In the Blink of an Eye: Event-based Emotion Recognition (Supplementary Material)

Haiwei Zhang*
Dalian University of Technology
Dalian, China
haiweizhang32009182@mail.dlut.edu.cn

Jiqing Zhang*
Dalian University of Technology
Dalian, China
jqz@mail.dlut.edu.cn

Bo Dong†
Princeton University
Princeton, USA
bo.dong@princeton.edu

Pieter Peers
College of William & Mary
Williamsburg, USA
ppeers@siggraph.org

Wenwei Wu
Dalian University of Technology
Dalian, China
wuwenwei0206@mail.dlut.edu.cn

Xiaopeng Wei†
Dalian University of Technology
Dalian, China
xpwei@dlut.edu.cn

Felix Heide
Princeton University
Princeton, USA
fheide@cs.princeton.edu

Xin Yang†
Key Laboratory of Social Computing and Cognitive Intelligence of Ministry of Education, Dalian University of Technology
Dalian, China
xinyang@dlut.edu.cn

CCS CONCEPTS

• **Computing methodologies → Computer vision; Supervised learning by classification; Spiking neural networks**.

KEYWORDS

Event-based cameras, eye-based emotion recognition

ACM Reference Format:
Haiwei Zhang, Jiqing Zhang, Bo Dong, Pieter Peers, Wenwei Wu, Xiaopeng Wei, Felix Heide, and Xin Yang. 2023. In the Blink of an Eye: Event-based Emotion Recognition (Supplementary Material). *ACM Trans. Graph.* 1, 1 (August 2023), 4 pages. https://doi.org/10.1145/3588432.3591511

*Equal contribution.
†Corresponding authors.

Authors' addresses: Haiwei Zhang, Dalian University of Technology, Dalian, China, haiweizhang32009182@mail.dlut.edu.cn; Jiqing Zhang, Dalian University of Technology, Dalian, China, jqz@mail.dlut.edu.cn; Bo Dong, Princeton University, Princeton, USA, bo.dong@princeton.edu; Pieter Peers, College of William & Mary, Williamsburg, USA, ppeers@siggraph.org; Wenwei Wu, Dalian University of Technology, Dalian, China, wuwenwei0206@mail.dlut.edu.cn; Xiaopeng Wei, Dalian University of Technology, Dalian, China, xpwei@dlut.edu.cn; Felix Heide, Princeton University, Princeton, USA, fheide@cs.princeton.edu; Xin Yang, Key Laboratory of Social Computing and Cognitive Intelligence of Ministry of Education, Dalian University of Technology, Dalian, China, xinyang@dlut.edu.cn.

0730-0301/2023/8-ART $15.00
https://doi.org/10.1145/3588432.3591511

1 EVENT AGGREGATION

Event-based cameras output asynchronous events, which are incompatible with CNN-based architectures. As such, captured events are normally aggregated into event frames first. We adopt the aggregation algorithm of Zhang *et al.*[Zhang et al. 2021] to accumulate events into event frames. Specifically, in a given time interval, a set of events can be defined as:

$$\mathcal{E} = \{e_k\}_{k=1}^{N} = \{[x_k, y_k, t_k, p_k]\}_{k=1}^{N}, \tag{1}$$

where (x_k, y_k) denotes the pixel location of event e_k; t_k is the timestamp; p is the polarity which means the sign of bright change, with +1 and −1 representing the positive and negative events, respectively.

Then, the events captured within a time window are aggregated as follows:

$$E(x, y, t) = \lfloor \frac{p_k \times \delta(t_{max} - t_k) + 1}{2} \times 255 \rfloor \tag{2}$$

$$t_{max} = max(t_k \times \delta(x - x_k, y - y_k)) \tag{3}$$

where δ is the Dirac delta function.

