ISSAI

NAZARBAYEV UNIVERSITY | Institute of Smart Systems and Artificial Intelligence

Face and Facial Landmark Detection for Event-based Imaging

Tomiris Rakhimzhanova

Department of Robotics, School of Engineering and Digital Sciences

Nazarbayev University

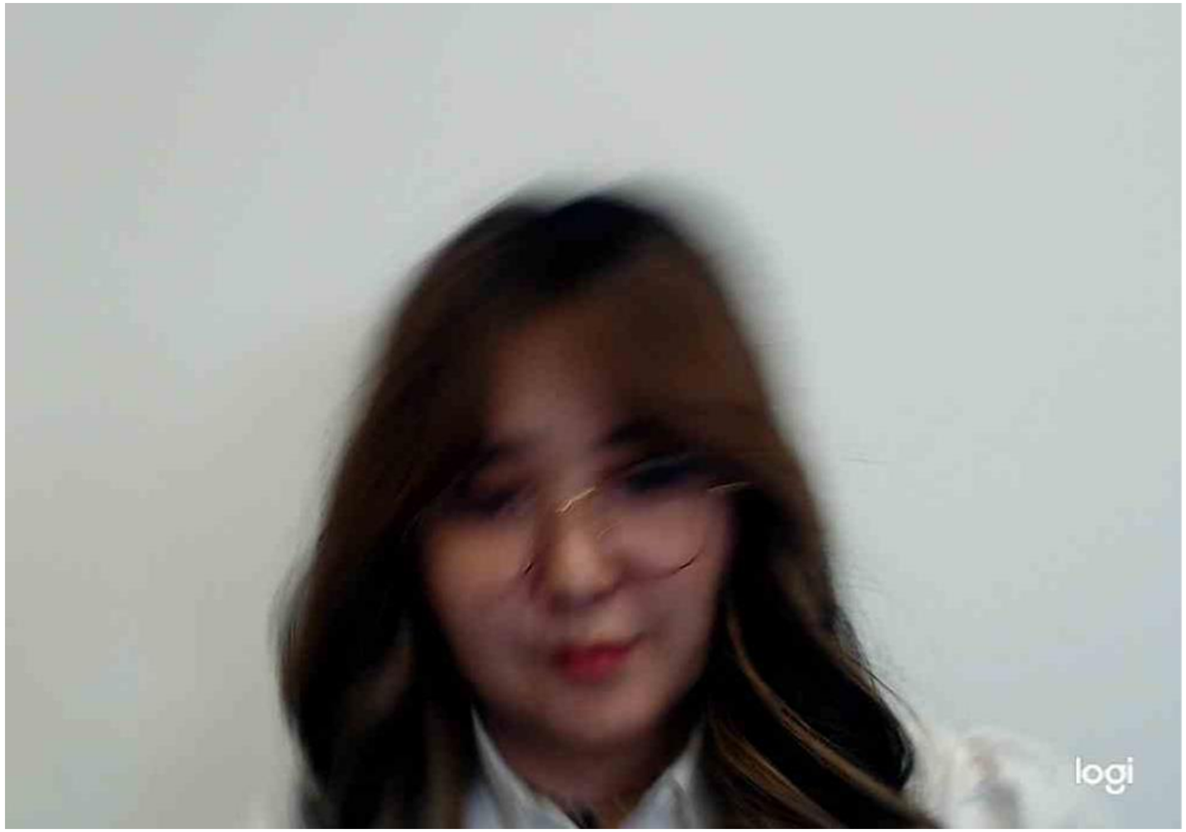
issai.nu.edu.kz

Outline

- ❖ Introduction
- ❖ Event-based Imaging for Robotics
- ❖ Face and Facial Landmarks Detection
- ❖ Thesis Objectives
- ❖ Faces in Event Streams (FES) Dataset
- ❖ Methodology
- ❖ Results and Experiments
- ❖ Conclusion

Despite significant advances in imaging, frame-based cameras still have a number of shortcomings.

Latency & Motion Blur



Dynamic Range

Event-based Imaging

- ❖ Bioinspired sensors that measures only brightness changes in the scene
- ❖ Low-latency (~1 μs)
- ❖ No motion blur

Characteristics	Frame-based camera	Event camera
Update rate	syncronous	aynschronous
Latensy	yes	≈ 0
Dynamic range	53 db	>120 db
Motion blur	exist	absent
Temporal resolution	low	high

Fig.1 Comparison between conventional and event based camera (Adapted from [2])

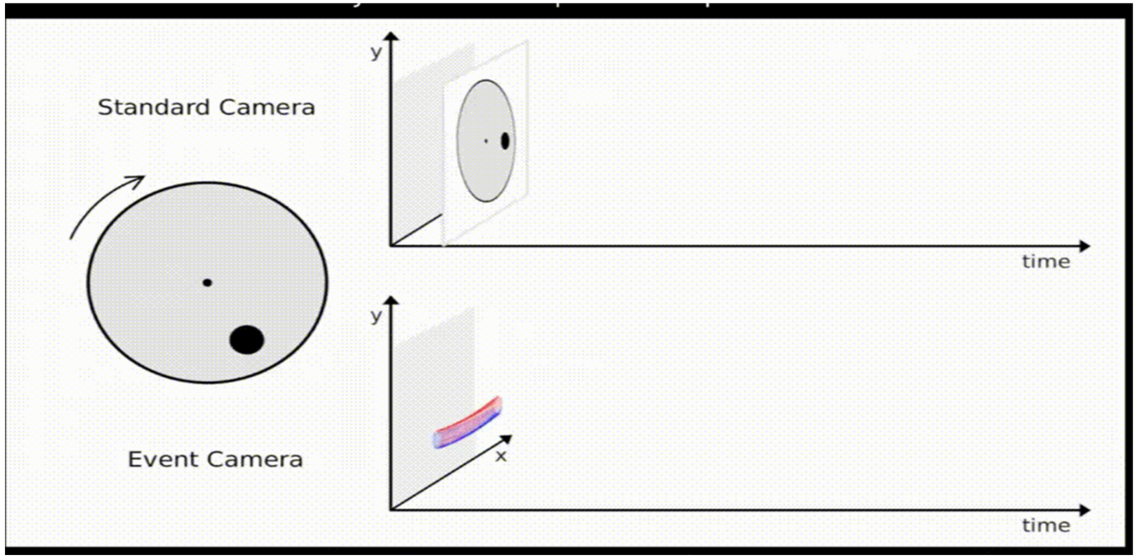


Fig.2 Difference between outputs of cameras. (Retrieved from [4])

- ❖ Ultra-low power (mW)
- ❖ High dynamic range >120 db

Operating Principles of Event Cameras

- ❖ Similar to the human retina work;
- ❖ The light first hits the photoreceptor of the pixel;
- ❖ Each peak event is then processed in a bipolar cell;
- ❖ The signal voltage values are compared by the comparators in the third step.

