Benchmarking Microsaccade Recognition with Event Cameras: A Novel Dataset and Evaluation

Waseem Shariff*[1]
waseem.shariff@universityofgalway.ie
Timothy Hanley[2]
t.hanley4@universityofgalway.ie
Maciej Stec[1]
m.stec1@universityofgalway.ie
Hossein Javidnia[3]
hossein.javidnia@tcd.ie
Peter Corcoran[1]
peter.corcoran@universityofgalway.ie

[1] C3I Imaging Lab, School of Engineering, University of Galway, Ireland
[2] School of Computer Science, University of Galway, Ireland
[3] Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland

Abstract

Microsaccades are small, involuntary eye movements essential for visual perception and neural processing. Traditional microsaccade research often relies on eye trackers and frame-based video analysis. While eye trackers offer high precision, they can be expensive and have limitations in scalability and temporal resolution compared to alternative methods. In contrast, event-based sensing offers a more efficient and precise alternative, as it captures high-resolution spatial and temporal information with minimal latency. This work introduces a pioneering event-based microsaccade dataset, designed to support efforts in studying small eye movement data within cognitive computing. Using Blender, we generate high-fidelity renderings of eye movement scenarios and simulate microsaccades with angular displacements ranging from 0.5° to 2.0°, divided into seven distinct classes. These sequences are simulated into event streams using v2e, preserving the temporal dynamics of real microsaccades. The event streams, with durations as small as 0.25 milliseconds and as large as 2.25 milliseconds, have such high temporal resolution, which prompted us to investigate whether SNNs could effectively detect and classify these movements. We evaluate the proposed dataset using Spiking-VGG11, Spiking-VGG13, and Spiking-VGG16, and further introduce Spiking-VGG16Flow - an optical flow - enhanced variant-implemented with SpikingJelly. Across experiments, these models achieve an average accuracy of approximately 90%, effectively classifying microsaccades based on angular displacement, irrespective of event count or duration. These results highlight the suitability of SNNs for fine motion classification and establish a benchmark for future research in event-based vision. To facilitate further work in neuromorphic computing and visual neuroscience, both the dataset and trained models will be released publicly. The dataset and code will be made available here- https://waseemshariff126.github.io/microsaccades/

© 2025. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

1 Introduction

Microsaccades are small, involuntary eye movements typically under 1 to 2 in amplitude, distinct from larger, voluntary saccades due to their subtlety and small movements. While they share some kinematic properties with larger saccades, such as being conjugate in nature [7] and following a similar amplitude-velocity relationship [5], microsaccades are much finer in movement and serve a different function. Predominantly binocular, they often exhibit a horizontal directional bias [13]. These brief, involuntary movements play a crucial role in maintaining visual clarity during fixation, subtly shifting the retinal image to prevent fading, a phenomenon that can occur when a stimulus remains stationary on the retina [15]. Unlike larger saccades, which are voluntary and are used for noticeable shifts in visual attention, microsaccades help preserve perceptual stability and visual sensitivity without any obvious change in gaze direction. The changes are very small, yet they are essential for preserving the clarity of our visual experience.

Microsaccades are attracting increased attention due to their sensitivity to underlying cognitive processes. Variations in their frequency, direction, and amplitude have been linked to attentional shifts, mental workload, and other internal neural states [15]. As brief, involuntary movements, they offer a noninvasive means of probing attention and cognitive function, making them relevant to both basic neuroscience and applied computational modeling [7, 15, 18]. They are also being explored as potential indicators for neurological and psychiatric conditions such as Parkinson's disease, schizophrenia, and attention deficit disorders [2, 13, 33]. Despite this promise, microsaccades are difficult to analyze due to their small magnitude and transient nature. Current detection methods rely heavily on high-resolution eye trackers that produce numerical gaze coordinates, limiting the use of image- or event-based techniques. The absence of visual or spatiotemporal datasets tailored specifically for microsaccade detection further hinders the development of advanced models. In particular, it remains unclear whether artificial neural networks, especially those designed for vision tasks, can reliably detect and interpret such subtle motion signals amid noise and sparse data.

