International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2024)
S. Hasegawa, N. Sakata and V. Sundstedt (Editors)

Learning-based Event-based Human Gaze Tracking with Blink Detection

Mao Kanno[1] and Mariko Isogawa[1]

[1]Graduate School of Science and Technology, Department of Open Environmental Science, Major in Information Engineering, Keio University, Japan

Abstract

This paper proposes an eye-tracking system using a CNN-LSTM network that utilizes only event data. This method holds potential for future applications in a wide range of fields, including AR/VR headsets, healthcare, and sports. Compared to traditional frame-based camera methods, our proposed approach achieves high FPS and low power consumption by utilizing event cameras. To improve the estimation accuracy, our gaze estimation system incorporates a blink detection, which was absent in existing systems. Our results shows that our method achieves better performance compared to existing studies.

1. Introduction

Eye tracking is a task that estimates where a person is looking at. Since eye tracking can be applied in a wide range of application such as AR/VR [AMF23, CKK19], medical field [NLB*10], and sports analysis [WWHA21, KNM18], it has been extensively studied. Many of the gaze estimation methods proposed so far use images or video captured by standard RGB cameras as input. However, RGB-based methods have several limitations. They are prone to failure in low-light conditions or when occlusion occurs. Additionally, even when using high-frame-rate cameras, the frame rate is typically limited to around 1000 fps. Consequently, the applicability of these methods to the aforementioned applications is restricted.

One possible solution to address these challenges is the use of event-based cameras. Event-based cameras, inspired by the retina of biological organisms, are sensors that detect and output data only for changes in the brightness of the observed scene [GDO*19]. Using events captured by event-based cameras allows the methods to operate at a relatively high FPS while maintaining low power consumption and capability of low-light conditions.

While there have indeed been studies that use data captured by event-based cameras for eye tracking, previous research has also utilized data captured by frame-based cameras. As a result, these studies have not fully exploited the high frame rate, low power and capability of low-light conditions advantages of event-based cameras. For this reason, we used only the data that captured by the event-based cameras.

To perform eye-tracking using only sparse event data without relying on any frame-based image data, this study we leverage combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks which can surpass the traditional eye-tracking methods.In addition, we introduced blink detection into the system. Existing methods track human gaze even when the subject is blinking or not focusing on anything, which reduces estimation accuracy. Therefore, we implemented a system that skips the frames where the subject is blinking during both training and inference phases.

2. Related Work

2.1. Eye Tracking

Eye tracking technology has seen significant advancements and diverse applications in recent years. It is extensively used in fields such as AR/VR, medical applications, accident prevention, and sports [KAS*19, NLB*10, AKNW21, VTMB20, CCHK23]. The ability to accurately track eye movements provides valuable insights into cognitive processes, user behavior, and attention. Traditional eye tracking often face challenges in terms of robustness and accuracy under varying lighting conditions and head movements [BBD24, AT17]. Recent developments in eye tracking algorithms and the integration of advanced sensors, such as event cameras, have shown promise in overcoming these limitations [AMK*21, FGK*22]. By leveraging the high temporal resolution and low latency of event cameras, eye tracking systems can achieve more precise and reliable performance, particularly in dynamic and challenging environments.

2.2. Event-Based Camera

Event cameras, renowned for their high-speed sensing capabilities, have been widely applied in various domains. Notably, they have found applications in object tracking [MFPA18], visual odometry [RHGS17], human pose tracking [ZMZ*23], 3D reconstruc-

