than 2D features, *i.e.* IWE pixel coordinates. Thus, we propose an EMin framework for event-based model estimation that does not require projection onto any sub-space. The proposed framework relies on a family of entropy loss functions which measure the dispersion of features related to events in a feature space (Section 3.2). The EMin procedure complexity is quadratic with the number of events, which will motivate an efficient Approximate Entropy Minimisation (AEMin) solution, whose complexity is only linear with the number of events (Section 3.3).

3.1 Intuition from Pairwise Potentials

Events are optimally modelled if in the temporal-image space they are aligned along point trajectories that reveal strong edge structures [6,7]. This means that events become more concentrated (on the edges) and thus, their average pairwise distance is minimised (*e.g.* the average distance is zero if all events are projected onto the same point). This can be measured by using fully connected Conditional Random Field (CRF) models [13], which are a standard tool in computer vision, *e.g.* with applications to semantic image labelling [12]. An image is represented as a graph, each pixel is a node, and the aim is to obtain a labelling for each pixel that maximises the corresponding Gibbs energy

$$E_{\text{Gibbs}} = \sum_{i}^{N_p} \psi_u(l_i) + \sum_{i,j}^{N_p} \psi_p(l_i, l_j), \tag{5}$$

where ψ_u is the unary potential computed independently for each pixel and ψ_p is the pairwise potential. The unary potential incorporates descriptors of a pixel, *e.g.* texture and color, whereas the pairwise potential can be interpreted as a similarity measure between pixels, which can take the form

$$\psi_p(l_i, l_j) = \xi(l_i, l_j) \sum_m^M \omega_m \mathcal{K}_m(\mathbf{f}_i, \mathbf{f}_j), \tag{6}$$

where ξ is some label compatibility function, ω_m represents the weight of the contribution of the kernel $\mathcal{K}_m$ and $\mathbf{f}$ is a d-dimensional feature vector in an arbitrary feature space.

This model can be used to measure the similarity between events, by representing each event as a node of a graph, and then maximising the similarity measure. According to Eq. (6), this is equivalent to maximising the pairwise potentials between all events in the case where we consider one kernel (M=1) and the feature vector corresponds to the modelled event coordinates in the IWE ($\mathbf{f}_k = \mathbf{x}'_k$). This can be formalised by defining the *Potential* energy which we seek to minimise according to parameters $\boldsymbol{\theta}$ as

$$P(\mathbf{f}; \boldsymbol{\theta}) \coloneqq -\frac{1}{N_e^2} \sum_{i,j}^{N_e} \mathcal{K}_{\boldsymbol{\Sigma}}(\mathbf{f}_i, \mathbf{f}_j; \boldsymbol{\theta}). \tag{7}$$

The *Potential* energy measures the average dispersion of events relative to each other in the feature space, *e.g.* if the events are more concentrated, then Eq. (7) is minimised. Based on this formulation, the modelled events are not required to be projected onto any particular space and we can handle feature vectors of arbitrary dimensions. For easy exposition, we will consider the d-dimensional multivariate Gaussian kernel parameterised by the covariance matrix $\boldsymbol{\Sigma}$

$$\mathcal{K}_{\boldsymbol{\Sigma}}(\mathbf{f}_i, \mathbf{f}_j; \boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{f}_i - \mathbf{f}_j)^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{f}_i - \mathbf{f}_j)\right)}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}}. \tag{8}$$

3.2 Description of Framework

Another well-known measure of dispersion is the entropy, which is at the core of the proposed optimisation framework. By interpreting the kernel $\mathcal{K}_{\boldsymbol{\Sigma}}(\mathbf{f}_i, \mathbf{f}_j; \boldsymbol{\theta})$ as a conditional probability distribution $p(\mathbf{f}_i | \mathbf{f}_j, \boldsymbol{\theta})$, we can consider the corresponding *Sharma-Mittal* entropies [20]

$$H_{\alpha,\beta}(\mathbf{f}; \boldsymbol{\theta}) \coloneqq \frac{1}{1-\beta} \left[\left(\frac{1}{N_e^2} \sum_{i,j}^{N_e} \mathcal{K}_{\boldsymbol{\Sigma}}(\mathbf{f}_i, \mathbf{f}_j; \boldsymbol{\theta})^\alpha \right)^\gamma - 1 \right], \tag{9}$$

Algorithm 1 Entropy Minimisation Framework

Input: Set of events $\mathcal{E} = \{e_k\}_k^{N_e}$.
Output: Estimated model parameters $\boldsymbol{\theta}^*$.
Procedure:
1: Initialise the candidate model parameters $\boldsymbol{\theta}$.
2: Model events according to parameters $\boldsymbol{\theta}$, Eq. (13).
3: Compute the entropy $E(\mathbf{f};\boldsymbol{\theta})$ (Sections 3.2 and 3.3).
4: Find the best parameters $\boldsymbol{\theta}^*$ by minimising the entropy $f(\boldsymbol{\theta})$, Eq. (14).