Despite the promise of microsaccades as indicators of cognitive and neurological states, their detection remains technically challenging [3]. Traditional frame-based cameras, such as standard RGB sensors, lack the temporal resolution needed to reliably capture these rapid, subtle eye movements, which occur on the order of milliseconds [16]. Achieving sufficiently high frame rates (e.g., 1000 Hz) demands expensive hardware, large memory capacity, and generates vast amounts of mostly redundant data, complicating efficient analysis [16]. To address these limitations, event-based cameras also known as dynamic vision sensors (DVS) have emerged as a powerful alternative. Unlike conventional cameras that capture full image frames at fixed intervals, event cameras asynchronously record changes in brightness with microsecond-level precision and operate at effective sampling frequencies of around 10 kHz [1, 12, 29]. This approach offers exceptionally high temporal resolution and data efficiency, making event cameras particularly well-suited for detecting rapid, subtle movements such as microsaccades. Previous studies have demonstrated the effectiveness of event cameras for high-speed eye tracking [1, 20, 35], further motivating their use in this domain.

Many event-based studies indicate that Spiking Neural Networks (SNNs) are particularly well-suited for processing event camera data due to their biological plausibility and dynamic neural behavior [6, 12, 29].However, while these models have demonstrated success in capturing spatiotemporal dynamics, it remains unclear whether they can reliably detect the small, brief eye movements characteristic of microsaccades, which typically last between 0.5

and 2.5 milliseconds. To investigate this, we frame the classification of eye motion angles based on angular bins, not to replicate biological function, but to evaluate whether neural networks can recognize such subtle motion patterns. This proof-of-concept task lays the groundwork for future, more complex objectives like trajectory reconstruction or cognitive state inference.

In this work, we introduce the first synthetic microsaccades dataset created using Blender. It comprises 175,000 annotated sequenced samples containing event streams for left-eye and right-eye microsaccades, over seven amplitudes ranging from 0.5° to 2°. This data set provides a controlled environment for systematic evaluation and supports reproducibility in neuromorphic vision.

We benchmarked the dataset using Spiking Neural Networks (SNNs), focusing on VGG-based models for their strong performance in visual tasks and compatibility with event-driven data [22]. By leveraging the publicly available repository [22], we augmented the data and integrated event data with spiking models, building upon existing frameworks and improving the representation of the rapidly changing microsaccade signals. Additionally, we introduced Spiking-VGG16Flow, a variant of Spiking-VGG16 enhanced with optical flow estimation to improve sensitivity to subtle microsaccadic movements. This allowed a direct comparison against standard Spiking VGG models. Our classification results provide a solid baseline, demonstrating both the promise and the challenges of using SNNs for microsaccade detection.

Contributions

- We present the first synthetic event-based microsaccade dataset generated using Blender, featuring 175,000 annotated event streams that simulate realistic horizontal eye movements.
- We propose the Spiking-VGG16Flow variant, an optical-flow-augmented adaptation of Spiking-VGG16, specifically designed to investigate whether microsaccade classification models rely on event count or true motion cues.
- We demonstrate that models trained on synthetic data can generalize to real event-based recordings (EV-Eye), indicating the potential of simulation-driven learning in data-scarce domains.

2 Background

2.1 Conventional Approaches to Microsaccade Detection and Analysis

Microsaccades are critical for maintaining visual clarity during periods of visual fixation. They occur involuntarily and share kinematic similarities with larger, voluntary saccades, such as their conjugate nature and the amplitude-velocity relationship [5, 7, 18]. While microsaccades are often binocular, monocular events are typically excluded from analysis due to their perceived unreliability [24]. This approach is evident in studies such as Engbert and Mergenthaler (2006) and Troncoso et al. (2008), which recorded from both eyes but retained only binocular microsaccades in their analysis [5, 32]. This suggests that even if we record from just one eye, microsaccades can still be detected and classified, as long as they are part of a binocular event.

Despite their similarities, microsaccades differ signi cantly from larger saccades in terms of volitional control. Whereas saccades are often voluntary and goal-directed, microsaccad are considered involuntary. One of the strongest indicators of this involuntary nature is tha microsaccadic onset times follow an exponential distribution, suggesting stochastic gener tion [30]. The operational de nition of a microsaccade also varies slightly across studies though most adopt a hard upper bound of 1.2^{0} for amplitude [13]. Historically, reported sizes ranged from as small as 0.1^{0} [33] and 15^{0} [2] to over 60^{0} [4] in later studies. In this paper, we de ne microsaccades as having an amplitude of up to 2 degrees.

These ndings were largely enabled by the use of high-frequency EyeLink eye tracking systems [31]. Engbert and Kliegl (2003) relied on a system with a 250 Hz sampling rate [8], while later studies, such as Laubrock et al. (2005) and Sinn and Engbert (2016), use the EyeLink II system at 500 Hz [19, 30]. More recent experiments by Krejtz et al. (2020) and Schneider et al. (2021) employed the EyeLink 1000 system, which offers a 1000 H sampling rate [18, 27]. These tools, along with velocity-based detection algorithms like the Engbert & Kliegl method, have become the de facto standard for microsaccade analysis.