DOI: 10.2312/egve.20241367

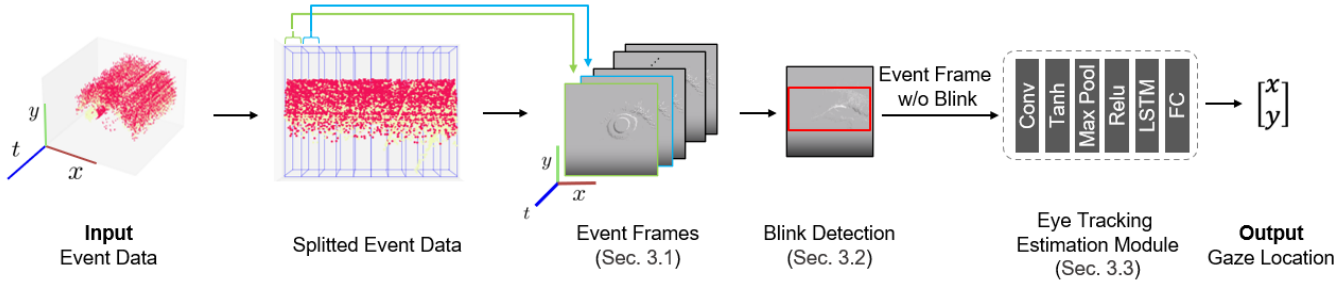

Figure 1: *Proposed framework. The input consists only of event data captured by an event-based camera. At the initial stage, event frames are created based on the number of events that occurred. Among these event frames, frames where blinks occur are detected and skipped during training and inference. The model is CNN-LSTM-based and outputs 2D coordinates.*

tion [BWCK20], SLAM [HZZT24], and hand tracking [DSN21]. These fields benefit greatly from the event camera's ability to capture dynamic scenes with exceptional temporal resolution and minimal latency. The characteristics of event-based cameras, such as their proficiency in handling high-speed motions and challenging lighting conditions, have paved the way for advancements in the robustness and accuracy of visual perception systems. These advantages underscore the growing adoption of event-based cameras in numerous fields, demonstrating their potential to surpass traditional frame-based cameras in specific scenarios.

3. Proposed Methods

Our system is built as illustrated in the Figure 1. First, we format the event data $E = [x_k, y_k, t_k, p_k]_{k=1}^{N}$ as event frames $F_t(y,x) = [F_1, F_2 ..., F_N]_{k=1}^{N}$ for input to the model. Then, we perform blink detection from these event frames and skip the frames where blinks occurred during both the training and inference phases, as this approach is expected to improve the accuracy of eye tracking.The frames that are not skipped are used as input to the CNN-LSTM model for training. The output of this model $a = [x'_k, y'_k]_{k=1}^{N}$ is the two-dimensional coordinates of a gaze point at a certain distance. The following subsections describe about the event frame that our model uses as input in Sec. 3.1, and blink detection in Sec. 3.2. Then, Sec. 3.3 explains out Eye Tracking Estimation Module.

3.1. Event Frame

The input to our network, called event frames, is created from the event data captured by the event-based camera. Event frames are created based on the number of events that occurred. If the events measured by the event camera exceed the set threshold, an event frame is created using the event data measured up to that point. Event frames make use of each pixel value according to the polarity of each event. To retain temporal information, they utilize the event's timestamp and polarity.

We adopt a method to effectively represent asynchronous event data as synchronous frame-like structures [ZYCD18, TWH*22].

We achieve this by using a this approach, where each event distributes its polarity p to the two closest event data points. Given N input events $\mathbf{E} = [x_k, y_k, t_k, p_k]_{k=1}^{N}$ and the temporal bin T, this approach first scales the timestamps to the range $[0, T-1]$, and then generates event frames V with dimensions $T \times H \times W$ as follows:

$$t_k^* = \frac{T-1}{t_N - t_1}(t_k - t_1)$$

$$V(t,y,x) = \sum_k p_k \max(0, 1 - |t - t_k^*|).$$

Furthermore, these outputs $V(t,y,x)$ are extracted one dimension at a time along the temporal axis and used as event frames $F_t(y,x)$:

$$F_t(y,x) = V(t,y,x), \quad t = 0, 1, \ldots, T-1$$

Thus, each pixel in these event frames F_t contains a temporally weighted and summed value. This method creates event frames that incorporate the temporal information of the events.

3.2. Blink Detection

There are several challenges in eye-tracking methods where estimation is performed even for frames where blinks occur. Since the gaze is not observed during a blink, the blink not only degrades the estimation accuracy but also unnecessarily increases power consumption and computational costs. Therefore, following the previous work that shows the effectiveness of incorporating blink detection into pupil tracking [ZSZ*24], we also introduce blink detection into our framework.