where $\alpha > 0,\ \alpha, \beta \neq 1$ and $\gamma = \frac{1-\beta}{1-\alpha}$. This family of entropies tends in limit cases to *Rényi* R_α $(\beta \to 1)$, *Tsallis* T_α $(\beta \to \alpha)$ and *Shannon* S $(\alpha, \beta \to 1)$ entropies:

$$R_\alpha(\mathbf{f};\boldsymbol{\theta}) \coloneqq \frac{1}{1-\alpha} \log \left(\frac{1}{N_e^2} \sum_{i,j}^{N_e} \mathcal{K}_{\boldsymbol{\Sigma}}(\mathbf{f}_i, \mathbf{f}_j; \boldsymbol{\theta})^\alpha \right), \tag{10}$$

$$T_\alpha(\mathbf{f};\boldsymbol{\theta}) \coloneqq \frac{1}{1-\alpha} \left(\frac{1}{N_e^2} \sum_{i,j}^{N_e} \mathcal{K}_{\boldsymbol{\Sigma}}(\mathbf{f}_i, \mathbf{f}_j; \boldsymbol{\theta})^\alpha - 1 \right), \tag{11}$$

$$S(\mathbf{f};\boldsymbol{\theta}) \coloneqq \frac{1}{N_e^2} \sum_{i,j}^{N_e} \mathcal{K}_{\boldsymbol{\Sigma}}(\mathbf{f}_i, \mathbf{f}_j; \boldsymbol{\theta}) \log \mathcal{K}_{\boldsymbol{\Sigma}}(\mathbf{f}_i, \mathbf{f}_j; \boldsymbol{\theta}). \tag{12}$$

Gallego *et al.* [4] proposed a loss function based on the image entropy, which measures the dispersion over the distribution of accumulated events in the IWE. Instead, we propose to measure the dispersion over the distribution of the modelled events in the feature space (by considering the distributions are given by each modelled event). The proposed measure is actually closer in spirit to the spatial autocorrelation loss functions proposed in [4].

The proposed framework then follows a similar flow to that of the CMax framework [6], which is summarised in Algorithm 1. A set of events $\mathcal{E}$ is modelled according to candidate parameters $\boldsymbol{\theta}$

$$\mathbf{f}_k = \mathcal{M}(e_k;\boldsymbol{\theta}), \tag{13}$$

where $\mathbf{f}_k$ is the resultant feature vector in a d-dimensional feature space associated with event e_k and $\mathcal{M}$ is a known model. Then, we find the best model parameters $\boldsymbol{\theta}^*$, by minimising the entropy in the feature space

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} E(\mathbf{f};\boldsymbol{\theta}), \tag{14}$$

where $E(\mathbf{f};\boldsymbol{\theta})$ is one of the proposed loss functions.

3.3 Efficient Approximation

The complexity of the proposed framework is quadratic with the number of events and thus is more computationally demanding than the CMax framework [6], which is linear with the number of events and the number of pixels of

the IWE. To overcome the increased complexity, we propose an efficient approximation to compute the entropy functions. This is achieved by approximating the kernel $\mathcal{K}_{\mathbf{\Sigma}}$ (Eq. 8) with a truncated version $\mathbf{K}_{\mathbf{\Sigma}}$, where values beyond certain standard deviations are set to zero, and then convolve each feature vector with $\mathbf{K}_{\mathbf{\Sigma}}$. This idea is illustrated in Fig. 2. To achieve linear complexity with the number of events, we asynchronously convolve each feature vector using the event-based convolution method proposed by Scheerlinck *et al.* [19].

For simplicity, assuming the kernel $\mathbf{K}_{\mathbf{\Sigma}}$ has size κ^d, evaluating the approximate functions has a complexity of $O(N_e \kappa^d)$, which is linear with the number of events. The complexity may still be exponential with the number of dimensions, if we do not consider efficient high-dimensional convolution operations that reduce the computational complexity to become linear with the number of dimensions, *e.g.* [1]. Although we need to discretise the feature space, the computational complexity of the AEMin approach is independent of the actual number of discretised bins. In contrast, the complexity of the CMax [6] framework is also linear with the number of discretised bins (*e.g.* number of image pixels N_p). This means that although directly extending the CMax framework to handle higher dimensions is possible, in practise it would be inefficient and not scalable, without considering sophisticated data structures.