Efforts to move beyond hardware-based tracking are limited. For example, [3] used a CNN to segment the iris in 96 FPS video and detect microsaccades by thresholding ey velocity. However, microsaccades occur within a 2 ms window, requiring at least 500 FPS, s their approach lacks suf cient temporal precision. Computational methods for microsaccad detection are emerging but remain underexplored, and the lack of publicly available datase hampers progress. This gap led to the creation of our own synthetic dataset using Blend for controlled microsaccadic simulation.

2.2 Event-based Microsaccade Detection

While numerous studies [1, 17, 20, 21, 25, 26, 35, 35] have explored eye detection an tracking using event-based cameras, microsaccade detection in this domain remains larg unexplored. To our knowledge, there has been no established work that systematically inve tigates microsaccades using event-cameras. Most recently, Iddrisu et al. (2024) underscor this research gap in their review, noting the lack of prior work on event-driven microsaccad detection within the context of event-based eye analysis [16].

3 Dataset and Preprocessing

This section describes the creation of a synthetic microsaccade dataset using Blender-rend RGB frames, followed by the conversion of these frames into event-based streams for m crosaccade classi cation. The dataset is designed to provide controlled variations in motio and timing for model training and evaluation.

3.1 Blender RGB Frame Generation

To generate microsaccade data, we use Blender to render the RGB frames that simula horizontal eye movements. Each microsaccade instance is modeled as a brief horizon shift, where the eye moves in a random direction and then returns to its original position mimicking the rapid and transient microsaccade movements typically observed during visua xation. Movement is further de ned in terms of angular displacement, and we established

seven distinct classes for microsaccades, each corresponding to a speci c angular rang
detailed in Table 1.

The left and right eyes exhibit symmetric motion patterns, so we rst generate the s
quences for the left eye. Afterward, these sequences are ipped to simulate the right e
This approach captures the symmetrical nature of microsaccadic movements while maint
ing consistency across both eyes. The motion of the eye is represented by a boomerang
pattern, where the eye moves in one direction (e.g., leftward) and then returns to its orig
position (e.g., rightward), simulating the back-and-forth nature of a microsaccade.

In Blender, we simulate eye movements by rotating the virtual eyeball while keepi
the camera xed relative to the head. The eyeball's rotation angles are carefully adjuste
replicate realistic microsaccadic velocities, producing motion consistent with natural x
tional eye movements. The total number of frames generated for each class is proporti
to the angular displacement, as shown in Table 1. The class-speci c frame counts are
signed to accurately represent the motion duration for each class. To isolate amplitude a
sole distinguishing factor, we kept scene layout, lighting, textures, and background const
enabling focused analysis of event responses to ne motion variations. A visualization
a 1.0 microsaccade sequence is shown in Figure 1, where the rst and third rows disp
the Blender-rendered frames cropped around the eye region, and the second and fourth
show the corresponding event streams. Additional details of the Blender-based microsac
data generation process are provided in the supplementary material.

3.2 Event Simulation from RGB to Event

To simulate event-based vision sensors, we convert the rendered RGB frames into e
streams using an event simulator, speci cally v2e [14]. This conversion process mi
ics the asynchronous output of a Dynamic Vision Sensor (DVS), generating a series
events with timestamps, spatial coordinates (x, y), and polarity values (increasing light in-
tensity/decreasing light intensity). v2e has proven to be an effective tool for generat
synthetic event-based data and is increasingly utilized in the evaluation of neuromorp
vision models. Numerous studies have employed v2e-generated datasets to demonstra
bust model performance and generalization across various tasks such as driver distra
[28, 34], Driver monitoring system [26], MVSEC dataset [14] and low-light image enhanc
ment (LIE) dataset [23]

The time intervals for event generation are aligned with the time taken for each cla
as presented in Table 1. These intervals re ect the motion duration for each class, rang
from 0.25 ms for the smallest microsaccades (0.5°) to 2.25 ms for the largest microsacca
(2.0°). The duration is important, as it ensures that the event stream properly simulates
motion dynamics associated with different angular displacements, while overlapping th
durations helps prevent the model from learning solely on the basis of timing cues, fo
ing it to focus on the underlying motion dynamics. To further enhance the simulation a
mitigate bias from varying event counts across different classes, we also employ an ev
count overlapping strategy. Since event count is in uenced by the magnitude of the moti
larger microsaccades typically produce more events. To address this, we randomly res
ple each event sequence after event simulation to introduce overlapping event counts ac
angular classes (more information in supplementary material). This ensures that each c
cannot be solely distinguished by its event count while maintaining temporal consiste
and producing an overall more balanced dataset.