Unlike existing work that uses frame-based video supplemented by event data for blink detection [ZSZ*24], we propose a blink detection method using only event data. Our blink detection is inspired by the existing research [LIB20]. According to this research, it has been found that the number of events increases when a blink occurs. In our study, we developed a blink detection module by utilizing this characteristic. As shown in the red box in Figure 2, the area is designated near the center of the event frame. We determine whether a blink is occurring based on whether the number of events within this area exceeds a predetermined threshold. In frames where a blink is occurring, as shown in Figure 2, the

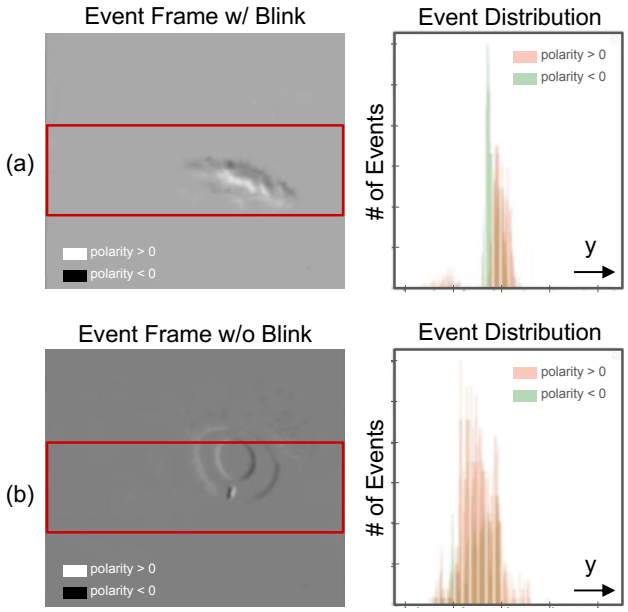

Figure 2: *Event frames and its event distributions (a) with or (b) without blink. In the event frame, white pixels represent event data with a polarity greater than 0, while black pixels represent event data with a polarity less than 0. Additionally, the adjacent graph shows the distribution of events along the y-axis. Orange indicates events with a polarity greater than 0, and green indicates events with a polarity less than 0. The differences in these distributions are used to detect whether a blink is present in a frame.*

events are concentrated within the red box in the event frame. Consequently, in the graph showing the distribution of events along the y-axis, events are concentrated near the center. Conversely, in frames where a blink is not occurring, as shown in Figure 2, the events do not fit within the red box and are dispersed across the entire frame. As a result, the graph also shows that events are not concentrated near the center. We detect blinks using these differences. In the created event frames, we set the aforementioned red box, and if the number of events within it exceeded the threshold, we considered it a blink. For event frames where this was not the case, we assumed no blink was occurring and proceeded with eye tracking.

3.3. Eye Tracking Estimation Module

We propose an Eye Tracking Estimation Module that takes event frames as input and outputs the 2D coordinates of the gaze direction. This module, illustrated in Figure 1 , consists of a 2D convolutional CNN, activation functions, pooling layers, an LSTM, and a fully connected layer.

First, we input the 2D frame $F_t(y,x) = [F_1, F_2 ..., F_N]_{k=1}^{N}$ into the 2D CNN, which processes it to generate feature maps. The resulting feature maps are then passed through a Tanh activation function to constrain their range between -1 and 1. Next, the output of the Tanh activation function is input to a max-pooling layer. The tensor obtained through pooling is then passed through a ReLU activation function and reshaped into a form suitable for input to the LSTM. In the LSTM layer, the hidden state and cell state are updated using the input frame.

The output of the LSTM is then flattened and fed into a fully connected layer. The final output $a = [x'_k, y'_k]_{k=1}^{N}$ is calculated through a linear transformation, producing the 2D coordinates of the gaze direction. Thus, the proposed module takes the event frame as input and outputs the 2D coordinates of the gaze direction. Our module effectively utilizes the spatio-temporal information of event data by performing spatial feature extraction through the CNN, temporal information integration through the LSTM, and final gaze direction estimation through the fully connected layer.