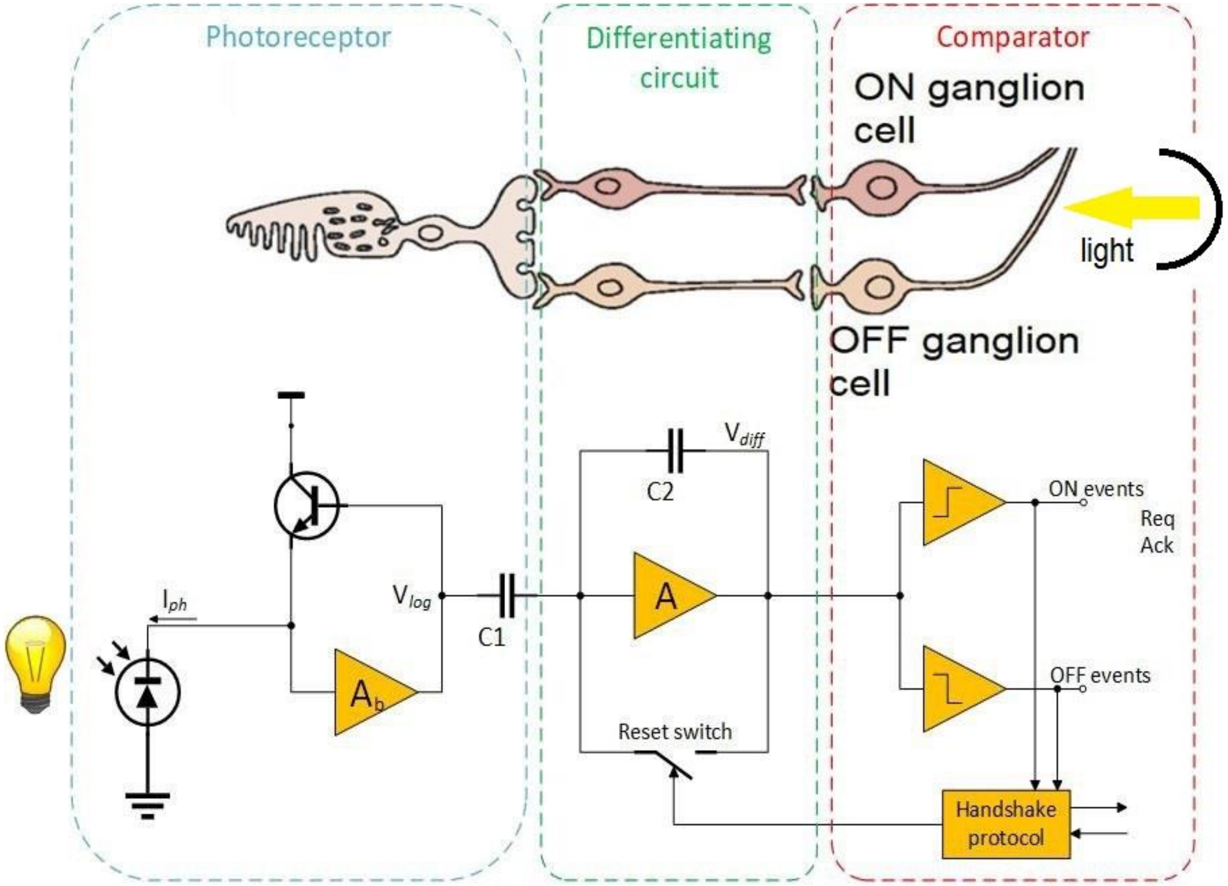


Figure 2-2: Pixel technical diagram of DAVIS event-based sensor. Adapted from [19]

Operating Principles of Event Cameras

Mathematical representation of data detection by pixel:

$$Log\left(I_{x,y,t+\Delta t}\right) - \log\left(I_{x,y,t+\Delta t}\right) \geq pC$$

Set of ON and OFF events:

$$p(x,y,t) = \begin{cases} ON\ if\ I(x,y,t) - T(x,y) > 0 \\ OFF\ if\ I(x,y,t) - T(x,y) < 0 \end{cases}$$