2 LEAKY INTEGRATE-AND-FIRE (LIF) SPIKING MODEL

A sequence of spikes is called a spike train and is defined as $s(t) = \Sigma_{t^{(f)} \in \mathcal{F}} \delta(t - t^{(f)})$, where $\mathcal{F}$ represents the set of times at which the individual spikes occur [Shrestha and Orchard 2018]. Given N^l incoming spike trains at layer l at timestamp t, the output of the

i-th LIF neuron of next layer at timestamp t, $s_i^{l+1}(t)$, is mathematically defined as:

$$
\begin{aligned}
s_i^{l+1}(t) &= f_s\left(v_i^{l+1}(t)\right), \\
v_i^{l+1}(t) &= \sum_{j=1}^{N^l} w_{ij} s_j^l(t) + \\
&\quad v_i^{l+1}(t-1) f_d\left(s_i^{l+1}(t-1)\right) + b_i^{l+1}, \\
f_d(s(t)) &= \begin{cases} D & s(t) = 0 \\ 0 & s(t) = 1, \end{cases}
\end{aligned} \tag{4}
$$

where w_{ij} is the synaptic weight between the j-th neuron on the l-th layer and the i-th neuron on the layer $l+1$; b_i^{l+1} is an adjustable bias; and D is a constant. The operator $f_s(\cdot)$ is a spike function defined as:

$$
f_s(v) : v \rightarrow s, s(t) := s(t) + \delta(t - t^{(f+1)}), \tag{5}
$$

$$
t^{(f+1)} = \min\{t : v(t) = \Theta, t > t^{(f)}\}, \tag{6}
$$

where Θ is the membrane potential threshold. Note that $f_s(\cdot)$ is non-differentiable.

In our experiment, when receiving membrane potentials X^t, this SNN layer outputs updated spikes, P^t, and updates the recorded membrane potential V^t, as follows:

$$
\begin{aligned}
P^t &= h(V^t - \Theta), \\
V^t &= \alpha V^{t-1}(1 - P^{t-1}) + X^t, \\
h(x) &= \begin{cases} 1 & x >= 0 \\ 0 & x < 0 \end{cases},
\end{aligned} \tag{7}
$$

where Θ is set to 0.3 in all our experiments. The parameter α is a decay factor used for achieving hyperpolarization. The potential value V^t is updated such that, for a spike at timestamp $t-1$, the membrane potential should be reset to 0 by scaling $1 - P^{t-1}$. Essentially, $\alpha(1-P^{t-1})$ corresponds to $f_d(\cdot)$ in Equation 4. Note that the item $\sum_{j=1}^{N^l} w_{ij} s_j^l(t)$ of Equation 4 is replaced by a CNN-based layer, and X^t is the corresponding item here.

3 NETWORK DETAILS

In Figure 1, we show more details of the proposed Spiking Eye Emotion Network (SEEN).

4 EMOTION CLASSIFICATION DATASETS

In Table 1, we report the statistics of widely used emotion recognition datasets. Our SEE dataset is the only one that provides the metadata of lighting conditions and the only one that provides events and corresponding intensity frames.

Our dataset collection follows all required human subject regulations of our institutions; all subjects gave their informed consent to image their eye region and facial expressions (with the possibility of identification) to be used for research purposes and publication.

5 TRAINING SETUP

SEEN is implemented in PyTorch [Paszke et al. 2019] and trained with stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 1e−3. We train SEEN for 180 epochs with a batch size of 32 on an NVIDIA TITAN V GPU. We use the StepLR scheduler to moderate the learning rate. Specifically, the initial learning rate is set to 0.015, the step size is set to 1, and the decay rate is set to 0.94. For a fair comparison, we train or fine-tune all competing models on the same SEE dataset. Based on the requirements of different models, we resize the input accordingly: SEEN is resized to 90×90; All face-based models are resized to 112×112; Eyemotion and EMO inputs are resized to 299×299 and 64×64, respectively. On the SNN side, We use a spiking threshold of 0.3 and a decay factor of 0.2 for all SNN neurons.

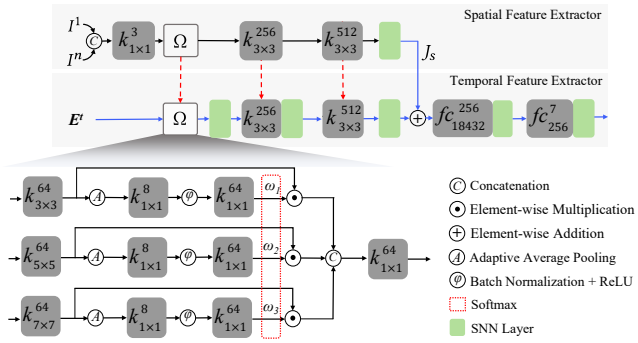

Figure 1: More details of our spatial feature extractor *S* and temporal feature extractor *T*. $k_{i\times i}^{c}$ denotes the convolutional layer where kernel size is $i \times i$, and the number of output channels is c; fc_{c1}^{c2} denotes the fully connected layer where the number of input and output channels are $c1$ and $c2$, respectively.