The *Approximate Potential* energy can be expressed as

$$\tilde{P}(\mathbf{f};\boldsymbol{\theta}) \coloneqq -\frac{1}{N_e^2} \sum_{k,l}^{N_e} \mathbf{K}_{\mathbf{\Sigma}} * \delta_{\mathbf{f}_k}(\mathbf{f}_l), \tag{15}$$

where each feature $\mathbf{f}_k$ is convolved with the truncated kernel $\mathbf{K}_{\mathbf{\Sigma}}$ in the feature space. Similarly, the *Sharma-Mittal*, *Rényi* and *Tsallis* entropies can be approximately expressed based on the kernel $\mathbf{K}_{\mathbf{\Sigma}}^{\alpha}$ and the *Shannon* entropy can be approximately expressed based on the kernel $\mathbf{K}_{\mathbf{\Sigma}} \odot \log \mathbf{K}_{\mathbf{\Sigma}}$:

$$\tilde{H}_{\alpha,\beta}(\mathbf{f};\boldsymbol{\theta}) \coloneqq \frac{1}{1-\beta} \left[\left(\frac{1}{N_e^2} \sum_{k,l}^{N_e} \mathbf{K}_{\mathbf{\Sigma}}^{\alpha} * \delta_{\mathbf{f}_k}(\mathbf{f}_l) \right)^{\gamma} - 1 \right], \tag{16}$$

$$\tilde{R}_{\alpha}(\mathbf{f};\boldsymbol{\theta}) \coloneqq \frac{1}{1-\alpha} \log \left(\frac{1}{N_e^2} \sum_{k,l}^{N_e} \mathbf{K}_{\mathbf{\Sigma}}^{\alpha} * \delta_{\mathbf{f}_k}(\mathbf{f}_l) \right), \tag{17}$$

$$\tilde{T}_{\alpha}(\mathbf{f};\boldsymbol{\theta}) \coloneqq \frac{1}{1-\alpha} \left(\frac{1}{N_e^2} \sum_{k,l}^{N_e} \mathbf{K}_{\mathbf{\Sigma}}^{\alpha} * \delta_{\mathbf{f}_k}(\mathbf{f}_l) - 1 \right), \tag{18}$$

$$\tilde{S}(\mathbf{f};\boldsymbol{\theta}) \coloneqq \frac{1}{N_e^2} \sum_{k,l}^{N_e} (\mathbf{K}_{\mathbf{\Sigma}} \odot \log \mathbf{K}_{\mathbf{\Sigma}}) * \delta_{\mathbf{f}_k}(\mathbf{f}_l), \tag{19}$$

where $\odot$ represents the Hadamard product and log represents the natural logarithm applied element-wise.

4 Experiments and Results

In this section, we test our framework to estimate several models by providing qualitative and quantitative assessments. In every experiment, we assume that the camera is calibrated and lens distortion has been removed. We also assume that each batch of events spans a short time interval, such that the model parameters can be considered constants (which is reasonable, since events can be triggered with a microsecond temporal resolution). For brevity, we will only consider the *Tsallis* entropy (11) and respective approximation (18) where $\alpha = 2$. Note that α and β are not tunable parameters since each one specifies an entropy.

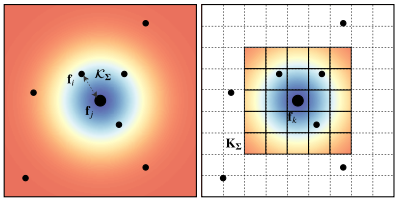

Fig. 2. In the EMin approach, all features pairwise distances are computed (left). In the AEMin approach, the feature space is discretised and each feature $\mathbf{f}_k$ is convolved with a truncated kernel $\mathbf{K}_{\boldsymbol{\Sigma}}$ (right)

The proposed AEMin requires that we discretise the feature space and perform convolution operations. We use bilinear interpolation/voting for each feature vector $\mathbf{f}_k$ to update the nearest bins, as suggested in [7]. We asynchronously convolve each feature vector, using the event-based asynchronous convolution method proposed by Scheerlinck *et al*. [19]. This convolution method was also used in the custom implementation of the CMax framework [6] and the decay factor was set to 0, to emulate a synchronous convolution over the entire space. We use a 3×3 Gaussian truncated kernel with 1 bin as standard deviation.

Our optimisation framework was implemented in C++ and we used the CG-FR algorithm of the scientific library GNU-GSL for the optimisation and the Auto-Diff module of the Eigen library for the (automatic) derivatives computation. [1] Additional models are provided in the supplementary material, as well as additional results (for more entropies), and practical considerations.

4.1 Motion Estimation in the Image Plane

We tested our framework using sequences from the dataset provided by Mueggler *et al*. [17]. The dataset consists of real sequences acquired by a DAVIS240 camera, each with approximately one minute duration and increasing motion magnitude. A motion-capture system provides the camera's pose at 200Hz and a built-in Inertial Measurement Unit (IMU) provides acceleration and angular velocity measurements at 1000Hz.

[1] Code publicly available: www.imperial.ac.uk/personal-robotics.

Optical Flow: The optical flow model has 2-DOF and can be parameterised by the 2D linear velocity on the image plane $\boldsymbol{\theta} = (u_x, u_y)^\mathsf{T}$, being expressed as

$$\mathbf{f}_k = \mathcal{M}(e_k; \boldsymbol{\theta}) = \mathbf{x}_k - \Delta t_k \boldsymbol{\theta}, \tag{20}$$

where $\mathbf{f}_k = \mathbf{x}'_k$ represents the image coordinates of the warped events, according to the notation in [6], and $\Delta t_k = t_k - t_{\text{ref}}$. For our framework, this is just a special case, where the feature vector $\mathbf{f}_k$ is 2-dimensional.