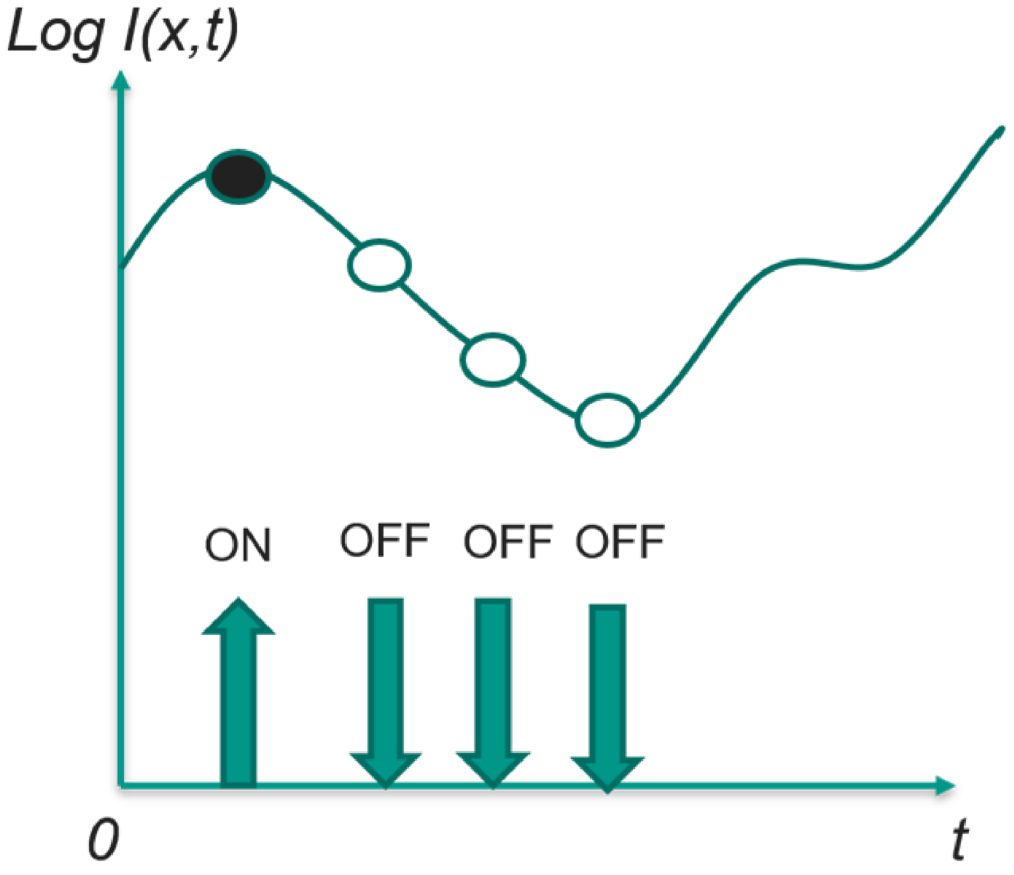


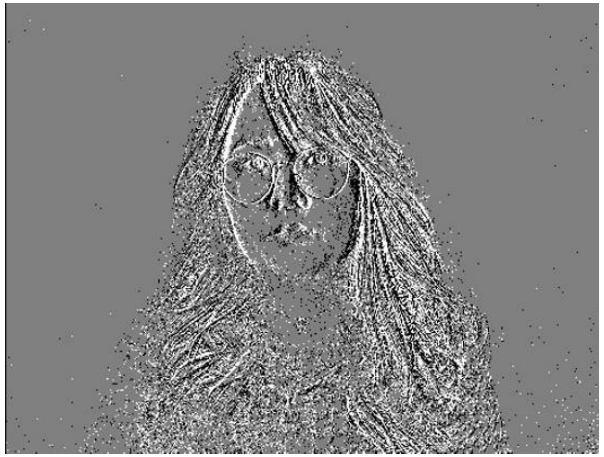
Fig.3 Graphical representation

Output Data Format

- Pixel location - x and y;
- p – ON (1) and OFF (0) events;
- t - timestamp in microseconds

x	y	p	t
71	55	1	48
288	55	1	48
278	54	1	49
13			[illegible]

Traditional vision algorithms cannot be used



Image–like visualization with accumulation time

Grayscale transform

Event-based Imaging for Robotics

In the article (Li et.al, 2020):

- ❖ constructed a robotic grasping dataset named Event-Grasping dataset ;
- ❖ developed a deep neural network for grasping detection that considers the angle learning problem as classification instead of regression.

Paper (Taunyazov et al, 2020):

- ❖ this work contributes an event-driven visual-tactile perception system;
- ❖ authors developed a novel biologically-inspired tactile sensor NewTouch;
- ❖ visual-tactile system (using the NeuTouch and Prophesee event camera).

Authors in the article (Mueggler et.al, 2015):

- ❖ proposes a method to predict collisions with objects thrown at a quadrotor using a pair of event-based sensors;
- ❖ demonstrated that method allows a quadrotor initiating evasive maneuvers early.

In the article (Vidal et.al, 2020):

- ❖ demonstrated the autonomous quadrotor flight using an event camera for state estimation, unlocking flight scenarios that were not reachable with traditional visual-inertial odometry;
- ❖ the first state estimation pipeline that fuses three sensors.

Paper (Gallego et al, 2020):

- ❖ presented an approach to track the 6-DOF pose of an arbitrarily moving event camera from an existing photometric depth map in natural scenes;
- ❖ compared the 6-DOF motion of the event camera with standard cameras.

Authors in the article (Falanga et.al, 2015):

- ❖ study the effects that perception latency has on the maximum speed a robot can reach to safely navigate through an unknown cluttered environment;
- ❖ showed the maximum latency that the robot can tolerate to guarantee safety.

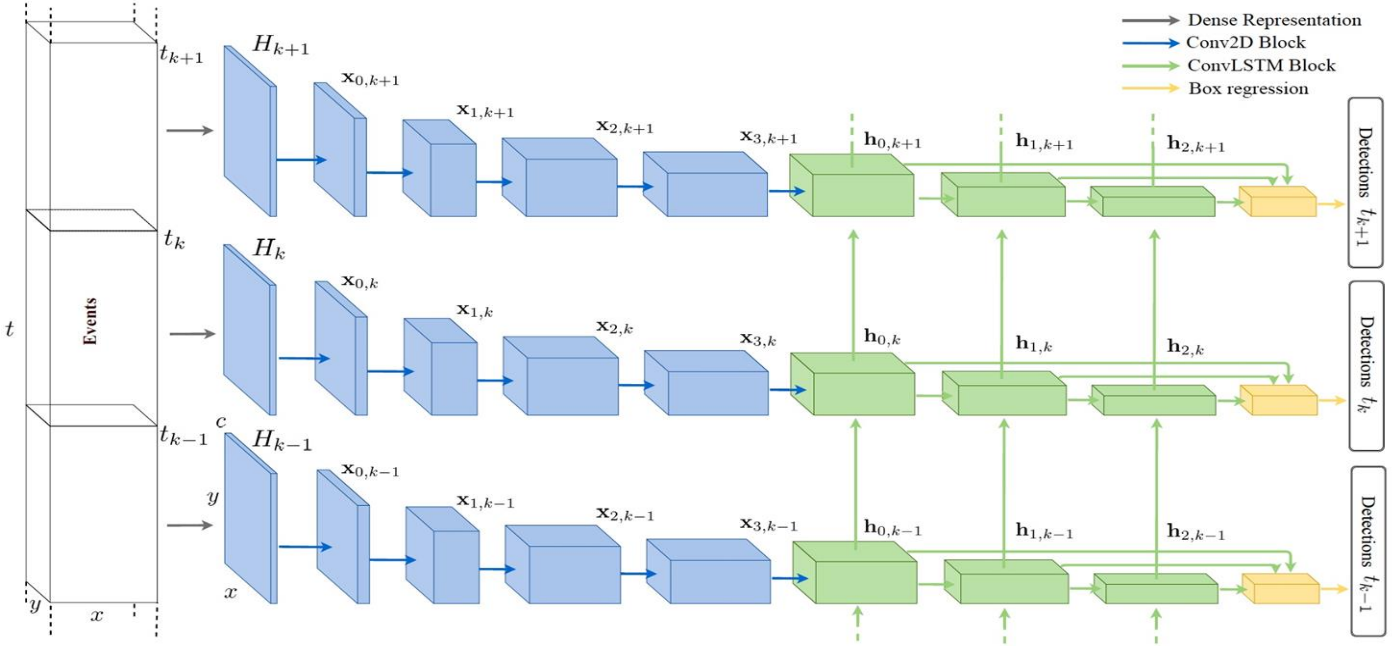
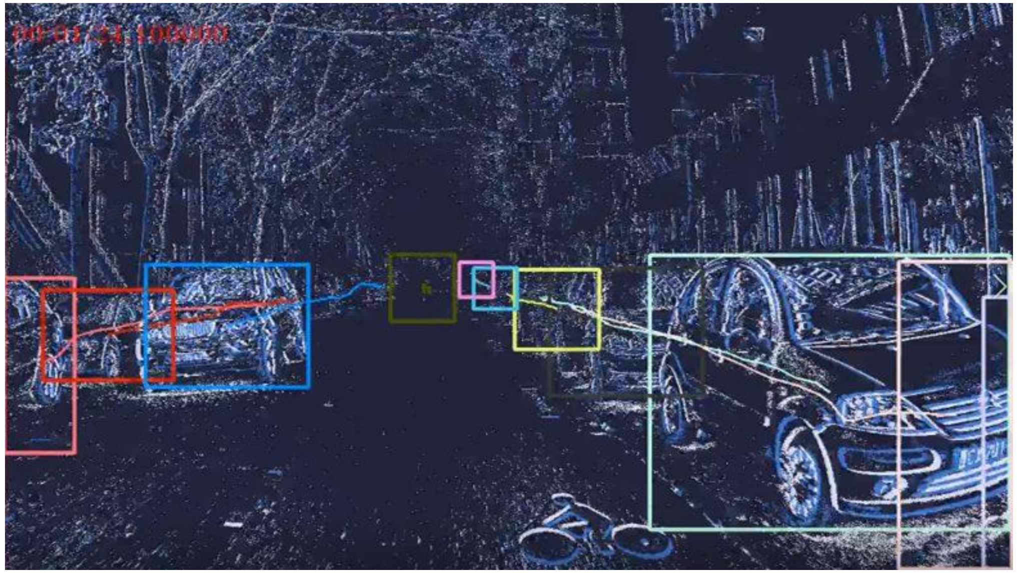