In total, each microsaccade class generates ve distinct event sequences by rando

sampling durations within the speci ed time interval range for that class. These event se quences are resampled a further ve times, as described in the supplementary material, e panding the dataset and ensuring that the model learns to classify based on motion dynam rather than duration or event count. The resampling process further re nes the dataset, r sulting in a more robust set of event samples for model training.

Figure 1: Visualization of a 10 microsaccade sequence, showing Blender-rendered frames (top and bottom rows) and corresponding event streams (middle and fourth rows). A gri is overlaid for clarity, with red and green pixels representing positive and negative polarity events, respectively. All rows should be read from left to right to observe the tempora evolution of the microsaccade.

Table 2 summarizes the key properties of our synthetic microsaccade dataset[1]. The original event frame resolution is 800600 pixels, from which we extract a centered ROI of 440 300 pixels to focus on the eye region and reduce background noise. The dataset co tains a total of 175,000 sequences, evenly split between the left and right eyes (87,500 eac We de ne seven horizontal microsaccade classes, with an equal number of samples per cla to ensure balanced training and evaluation.

4 Spiking VGG Architectures and Motion-Regularized VGGFlow

We explore the problem of microsaccade classi cation using spiking convolutional architec tures, beginning with spiking variants of VGG11, VGG13, and VGG16, adapted from [22]. These models incorporate leaky integrate-and- re (LIF) neurons and surrogate gradient trai ing, providing strong baselines for classifying microsaccades of varying magnitudes in

[1] https://waseemshariff126.github.io/microsaccades/

Class (°)	Frames	Angle (°)	Time (ms)
0.5	7	0.000–0.625	0.25–0.75
0.75	9	0.625–0.875	0.50–1.00
1.0	11	0.875–1.125	0.75–1.25
1.25	13	1.125–1.375	1.00–1.50
1.5	15	1.375–1.625	1.25–1.75
1.75	17	1.625–1.875	1.50–2.00
2.0	19	1.875–2.125	1.75–2.25

Table 1: Microsaccade classes with corresponding frame counts, angle ranges, and temporal resolution.

Property	Value
Original resolution	800 600
ROI resolution	440 300 (center)
Left eye sequences	87,500
Right eye sequences	87,500
Total sequences	175,000
Number of classes	7
Class distribution	Evenly split

Table 2: Synthetic microsaccade dataset details.

neuromorphic setting.

During early experimentation, we observed a strong performance across classes, b remained unclear whether the spiking models were truly learning the spatiotemporal namics of motion or simply leveraging differences in event count. Larger microsaccad naturally generate more events due to increased pixel activation, and we were concerned this correlation might inadvertently drive classi cation accuracy. To investigate this hypot esis and ensure that the model's decision-making was not based on such statistical bi we introduced an auxiliary motion-prediction objective into the spiking architecture.

Speci cally, we propose *spiking-VGGFlow*, a motion-regularized version of spiking VGG16 that incorporates an optical ow prediction branch during training. The model shar its convolutional backbone with spiking VGG16, but after the nal convolutional block, i bifurcates into two heads: (1) a classi cation head composed of fully connected spiking la ers trained with standard cross-entropy loss, and (2) a ow-prediction head that regres dense optical ow elds between adjacent binned event frames.

This additional supervision encourages the network to learn motion-aware represe tions rather than over tting to spatial event count. To compute ow targets, we rst divid the event stream into temporally binned frames (e.g., 10 bins per 3 ms), then apply Farneback algorithm [11] on the polarity-summed frames to produce dense ow maps. E target ow $F_t(x;y) = (u_t(x;y); v_t(x;y))$ represents horizontal and vertical displacements be tween frame t and $t+1$. The ow prediction head minimizes the mean squared error betwee predicted and ground truth ow vectors:

$$L_{ow} = \frac{1}{HW} \sum_{x=1}^{W} \sum_{y=1}^{H} (\hat{u}(x;y) \quad u(x;y))^2 + (\hat{v}(x;y) \quad v(x;y))^2 :$$

The overall training objective combines classi cation and ow supervision:

$$L_{total} = L_{class} + l \quad L_{ow} ;$$

where $l = 0:5$ balances the in uence of motion consistency.

Importantly, the ow branch is only active during training. At inference time, the mode reverts to the base spiking classi er without any added computational cost. As shown in S tion 5, this motion-supervised training improves generalization and robustness, sugges that the ow objective acts as an effective regularizer guiding the model toward more mea ingful spatiotemporal representations and reducing reliance on event-count correlations.