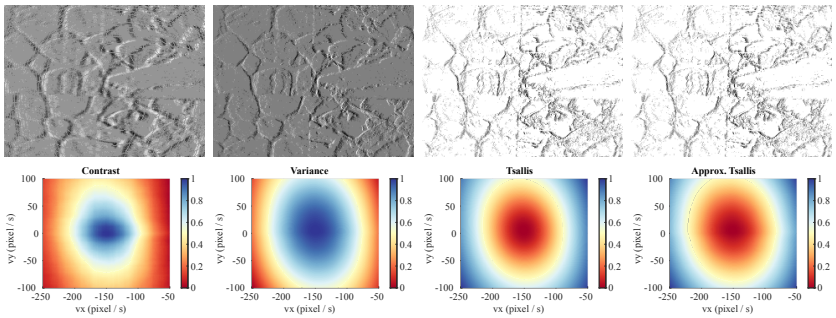

Fig. 3. Optical flow estimation between frames 17 and 18 of the `poster_translation` sequence [17]. Top row, from left to right: Original events projected onto the IWE, where the flow is dominated by the horizontal component $\left(\boldsymbol{\theta}^* \approx (-150, 0)^\mathsf{T}\right)$; Motion compensated events projected onto the IWE, according to the CMax framework [6] $\left(\boldsymbol{\theta}^* = (-150.3, 3.7)^\mathsf{T}\right)$, *Tsallis* entropy $\left(\boldsymbol{\theta}^* = (-150.8, 5.6)^\mathsf{T}\right)$ and corresponding approximate entropy $\left(\boldsymbol{\theta}^* = (-150.7, 3.8)^\mathsf{T}\right)$, respectively. Bottom row, from left to right: IWE contrast, Variance [6], *Tsallis* entropy and respective approximation profiles in function of the optical flow parameters, respectively

Fig. 3 shows the results of the optical flow estimation between frames 17 and 18 of the `poster_translation` sequence [17]. Similar results are obtained if we use the CMax framework [6]. Distinctively, however, by using an entropy measure, our framework minimises the modelled events' dispersion in a 2D feature space. The profiles of the *Tsallis* entropy and corresponding approximation in function of the optical flow parameters are also presented (two most right plots in the bottom row). We can see that both profiles are similar and the optical flow parameters are correctly estimated.

Rotational Motion: The rotational model has 3-DOF and can be parameterised by the 3D angular velocity $\boldsymbol{\theta} = (w_x, w_y, w_z)^\mathsf{T}$, being expressed as

$$\mathbf{f}_k = \mathcal{M}(e_k; \boldsymbol{\theta}) \propto \mathcal{R}^{-1}(t_k; \boldsymbol{\theta}) \begin{pmatrix} \mathbf{x}_k \\ 1 \end{pmatrix}, \tag{21}$$

where the rotation matrix $\mathcal{R} \in SO(3)$ can be written as a matrix exponential map of the angular velocity parameters $\boldsymbol{\theta}$ as

$$\mathcal{R}(t_k; \boldsymbol{\theta}) = \exp\left(\Delta t_k \hat{\boldsymbol{\theta}}\right), \tag{22}$$

where $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{3\times 3}$ is the associated skew-symmetric matrix.

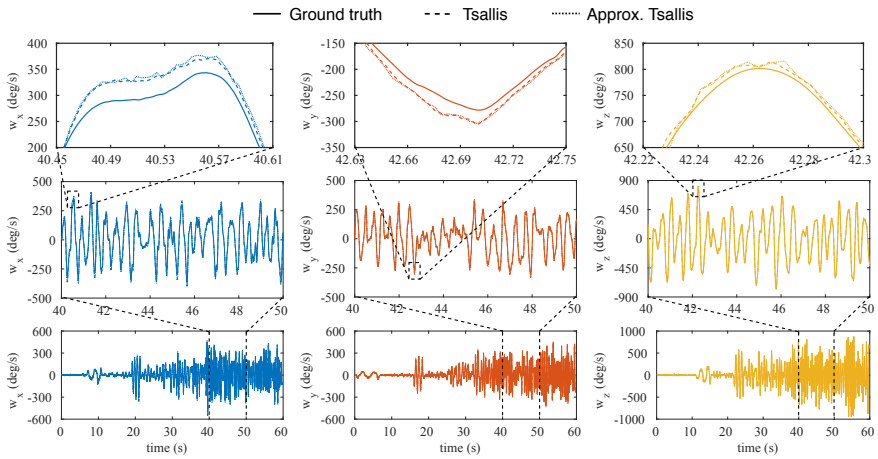

Fig. 4. Comparison of the angular velocity estimated against the ground truth from IMU measurements on the `poster_rotation` sequence [17], considering batches of 20000 events. Bottom row: Whole sequence. Middle and top rows: Zoomed-in plots of the corresponding bounded regions. Both angular velocity profiles are almost identical to the ground truth

Fig. 4 presents a comparison of the angular velocity parameters estimated by the proposed framework using the *Tsallis* entropy and the corresponding approximation against the ground truth provided by IMU measurements on the `poster_rotation` sequence [17]. Qualitatively, both angular velocity profiles are identical to the ground truth, even during peak excursions of approximately ± 940 deg/s, which correspond to rotations of more than 2.5 turns per second.