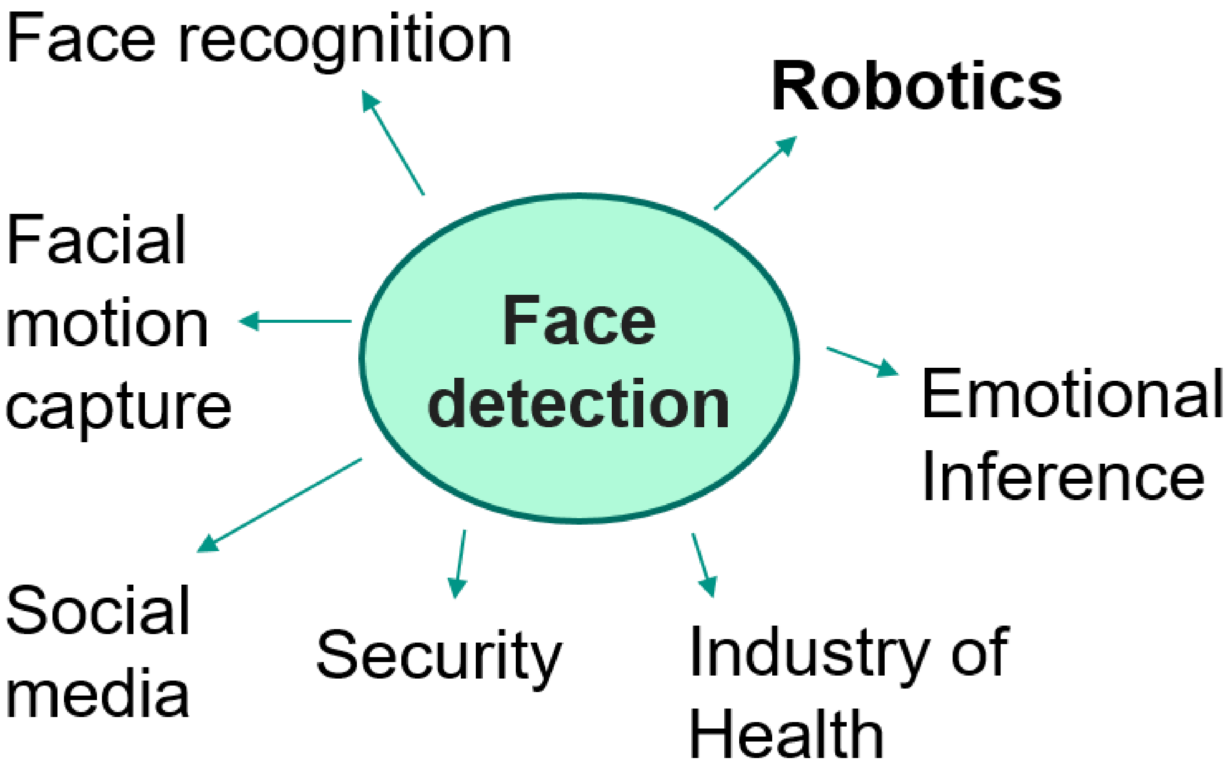
Fig.5 Prophesee architecture for object detection. Retrieved from [10].

The dataset contains more than 14 hours recordings of a 1 megapixel event camera and the it consist 7 classes: pedestrians, two wheelers, cars, trucks, buses, traffic signs, traffic lights



Perot et.al introduced of a novel architecture for event-based object detection. Authors showed that directly predicting the object locations is more efficient and more accurate than applying a detector on the gray-level images.

Face and Facial Landmarks Detection



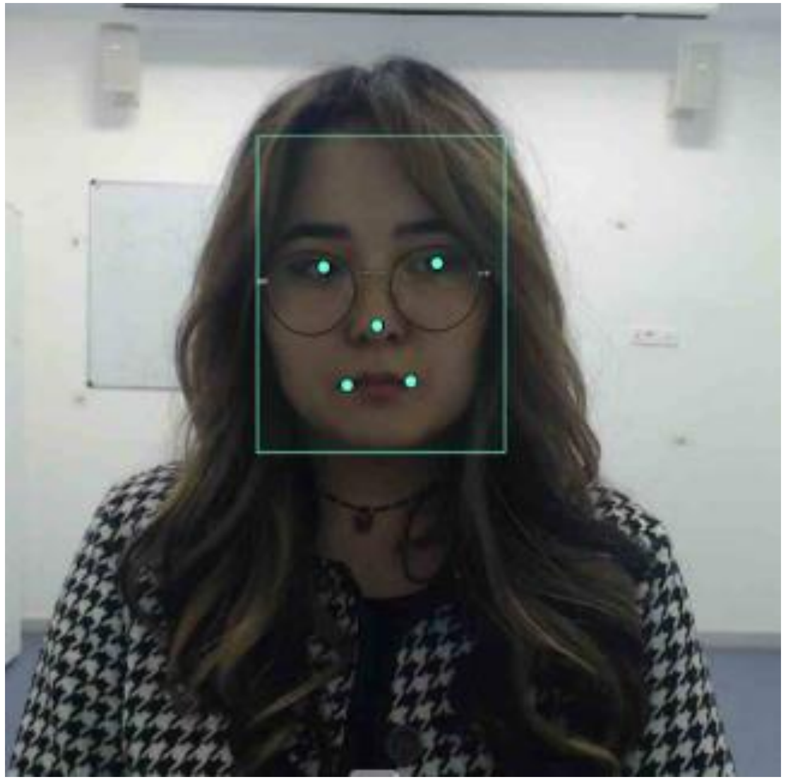


Fig.6 Face and facial landmarks detection

In the article (Barua et.al, 2016):

- limited datasets for face detection;
- used RGB images dataset;
- reconstructed frame-based images to event-based output;
- apply face detection on reconstructed gray-scale images.