4. Experimental Settings

4.1. Datasets

We use the dataset from the previous research on eye tracking using event-based cameras [AMK*21]. This dataset was captured using an event-based camera attached near the eyes, recording at a resolution of 346×260 while tracking a dot displayed on a 1920×1080 resolution 40-inch monitor located 40 cm away.. The dataset consists of event data from both eyes of 27 participants, along with frame data recorded at a frame rate of 25fps. In our study, we used only the event data, creating event frames that are updated at 200FPS for use as data in this paper.

4.2. Baseline Methods

To investigate the effectiveness of our method, we compare our method against existing event-based eye-tracking method [AMK*21] and existing network architecture for the task of lip reading using event data [TWH*22]. The former is selected for comparison with existing work that addresses the same task as ours, while the latter is used for comparison against existing networks that only utilize events. In the previous method [AMK*21], blink detection is also performed; the results of this study also indicate that blink detection is effective in eye tracking tasks from event cameras.

4.3. Evaluation Metrics

In this paper we used the Mean Squared Error (MSE) as evaluation metric. It indicates the error between the ground truth and the predicted values, and the execution time per inference.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - x'_i)^2 + (y_i - y'_i)^2\right]$$

4.4. Implementation Details

For learning-based methods, i.e., Ganchao Tan et al. [TWH*22] and our proposed method, Adam [KB14] optimizer was employed at a learning rate of $1e-3$.

Table 1: Comparison of Time and MSE for Single and Cross

Methods	Single Subject	Cross Subject
	MSE(↓)	
Angelopoulos et al. [AMK*21]	0.043	0.051
Tan et al. [TWH*22]	0.061	0.069
Ours	**0.030**	**0.033**

Table 2: Ablation Study

Methods	Single Subject	Cross Subject
	MSE(↓)	
Ours w/o Detection	0.032	0.037
Ours	0.030	0.033

Table 3: Run Time Comparison

Methods	Time(↓)
Angelopoulos et al. [AMK*21]	**0.021s**
Tan et al. [TWH*22]	0.052s
Ours w/o Detection	0.026s
Ours	0.027s

5. Experimental Results

We conducted three different experiments to investigate our method's efficacy: (1) a comparison against existing baselines; (2) a run time comparison against existing baselines to investigate the computational efficiency of our method. We also conducted (3) an ablation study to show the importance of the blink detection. Sec. 5.1 presents our results on accuracy and runtime, as well as a comparison with baseline methods. The ablation study is discussed in Sec. 5.2.

5.1. Comparison against baseline methods

Estimation Performance Comparison The experiments were conducted with two scenarios: Single Subject setting, where the training and inference data come from the same subject, and Cross Subject setting, where the training and inference data come from different subjects. The experimental results for each scenario are shown in Table. 1. As shown in Table 1, our method significantly outperforms existing approaches, demonstrating that high-accuracy eye-tracking using only event data is achievable with our approach. Compared to previous studies, we attribute the improved accuracy to the introduction of blink detection and the use of a network that incorporates an LSTM, which takes temporal information into account.

Run Time Comparison Top three lines of Table 3 shows the run time comparison results between our method and other baselines. Angelopoulos et al.'s method [AMK*21] was the fastest, while our method has achieved similar runtime. Although the network used for lip reading [TWH*22] can learn both spatial and fine temporal features effectively, it requires more inference time compared to our network, which limits the ability to fully exploit the high frame rate characteristic of event cameras. These results show that our method achieves a good balance between accuracy and computational cost.

5.2. Ablation Study

We conducted experiments to verify the effectiveness of introducing blink detection in the eye-tracking task and skipping frames where blinks were detected. The results are presented in the Table 2. As shown in Table 2, incorporating blink detection leads to more accurate eye-tracking. Here, please note that there is no significant difference in execution time as shown in the bottom two lines in Table 3. These findings confirm the effectiveness of integrating blink detection in eye-tracking systems.