A more detailed quantitative performance comparison is presented in Table 1. We compare the *Tsallis* and corresponding approximate entropy functions in terms of errors for each angular velocity component and the respective standard deviation and RMS, considering batches of 20000 events, on the `boxes_rotation` and `poster_rotation` sequences [17]. The exact entropy function achieves the best overall results, while the proposed efficient approximate entropy achieves competitive performance.

Table 1. Accuracy comparison on the `boxes_rotation` and `poster_rotation` sequences [17]. The angular velocity errors for each component $(e_{w_x}, e_{w_y}, e_{w_z})$, their standard deviation (σ_{e_w}) and RMS are presented in deg/s, w.r.t. IMU measurements, considering batches of 20000 events. The RMS error compared to the maximum excursions are also presented in percentage (RMS %), as well as the absolute and relative maximum errors. The best value per column is highlighted in bold

Sequence	Function	e_{w_x}	e_{w_y}	e_{w_z}	σ_{e_w}	RMS	RMS %	max	max %
`boxes_rotation`	Variance (3) [6]	7.49	6.87	**7.53**	10.73	10.80	1.15	65.05	6.92
	Tsallis (11)	**7.43**	**6.53**	8.21	**9.75**	**9.91**	**1.06**	**45.76**	**4.87**
	Approx. Tsallis (18)	7.93	7.57	7.87	10.25	10.33	1.10	47.40	5.04
`poster_rotation`	Variance (3) [6]	**13.26**	8.73	8.73	14.60	14.65	1.56	**62.41**	**6.64**
	Tsallis (11)	13.40	**8.66**	**7.88**	**13.39**	**13.65**	**1.45**	67.34	7.16
	Approx. Tsallis (18)	14.31	9.16	8.34	14.16	14.22	1.51	64.68	6.88

Motion in Planar Scenes: Motion estimation in planar scenes can be achieved by using the homography model. This model has 8-DOF and can be parameterised by the 3D angular velocity $\mathbf{w} = (w_x, w_y, w_z)^\mathsf{T}$, the up to scale 3D linear velocity $\bar{\mathbf{v}} = \mathbf{v}/s = (\bar{v}_x, \bar{v}_y, \bar{v}_z)^\mathsf{T}$ and the normalised 3D normal vector $\mathbf{n} = (n_x, n_y, n_z)^\mathsf{T}$ of the inducing plane $\boldsymbol{\pi} = (\mathbf{n}^\mathsf{T}, s)^\mathsf{T}$, being expressed as

$$\mathbf{f}_k = \mathcal{M}(e_k; \boldsymbol{\theta}) \propto \mathcal{H}^{-1}(t_k; \boldsymbol{\theta}) \begin{pmatrix} \mathbf{x}_k \\ 1 \end{pmatrix}, \quad (23)$$

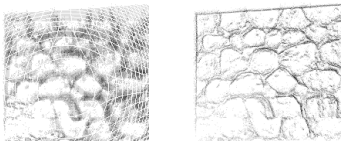

Fig. 5. Motion estimation in planar scenes. (Left) Original events projected. (Right) Modelled events projected onto the image plane

where $\boldsymbol{\theta} = (\mathbf{w}^\mathsf{T}, \bar{\mathbf{v}}^\mathsf{T}, \mathbf{n}^\mathsf{T})^\mathsf{T}$ are the model parameters and the homography matrix $\mathcal{H}$ can be written in function of the model parameters $\boldsymbol{\theta}$ as

$$\mathcal{H}(t_k; \boldsymbol{\theta}) = \mathcal{R}(t_k; \mathbf{w}) - \Delta t_k \bar{\mathbf{v}} \mathbf{n}^\mathsf{T}. \quad (24)$$

Fig. 5 shows the results of motion estimation in a planar scene, according to the homography model, between frames 673 and 674 of the `poster_6dof` sequence [17]. The following parameters were obtained: $\mathbf{w} = (0.81, -0.11, -1.47)^\mathsf{T}$, $\bar{\mathbf{v}} = (-0.19, 0.12, 1.73)^\mathsf{T}$, $\mathbf{n} = (-0.10, -0.14, -0.99)^\mathsf{T}$.