Face detection using eye blink:

Papers (Lenz et al, 2020) and (Cian Ryan et.al, 2020):

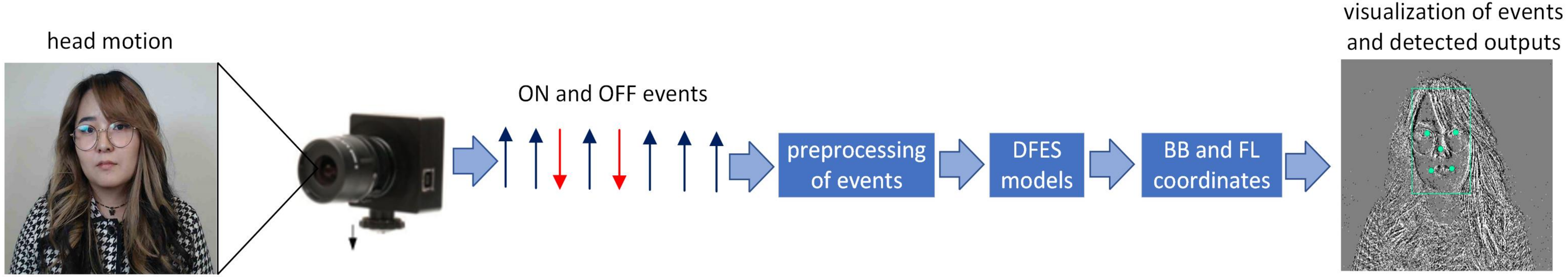
- algorithm for eye blink detection;
- using the area of eye blink detection, probabilistic places a bounding box for the face.

Papers [20],[21] identify problems with absent of large dataset of event-streams:

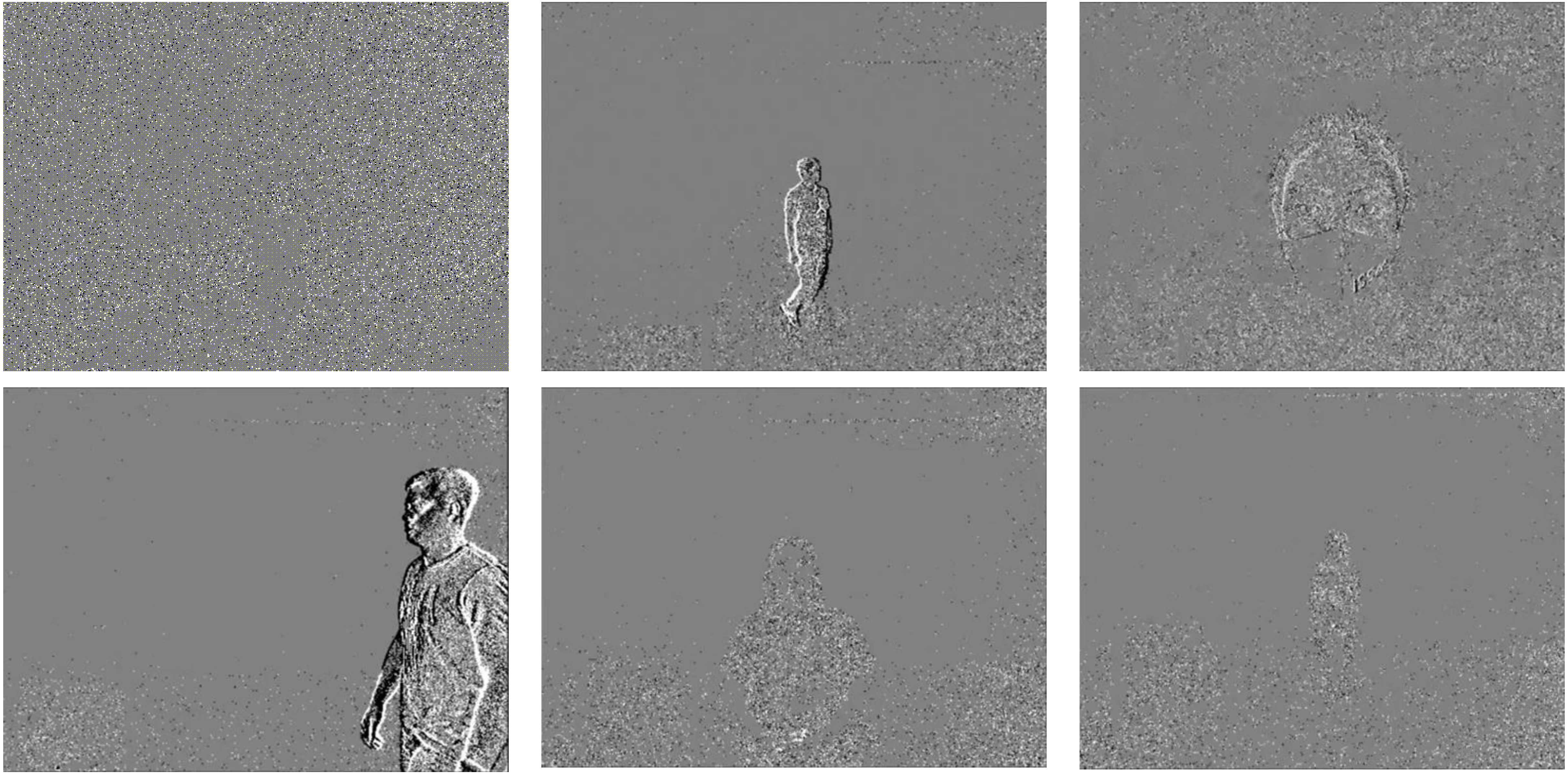
- propose the transformation of frame-based dataset images into images similar to event-based

Thesis Objectives

- Created and published the first rich and structured dataset of 689 minutes of machine learning-transformed event streams, captured at different lighting conditions, from different viewpoints and distances, with multiple people in the scene, and a greater number (73) and diversity of participants;
- For the first time, 12 research-based DFES models were created and trained for face and landmark detection that use outputs based directly on events;
- Experiments and comparative analysis of DFES models.