6 EVALUATION METRICS

For our evaluation, we adopt two widely used metrics for quantitatively assessing the emotion classification performance: Unweighted Average Recall (UAR) and Weighted Average Recall (WAR). UAR reflects the average accuracy of different emotion classes without considering instances per class, while WAR indicates the accuracy of overall emotions [Schuller et al. 2011]. Formally, they are defined as:

$$
UAR = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{TP_i}{TP_i + FN_i}, \tag{8}
$$

$$
WAR = \frac{TP + FN}{TP + TN + FP + FN}, \tag{9}
$$

where N_c is the total number of emotion classes; TP and FP are true and false positive, respectively; TN and FN are true and false negative, respectively.

7 WEIGHT COPY ON TRAINING SPEED

Our weight copy scheme does not only effectively address the spatial and temporal domain gap during training, but it also dramatically increases training speed; see Table 2. The experimental results show that with the experimental setting E13-S0, the training speed is almost doubled.

Table 1: We summarize commonly used emotion recognition datasets with different target areas, including face-based, single-eye based, and double-eye based datasets. Among them, four were collected with RGB-based conventional cameras, and three are collected with Infrared sensors. The proposed SEE dataset is the only one collected by event-based cameras. In addition, the SEE dataset is the only one that provides lighting conditions (LC).

Database	Sequence/Frame	Subjects	Condition	Emotion	Sensor Type	Target Area	LC
CK+[2010]	593/NA	123	Lab	8	RGB	Face	No
BU-4DFE[2008]	606/NA	101	Lab	6	RGB	Face	No
MMI [2005][2010]	2900/NA	25	Lab	7	RGB	Face	No
MUG [2010]	1462/NA	52	Lab	7	RGB	Face	No
Oulu-CASIA [2011]	2880/NA	80	Lab	6	Infrared	Face	No
Eyemotion [2019]	NA/50000	46	HMD	5	Infrared	Double Eyes	No
EMO [2020]	NA/39780	20	HMD	7	Infrared	Single Eye	No
SEE (Ours)	2405/128712	113	HMD	7	Grayscale + Event	Single Eye	Yes

Table 2: Effect of weight copy on training speed.

Train one epoch	E4-S0 E4-S1 E4-S3	E7-S0 E7-S1	E13-S0
without weight copy	8.3s	12.4s	20.6s
Ours	5.5s	7.2s	10.9s

8 GENDER AND AGE DISTRIBUTION EFFECTS

We conduct experiments to analyze the impact of gender and age. The results in Table 3 and 4 show that our approach offers more accurate results for female than male participants, and it is slightly more effective in younger age categories.

Table 3: Performance on the gender distribution.

Gender (Num)	Metrics	E4-S0	E4-S1	E4-S3	E7-S0	E7-S1	E13-S0
Male (66)	WAR	76.7	78.9	82.0	77.7	80.3	80.7
	UAR	77.4	79.5	82.7	78.4	80.9	81.3
Female (45)	WAR	81.8	84.1	86.2	82.6	85.8	84.7
	UAR	81.2	83.4	86.0	82.2	85.0	83.9

Table 4: Performance on the age distribution.

Age group (Num)	Metrics	E4-S0	E4-S1	E4-S3	E7-S0	E7-S1	E13-S0
[19,23] (80)	WAR	79.0	81.3	83.6	80.2	82.4	82.4
	UAR	79.4	81.7	84.0	80.6	82.6	82.6
[24,28] (31)	WAR	77.3	79.3	83.7	77.6	82.4	81.7
	UAR	78.0	80.2	84.3	78.5	83.3	82.7

9 ADDITIONAL ABLATION STUDY

In Table 5, we provide two additional ablation study: a) Impact of Input: Experiments (A) and (B); b) Influence of multiscale perception: Experiments (I)-(K).