4.2 Motion Estimation in 3D Space

As proof of concept, we synthetically generated events from the corners of a moving cube in 3D space and used our framework to estimate its 6-DOF motion parameters, without projecting the events onto the image plane. We also tested our framework on sequences from the dataset provided by Zhu *et al.* [27]. The dataset consists of real sequences acquired by a set of different sensors for event-based 3D perception. The set of sensors includes an event-based stereo system of two DAVIS346B cameras and a frame-based stereo system of two Aptina MT9V034 cameras at 20fps, which provides the depth of the events generated.

6-DOF: The 6-DOF can be parameterised by the 3D angular velocity $\mathbf{w} = (w_x, w_y, w_z)^\mathsf{T}$ and the 3D linear velocity $\mathbf{v} = (v_x, v_y, v_z)^\mathsf{T}$, being expressed as

$$\mathbf{f}_k = \mathcal{M}(e_k; \boldsymbol{\theta}) = \mathcal{S}^{-1}(t_k; \boldsymbol{\theta})\mathbf{z}_k, \tag{25}$$

where $\mathbf{z}_k$ and $\mathbf{f}_k$ represent the 3D coordinates of the original and modelled events, respectively. The matrix exponential map $\mathcal{S} \in SE(3)$ encodes the rigid body transformation and can be written in function of $\boldsymbol{\theta} = (\mathbf{w}^\mathsf{T}, \mathbf{v}^\mathsf{T})^\mathsf{T}$ as

$$\mathcal{S}(t_k; \boldsymbol{\theta}) = \exp\left[\Delta t_k \begin{pmatrix} \hat{\mathbf{w}} & \mathbf{v} \\ \mathbf{0}^\mathsf{T} & 0 \end{pmatrix}\right], \tag{26}$$

where $\hat{\mathbf{w}} \in \mathbb{R}^{3\times3}$ is the associated skew-symmetric matrix.

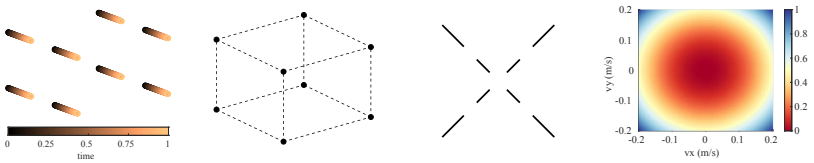

Fig. 6. 6-DOF estimation. From left to right: Events generated from the corners of a synthetic cube moving with $v_z = -1$ (m/s) in 3D space. Modelled events according to the optimal parameters $\mathbf{w} = (0, 0, 0)^\mathsf{T}$, $\mathbf{v} = (0, 0, -1)^\mathsf{T}$ (dashed lines included for better visualisation only). Original events projected onto the image plane. Projected *Tsallis* entropy profile at $\mathbf{w} = (0, 0, 0)^\mathsf{T}$, $v_z = -1$, in function of v_x and v_y

Fig. 6 illustrates a cube moving along the z-axis with $v_z = -1$ (m/s), whose corners generate events in 3D space. Our framework can accurately estimate the parameters of the moving cube, by modelling the events' trajectory according to the 6-DOF model. The CMax framework [6] can not estimate these parameters directly in 3D, because it requires the events to be projected onto the image plane. We also present the projected entropy profile in function of the velocities in the x and y direction, at $\mathbf{w} = (0, 0, 0)^\mathsf{T}$ and $v_z = -1$ (m/s). The profile exhibits a minimum at $\mathbf{v} = (0, 0, -1)^\mathsf{T}$, corresponding to the motion parameters.

In Fig. 7, we compare the original and modelled events from a moving 3D scene of the `indoor_flying1` sequence [27] (for illustration purposes, we consider the first 75000 events at the 24 second timestamp). We can retrieve the 6-DOF parameters by minimising the modelled augmented events' dispersion directly in 3D space, while also aligning the events to a reference frame.

We also present a quantitative evaluation in Table 2 on three sequences from the dataset provided by Zhu *et al.* [27]. Both the exact and approximate entropies achieve similar errors. The 6-DOF parameters were estimated from highly corrupted data since the events' 3D coordinates were obtained from depth measurements at 20fps. Moreover, in the `outdoor_driving_night1` sequence, the events generated are significantly influenced by several other relative motions,

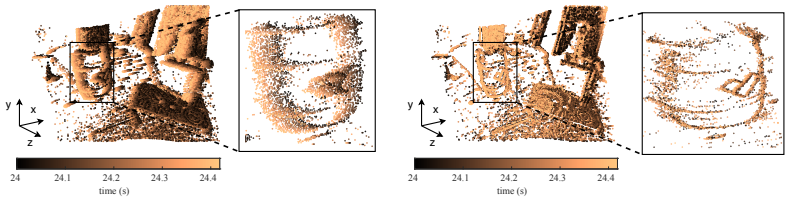

Fig. 7. 6-DOF estimation on the `indoor_flying1` sequence [27]. The trajectory of the original augmented events (left) can be modelled in 3D to retrieve the 6-DOF parameters of the moving camera, while also aligning the events (right)