Faces in Event Streams (FES) Dataset



Faces in Event Streams (FES) Dataset:

- ❖ Two major parts: controlled (laboratory) and uncontrolled (wild);
- ❖ 73 subjects: 31 female and 42 male participants;
- ❖ 59 experiments for each subject:
  - under bright and dim lighting conditions;
  - 50, 150, and 400 cm distances from the camera;
  - head postures and movements: left-right, up-down, circular movements of the head and counting;
  - walking: zigzag, walking toward the camera, and sideways;
  - Uncontrolled data were collected in indoor environments

	FES dataset
Duration	693 min
Participants	73
Resolution	480 x 360
Camera	Prophese PPS3MVCD
Environment	controlled, wild
Bounding box	✓
Facial landmarks	5 points

Dataset Annotation and Visualization:

- An image-like visualization of event streams is obtained by accumulating events over a short period of time (the accumulation time);
- Event streams were rendered by defining ON events as white pixels, OFF events as black pixels, and background as gray.
- Grayscale images obtained using Metavision software;
- The annotation was done by ISSAI laboratory moderators using CVAT annotation tool;

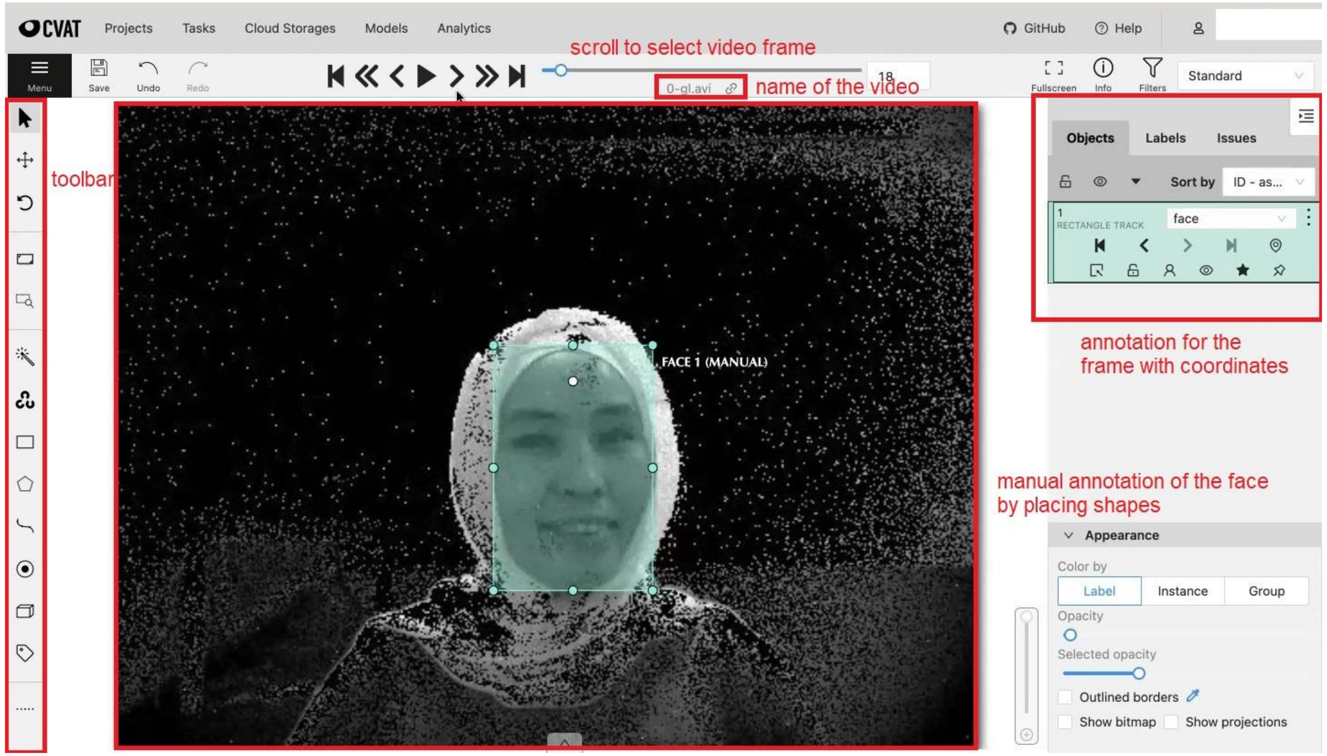


Fig. 3 Screenshots of the free CVAT toolkit (https://cvat.ai).

Deep Learning Model Architecture

- ❖ Event stream is represented as a sequence of events:

$$\boldsymbol{E}=\{e = (xi, yi, pi, ti)\}$$

- ❖ Sequence of events is transformed into a tensor map *Hk* using histogram preprocessing method;
- ❖ qk = the encoded information from the past stored as an internal state;
- ❖ The original feature extractor in our model was changed to the ResNet-18, ResNet-34, and ResNet-50 variants.

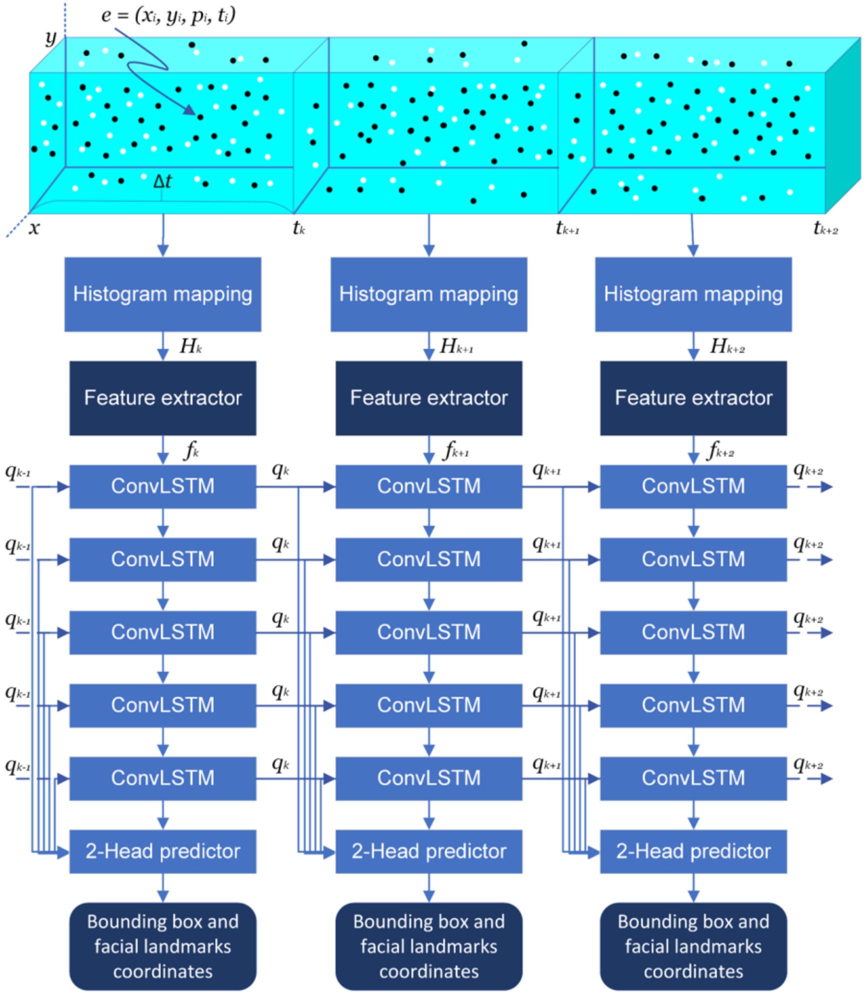


Fig. 8 Our model architecture. Adapted from [10].