6. Conclusion

This paper proposes an eye-tracking method that uses only event data. By utilizing only event data, we leverage the advantages of event cameras, such as high update rates and high power efficiency. Furthermore, by incorporating deep learning into eye-tracking, we achieve more accurate estimations. Additionally, we introduce blink detection into eye-tracking, skipping the inference of frames where blinks occur, thereby enhancing power efficiency and estimation accuracy. Our results demonstrate that our method outperforms existing eye-tracking methods.

Acknowledgement
This work was partially supported by JST Presto JPMJPR22C1 and Keio University Academic Development Funds.

References

[AKNW21] AHLSTRÖM C., KIRCHER K., NYSTRÖM M., WOLFE B.: Eye tracking in driver attention research—how gaze data interpretations influence what we learn. *Frontiers in Neuroergonomics 2* (2021). URL: https://www.frontiersin.org/journals/neuroergonomics/articles/10.3389/fnrgo.2021.778043, doi:10.3389/fnrgo.2021.778043. 1

[AMF23] ADHANOM I. B., MACNEILAGE P., FOLMER E.: Eye tracking in virtual reality: a broad review of applications and challenges. *Virtual Reality* (2023), 4–24. 1

[AMK*21] ANGELOPOULOS A. N., MARTEL J. N., KOHLI A. P., CONRADT J., WETZSTEIN G.: Event-based near-eye gaze tracking beyond 10,000 hz. *IEEE Transactions on Visualization and Computer Graphics 27*, 5 (2021), 2577–2586. doi:10.1109/TVCG.2021.3067784. 1, 3, 4

[AT17] ARAR N. M., THIRAN J.: Robust real-time multi-view eye tracking. *CoRR abs/1711.05444* (2017). URL: http://arxiv.org/abs/1711.05444, arXiv:1711.05444. 1

[BBD24] BARKEVICH K., BAILEY R., DIAZ G. J.: Using deep learning to increase eye-tracking robustness, accuracy, and precision in virtual reality. *Proceedings of the ACM on Computer Graphics and Interactive Techniques 7*, 2 (May 2024), 1–16. URL: http://dx.doi.org/10.1145/3654705, doi:10.1145/3654705. 1

[BWCK20] BAUDRON A., WANG Z. W., COSSAIRT O., KATSAGGELOS A. K.: E3D: event-based 3d shape reconstruction. *CoRR abs/2012.05214* (2020). URL: https://arxiv.org/abs/2012.05214, arXiv:2012.05214. 2

[CCHK23] COSSICH V. R. A., CARLGREN D., HOLASH R. J., KATZ L.: Technological breakthroughs in sport: Current practice and future potential of artificial intelligence, virtual reality, augmented reality, and modern data visualization in performance analysis. *Applied Sciences* (2023). URL: https://api.semanticscholar.org/CorpusID:265768819. 1

[CKK19] CLAY V., KONIG P., KOENIG S. U.: Eye tracking in virtual reality. *Journal of Eye Movement Research 12*, 1 (2019), 1–18. 1

[DSN21] DUARTE L., SAFEEA M., NETO P.: Event-based tracking of human hands. *Sensor Review 41*, 4 (Sept. 2021), 382–389. URL: http://dx.doi.org/10.1108/SR-03-2021-0095, doi:10.1108/sr-03-2021-0095. 2

[FGK*22] FENG Y., GOULDING-HOTTA N., KHAN A., REYSERHOVE H., ZHU Y.: Real-time gaze tracking with event-driven eye segmentation. *CoRR abs/2201.07367* (2022). URL: https://arxiv.org/abs/2201.07367, arXiv:2201.07367. 1

[GDO*19] GALLEGO G., DELBRÜCK T., ORCHARD G., BARTOLOZZI C., TABA B., CENSI A., LEUTENEGGER S., DAVISON A. J., CONRADT J., DANIILIDIS K., SCARAMUZZA D.: Event-based vision: A survey. *CoRR abs/1904.08405* (2019). URL: http://arxiv.org/abs/1904.08405, arXiv:1904.08405. 1