Table 2. Accuracy comparison on the `indoor_flying1` and `indoor_flying4` and `outdoor_driving_night1` sequences [27]. The average angular and linear velocity errors (e_w, e_v), their standard deviation $(\sigma_{e_w}, \sigma_{e_v})$ and RMS (RMS_{e_w}, RMS_{e_v}) are presented in deg/s and m/s, respectively. The RMS errors compared to the maximum excursions are also presented in percentage (RMS_{e_w} %, RMS_{e_v} %)

Sequence	**Function**	e_w	e_v	σ_{e_w}	σ_{e_v}	RMS_{e_w}	RMS_{e_v}	RMS_{e_w} %	RMS_{e_v} %
`indoor_flying1`	*Tsallis* (11)	2.22	0.13	2.87	0.16	3.03	0.17	7.47	19.88
	Approx. Tsallis (18)	2.08	0.10	2.40	0.12	2.70	0.12	6.66	14.80
`indoor_flying4`	*Tsallis* (11)	4.08	0.25	5.16	0.30	5.12	0.30	20.53	16.38
	Approx. Tsallis (18)	4.43	0.23	5.30	0.29	5.56	0.30	22.28	16.35
`outdoor_driving_night1`	*Tsallis* (11)	4.18	1.73	14.73	1.71	14.85	2.22	3.98	21.85
	Approx. Tsallis (18)	7.27	1.82	17.97	1.94	17.97	2.41	4.81	23.71

e.g. due to other cars moving in the field-of-view. Our framework is capable of coping with noise, provided it is not the predominant factor.

In Table 3, we compare our framework to a deep learning method [28] that predicts the egomotion of a moving camera from events, in terms of relative pose error ($\text{RPE} = \arccos \frac{t_{\text{pred}} \cdot t_{\text{gt}}}{\|t_{\text{pred}}\| \cdot \|t_{\text{gt}}\|}$) and relative rotation error ($\text{RRE} = \left\| \text{logm}\left(\mathbf{R}_{\text{pred}}^{\mathcal{T}} \mathbf{R}_{\text{gt}}\right) \right\|$, where logm is the matrix log). Our framework compares favourably possibly because the deep learning method also estimates the depth.

Table 3. Quantitative evaluation of our framework compared to Zhu *et al.* [28] on the `outdoor_driving_day1` sequence [27]

	ARPE (deg)	ARRE (rad)
Approx. Tsallis (18)	**4.44**	**0.00768**
Zhu *et al.* [28]	7.74	0.00867

4.3 Discussion and Limitations

We have demonstrated that the proposed EMin framework can be used to tackle several common computer vision estimation problems. The EMin approach can achieve better performance in rotational motion estimation. The proposed AEMin approach can still achieve competitive performance, whilst being computationally more efficient. Nevertheless, we consider that the frame-

work's capability of handling models whose outputs have arbitrary dimensions is its most relevant property. As proof of concept, we showed that our framework can estimate the 6-DOF of a moving cube in 3D space. The quantitative tests on three sequences support the potential of the proposed framework to estimate the parameters of models with arbitrary output dimensions.

Our framework estimates the model parameters by minimising an entropy measure of the resultant modelled events. In the limit, the entropy is minimised if all events are mapped onto the same point, which in practise can occur by trying to estimate up-to scale 3D linear velocities (*e.g.* homography model). Fig. 8 exemplifies this situation, where we show the original events and the modelled events according to the estimated parameters, in the first and second rows, respectively. The CMax framework [6] exhibits a similar limitation since the contrast of the IWE is also maximised if all events are warped onto a line.

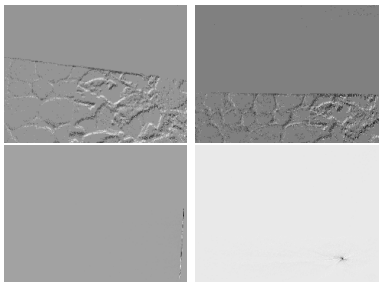

Fig. 8. Estimation failure cases. Original events (top row) and respective modelled events according to the best estimated parameters (bottom row), using (left) the variance [6] and (right) the *Approx. Tsallis* entropy, Eq. (18)

5 Conclusion

We have proposed a framework for event-based model estimation that can handle arbitrary dimensions. Our approach takes advantage of the benefits of the event-based cameras while allowing to incorporate additional sensory information, by augmenting the events, under the same framework. Additionally, since it can handle features of arbitrary dimensions, it can be readily applied to estimation problems that have output features of 4 or more dimensions, which we leave for future work. Thus, the proposed framework can be seen as an extension of previous motion compensation approaches. The exact EMin approach achieves the best performance, although its complexity is quadratic with the number of events. This motivated the proposed AEMin approach that achieves competitive performance while being computationally more efficient.