Methodology of experiments

- **Determination of the Optimal Accumulation Time**

Training models on a FES dataset with different accumulation times for choosing the optimal accumulation time

- **Training models for bounding box detection**.

The code for determining the architecture of the model was written using the PyTorch tool

- **Training models for bounding box and facial landmarks detection**

Adapting the code for adding facial landmarks detection.

- **Inference Time and Real-time Detection Experiment**

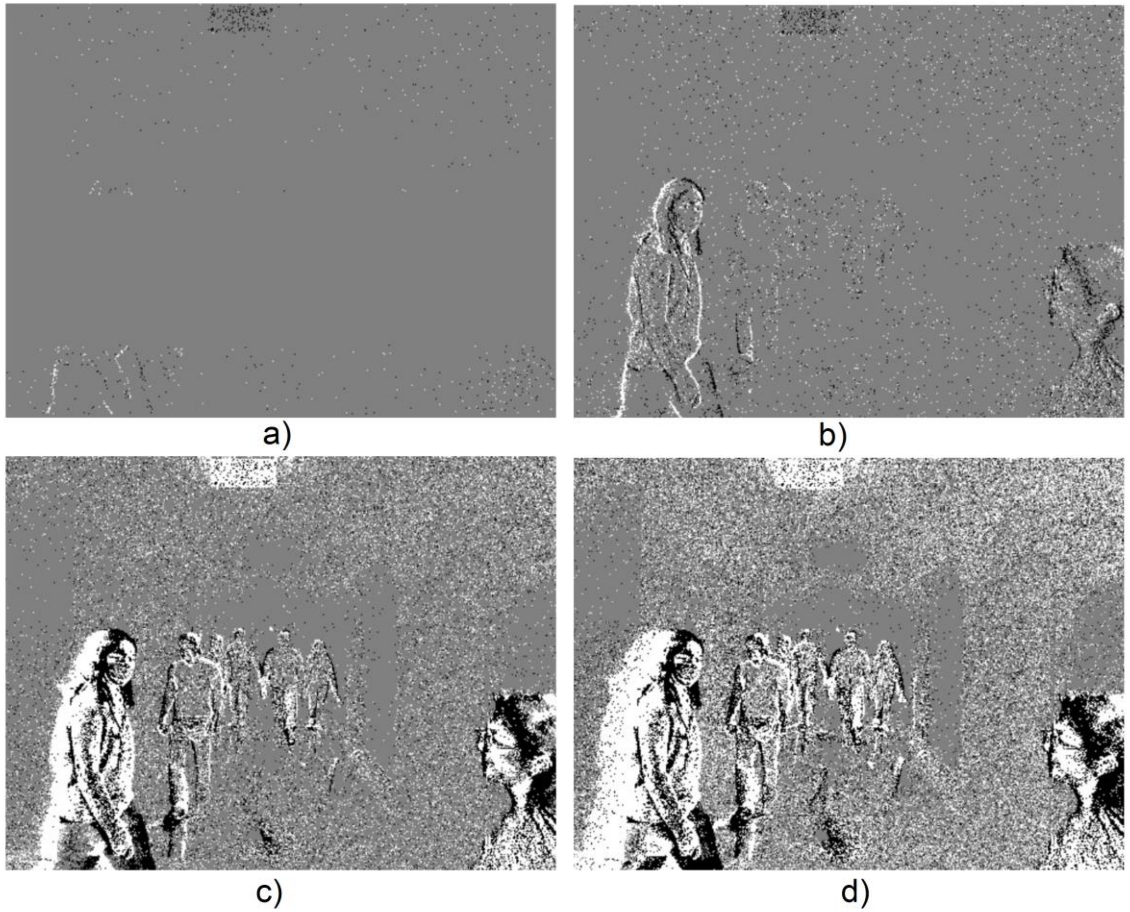


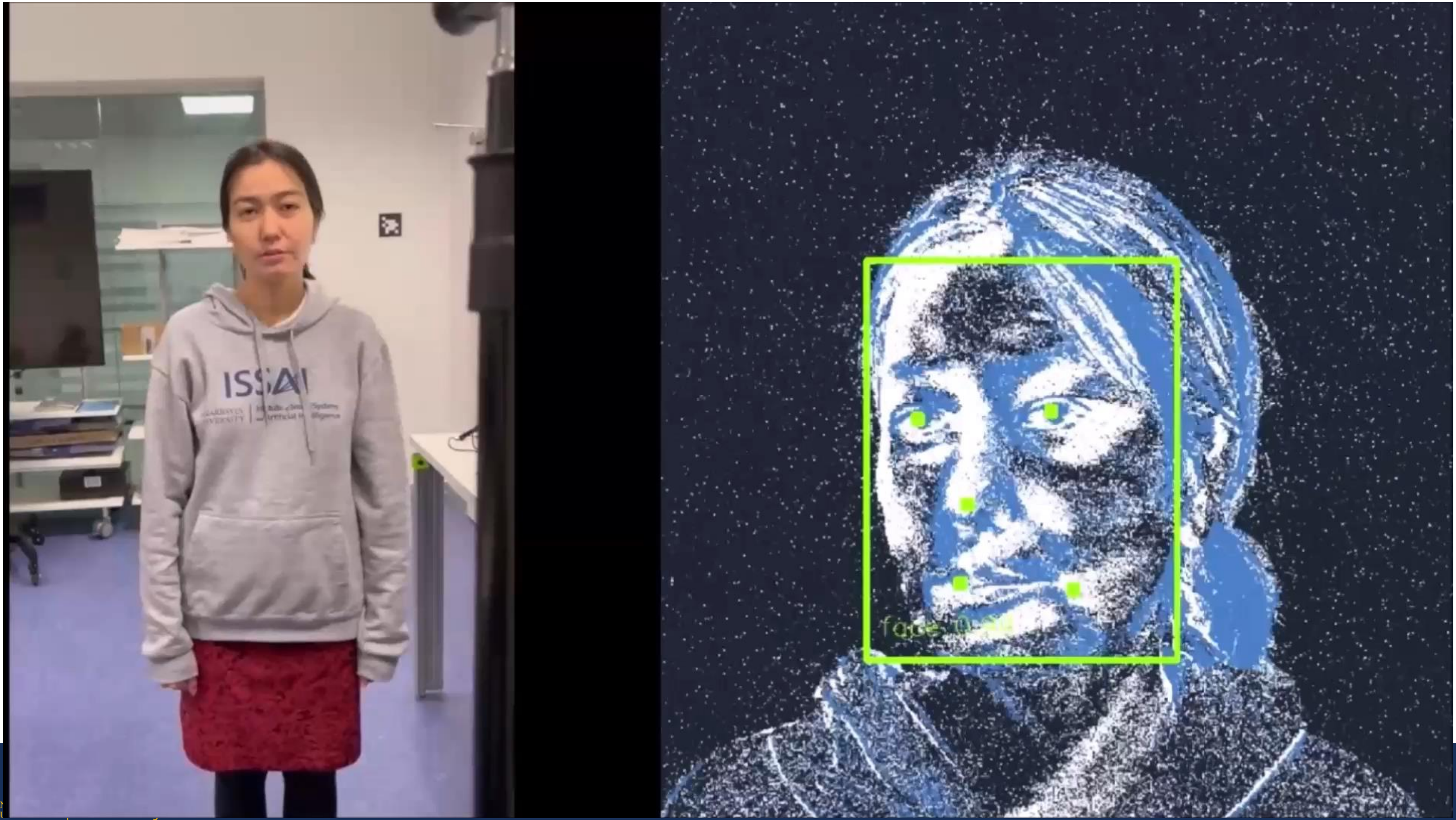
Figure 9. Data visualization of event streams at different accumulation times: a)200 μs, b) 5 ms, c) 33 ms, and d) 100 ms.