[HZZT24] HUANG K., ZHANG S., ZHANG J., TAO D.: Event-based simultaneous localization and mapping: A comprehensive survey, 2024. URL: https://arxiv.org/abs/2304.09793, arXiv:2304.09793. 2

[KAS*19] KOULIERIS G. A., AKŞIT K., STENGEL M., MANTIUK R. K., MANIA K., RICHARDT C.: Near-eye display and tracking technologies for virtual and augmented reality. *Computer Graphics Forum 38* (2019). URL: https://api.semanticscholar.org/CorpusID:84840630. 1

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 3

[KNM18] KLATT S., NOEL B., MEMMERT D.: Eye tracking in high-performance sports: Evaluation of its application in expert athletes. *International Journal of Computer Science in Sport 17*, 2 (2018), 182–203. 1

[LIB20] LENZ G., IENG S. H., BENOSMAN R.: Event-based face detection and tracking using the dynamics of eye blinks. *Frontiers in Neuroscience 14* (Jul 27 2020), 587. doi:10.3389/fnins.2020.00587. 2

[MFPA18] MITROKHIN A., FERMÜLLER C., PARAMESHWARA C., ALOIMONOS Y.: Event-based moving object detection and tracking. *CoRR abs/1803.04523* (2018). URL: http://arxiv.org/abs/1803.04523, arXiv:1803.04523. 1

[NLB*10] NEUHANN I., LEGE B., BAUER M., HASSEL J., HILGER A., NEUHANN T.: Static and dynamic rotational eye tracking during lasik treatment of myopic astigmatism with the zyoptix laser platform and advanced control eye tracker. *Journal of Refractive Surgery 26*, 1 (2010), 17–27. doi:10.3928/1081597X-20101215-03. 1

[RHGS17] REBECQ H., HORSTSCHAEFER T., GALLEGO G., SCARAMUZZA D.: Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time. *IEEE Robotics and Automation Letters 2*, 2 (2017), 593–600. doi:10.1109/LRA.2016.2645143. 1

[TWH*22] TAN G., WANG Y., HAN H., CAO Y., WU F., ZHA Z.-J.: Multi-grained spatio-temporal features perceived network for event-based lip-reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 20094–20103. 2, 3, 4

[VTMB20] VETTURI D., TIBONI M., MATERNINI G., BONERA M.: Use of eye tracking device to evaluate the driver's behaviour and the infrastructures quality in relation to road safety. *Transportation Research Procedia 45* (2020), 587–595. URL: https://doi.org/10.1016/j.trpro.2020.03.053, doi:10.1016/j.trpro.2020.03.053. 1

[WWHA21] WOOD G., WRIGHT D. J., HARRIS D., ATKINSON G.: Testing the construct validity of a soccer-specific virtual reality simulator using novice, academy, and professional soccer players. *Virtual Reality 25* (2021), 43–51. URL: https://doi.org/10.1007/s10055-020-00441-x, doi:10.1007/s10055-020-00441-x. 1

[ZMZ*23] ZOU S., MU Y., ZUO X., WANG S., CHENG L.: Event-based human pose tracking by spiking spatiotemporal transformer, 2023. URL: https://arxiv.org/abs/2303.09681, arXiv:2303.09681. 1

[ZSZ*24] ZHANG T., SHEN Y., ZHAO G., WANG L., CHEN X., BAI L., ZHOU Y.: Swift-eye: Towards anti-blink pupil tracking for precise and robust high-frequency near-eye movement analysis with event cameras. *IEEE Transactions on Visualization and Computer Graphics 30*, 5 (2024), 2077–2086. doi:10.1109/TVCG.2024.3372039. 2

[ZYCD18] ZHU A. Z., YUAN L., CHANEY K., DANIILIDIS K.: Unsupervised event-based learning of optical flow, depth, and egomotion. *CoRR abs/1812.08156* (2018). URL: http://arxiv.org/abs/1812.08156, arXiv:1812.08156. 2