Impact of Input. SEEN leverages both spatial and temporal cues for recognizing emotion. In experiments (A) and (B), the input of the spatial feature extractor S and the temporal feature extractor T is changed to events and frames, respectively. These experiments demonstrate the effectiveness of leveraging both domains. These two experiments also show that our SEEN can seamlessly work with pure event or intensity frames albeit less effectively.

Influence of SEEN components. We investigate the effectiveness of multiscale perception (experiments (L)-(N)). All three experiment groups show that SEEN with all components offers the best performance. We witness a significant performance degradation in (L)-(N), validating that multiscale cues and self-attention (*i.e.*, Ω) are essential for emotion classification accuracy.

REFERENCES

Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. 2010. The MUG facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE, 1–4.

Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. 2019. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1626–1635. https://doi.org/10.1109/WACV.2019.00178

Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*. IEEE, 5 pp.–. https://doi.org/10.1109/ICME.2005.1521424

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

B. Schuller, B. Vlasenko, F. Eyben, M. Wo?Llmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll. 2011. Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing* 1, 2 (2011), 119–131. https://doi.org/10.1109/T-AFFC.2010.8

Sumit B Shrestha and Garrick Orchard. 2018. Slayer: Spike layer error reassignment in time. *Advances in neural information processing systems* 31 (2018). 32nd Conference on Neural Information Processing Systems (NIPS), Montreal, CANADA, DEC 02-08, 2018.

Michel Valstar, Maja Pantic, et al. 2010. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*. Paris,

Table 5: Quantitative ablation comparisons show that: a) both spatial and temporal cues are essential for providing discriminative features; b) all components of SEEN contribute to the overall performance, except setting E4-S0 in row G. c) potential averaging is necessary for offering a more accurate performance.

		E4-S0		E4-S1		E4-S3	
	Networks	WAR	UAR	WAR	UAR	WAR	UAR
A	Both events	48.5	48.9	52.4	52.6	53.4	53.9
B	Both frame	62.3	62.3	64.6	64.5	69.9	70.1
C	w/o I^n	77.1	77.6	79.9	80.2	81.3	81.8
D	$I^n \rightarrow I^2$	76.4	76.9	80.1	80.6	81.8	82.2
E	$[I^1, ..., I^n]$	78.0	78.4	79.9	80.2	82.9	83.3
F	No weight copy	77.5	78.0	79.6	80.0	82.1	82.6
G	No Att. weight copy	78.7	79.2	80.7	81.1	83.0	83.2
H	SNN → CNN	50.2	50.2	53.2	53.2	55.7	55.6
I	SNN → LSTM	52.9	53.0	55.3	55.2	55.8	55.7
J	SNN → Transformer	69.2	69.8	73.6	74.2	77.1	77.3
K	SNN → 3D CNN	54.3	54.3	57.7	57.7	59.9	59.9
L	$\Omega_{(3)}$	70.1	70.6	74.2	74.5	76.4	76.9
M	$\Omega_{(3,5)}$	78.0	78.5	79.8	80.1	80.3	80.7
N	$\Omega_{(3,5,7)}$ w/o self-Att.	77.1	77.6	80.0	80.4	81.7	82.0
O	Last potential	76.6	77.2	78.8	79.2	81.1	81.7
P	Last spike	55.7	54.8	59.5	58.9	63.2	62.8
Q	Mean spike	63.5	63.2	64.1	63.6	69.7	69.5
R	Ours	78.6	79.1	80.9	81.3	83.6	84.1

France., J65–J70. 7th International Conference on Language Resources and Evaluation (LREC), Valletta, MALTA, MAY 17-23, 2010.

Hao Wu, Jinghao Feng, Xuejin Tian, Edward Sun, Yunxin Liu, Bo Dong, Fengyuan Xu, and Sheng Zhong. 2020. EMO: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. 448–461. https://doi.org/10.1145/3386901.3388917

L. Yin, X. Chen, S. Yi, T. Worm, and M. Reale. 2008. A high-resolution 3D dynamic facial expression database. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*. 1–6. https://doi.org/10.1109/AFGR.2008.4813324

Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. 2021. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13043–13052. https://doi.org/10.1109/ICCV48922.2021.01280

Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29, 9 (2011), 607–619. https://doi.org/10.1016/j.imavis.2011.07.002