Acknowledgements: Urbano Miguel Nunes was supported by the Portuguese Foundation for Science and Technology under Doctoral Grant with reference SFRH/BD/130732/2017. Yiannis Demiris is supported by a Royal Academy of Engineering Chair in Emerging Technologies. This research was supported in part by EPRSC Grant EP/S032398/1. The authors thank the reviewers for their insightful feedback, and the members of the Personal Robotics Laboratory at Imperial College London for their support.

References

1. Adams, A., Baek, J., Davis, M.A.: Fast high-dimensional filtering using the permutohedral lattice. Computer Graphics Forum **29**(2), 753–762 (2010) 7
2. Andreopoulos, A., Kashyap, H.J., Nayak, T.K., Amir, A., Flickner, M.D.: A low power, high throughput, fully event-based stereo system. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7532–7542 (2018) 1
3. Bardow, P., Davison, A.J., Leutenegger, S.: Simultaneous optical flow and intensity estimation from an event camera. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 884–892 (2016) 1
4. Gallego, G., Gehrig, M., Scaramuzza, D.: Focus is all you need: Loss functions for event-based vision. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 12280–12289 (2019) 3, 6
5. Gallego, G., Lund, J.E., Mueggler, E., Rebecq, H., Delbruck, T., Scaramuzza, D.: Event-based, 6-dof camera tracking from photometric depth maps. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(10), 2402–2412 (2018) 1
6. Gallego, G., Rebecq, H., Scaramuzza, D.: A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3867–3876 (2018) 1, 2, 3, 4, 6, 7, 8, 9, 11, 12, 14
7. Gallego, G., Scaramuzza, D.: Accurate angular velocity estimation with an event camera. IEEE Robotics and Automation Letters **2**(2), 632–639 (2017) 1, 3, 4, 8
8. Glover, A., Bartolozzi, C.: Robust visual tracking with a freely-moving event camera. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3769–3776 (2017) 1
9. Glover, A., Vasco, V., Bartolozzi, C.: A controlled-delay event camera framework for on-line robotics. In: IEEE International Conference on Robotics and Automation. pp. 2178–2183 (2018) 1
10. Haessig, G., Berthelon, X., Ieng, S.H., Benosman, R.: A spiking neural network model of depth from defocus for event-based neuromorphic vision. Scientific Reports **9**(1), 3744 (2019) 1
11. Kim, H., Leutenegger, S., Davison, A.J.: Real-time 3d reconstruction and 6-dof tracking with an event camera. In: European Conference on Computer Vision. pp. 349–364 (2016) 1
12. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems. pp. 109–117 (2011) 4
13. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning. pp. 282–289 (2001) 4
14. Manderscheid, J., Sironi, A., Bourdis, N., Migliore, D., Lepetit, V.: Speed invariant time surface for learning to detect corner points with event-based cameras. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 10245–10254 (2019) 1
15. Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., Von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J., Bernardino, A.: The icub humanoid robot: An open-systems platform for research in cognitive development. Neural Networks **23**(8-9), 1125–1134 (2010) 2

16. Mitrokhin, A., Fermüller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1–9 (2018) 1, 3
17. Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D.: The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. International Journal of Robotics Research **36**(2), 142–149 (2017) 8, 9, 10, 11
18. Rebecq, H., Gallego, G., Mueggler, E., Scaramuzza, D.: Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. International Journal of Computer Vision **126**(12), 1394–1414 (2018) 1, 3
19. Scheerlinck, C., Barnes, N., Mahony, R.: Asynchronous spatial image convolutions for event cameras. IEEE Robotics and Automation Letters **4**(2), 816–822 (2019) 7, 8
20. Sharma, B.D., Mittal, D.P.: New non-additive measures of inaccuracy. Journal of Mathematical Science **10**, 120–133 (1975) 5
21. Valeiras, D.R., Lagorce, X., Clady, X., Bartolozzi, C., Ieng, S.H., Benosman, R.: An asynchronous neuromorphic event-driven visual part-based shape tracking. IEEE transactions on neural networks and learning systems **26**(12), 3045–3059 (2015) 1
22. Vidal, A.R., Rebecq, H., Horstschaefer, T., Scaramuzza, D.: Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios. IEEE Robotics and Automation Letters **3**(2), 994–1001 (2018) 1
23. Weikersdorfer, D., Adrian, D.B., Cremers, D., Conradt, J.: Event-based 3d slam with a depth-augmented dynamic vision sensor. In: IEEE International Conference on Robotics and Automation. pp. 359–364 (2014) 2
24. Xie, Z., Chen, S., Orchard, G.: Event-based stereo depth estimation using belief propagation. Frontiers in Neuroscience **11**, 535 (2017) 1
25. Zhu, A.Z., Atanasov, N., Daniilidis, K.: Event-based feature tracking with probabilistic data association. In: IEEE International Conference on Robotics and Automation. pp. 4465–4470 (2017) 4
26. Zhu, A.Z., Chen, Y., Daniilidis, K.: Realtime time synchronized event-based stereo. In: European Conference on Computer Vision. pp. 433–447 (2018) 1, 3
27. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018) 11, 12, 13
28. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019) 1, 13