mAP_{50} results for Face Bounding Box Detection

			mAP50 Laboratory Testing Set				mAP_50 Wild Testing Set				mAP_50 Overall Testing Set			
Model	Feature extractor	Delta_t	Large	Medium	Small	Overall	Large	Medium	Small	Overall	Large	Medium	Small	Overall
$DFES_{BB}$	Original	33 ms	0.375	0.4	0.328	0.353	0.57	0.13	0.06	0.1	0.375	0.4	0.328	0.353
$DFES_{BB}$	Original	50 ms	**0.99**	**0.978**	**0.97**	**0.978**	**0.919**	0.273	0.138	0.146	**0.99**	**0.976**	0.8	0.93
$DFES_{BB}$	Original	100 ms	0.989	0.973	0.964	0.977	0.8	0.231	0.133	0.15	0.989	0.964	0.8	0.927
$DFES_{BB}$	ResNet-18	50 ms	**0.99**	0.974	**0.97**	**0.978**	0.83	0.3	**0.149**	0.165	**0.99**	0.97	**0.827**	**0.936**
$DFES_{BB}$	ResNet-34	50 ms	0.989	0.962	0.952	0.965	0.794	**0.436**	0.17	**0.182**	**0.99**	0.969	0.8	0.931
$DFES_{BB}$	ResNet-50	50 ms	0.988	0.964	0.9	0.957	0.73	0.12	0.05	0.1	0.988	0.96	0.715	0.884
$DFES_{FL+BB}$	Original	33 ms	0.371	0.397	0.38	0.37	0.599	0.443	0.26	0.252	0.369	0.393	0.325	0.347
$DFES_{FL+BB}$	Original	50 ms	0.989	**0.978**	**0.871**	**0.973**	0.728	**0.782**	0.482	0.528	0.989	**0.97**	0.7	**0.918**
$DFES_{FL+BB}$	Original	100 ms	0.989	0.976	0.7	0.937	0.64	0.7	**0.645**	**0.653**	0.989	0.949	0.575	0.868
$DFES_{FL+BB}$	ResNet-18	50 ms	**0.99**	0.969	0.8	0.96	0.72	0.75	0.47	0.5	**0.99**	0.96	0.7	0.9
$DFES_{FL+BB}$	ResNet-34	50 ms	**0.99**	**0.978**	0.869	0.966	**0.789**	0.75	0.498	0.54	**0.99**	**0.97**	**0.72**	0.912
$DFES_{FL+BB}$	ResNet-50	50 ms	0.985	0.928	0.75	0.925	0.184	0.282	0.124	0.138	0.984	0.873	0.52	0.8

MNE results for Facial Landmarks Detection

			NME Laboratory Testing Set				NME Wild Testing Set				NME Overall Testing Set			
Model	Feature extractor	Delta_t	Large	Medium	Small	Overall	Large	Medium	Small	Overall	Large	Medium	Small	Overall
$DFES_{FL+BB}$	Original	33 ms	**0.335**	**0.298**	**0.577**	**0.394**	16.69	15	15.87	15.74	**0.358**	1.5	5.387	2.52
$DFES_{FL+BB}$	Original	50 ms	0.398	0.342	0.6	0.44	16.09	**12.01**	**13.8**	**13.5**	0.432	1.44	4.85	2.338
$DFES_{FL+BB}$	Original	100 ms	0.57	0.45	0.83	0.61	**9.965**	14.23	14.72	14.73	0.6	**1.35**	**3.74**	**1.99**
$DFES_{FL+BB}$	ResNet-18	50 ms	0.414	0.373	1.276	0.656	16.8	12.7	15.9	15.3	0.45	1.615	5.9	2.786
$DFES_{FL+BB}$	ResNet-34	50 ms	0.383	0.325	0.6	0.42	17.9	12.5	14	13.7	0.414	1.638	4.79	2.365
$DFES_{FL+BB}$	ResNet-50	50 ms	0.84	1.98	3.03	1.8	16.54	14.23	15.46	15.32	0.92	3.281	6.98	3.65


ISSAI
face

References

[1] D. J. Griffiths and A. Wicks, "High Speed High Dynamic Range Video," in IEEE Sensors Journal, vol. 17, no. 8, pp. 2472-2480, 15 April15, 2017, doi: 10.1109/JSEN.2017.2668378.

[2] Barua, S., Miyatani, Y., & Veeraraghavan, A. (2016). Direct face detection and video reconstruction from event cameras. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/wacv.2016.7477561

[3] 3 G. Gallego et al., "Event-Based Vision: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 154-180, 1 Jan. 2022, doi: 10.1109/TPAMI.2020.3008413.

[4] Ruolin Sun, Dianxi Shi, Yongjun Zhang, Ruihao Li, Ruoxiang Li, "Data-Driven Technology in Event-Based Vision", Complexity, vol. 2021, 19, 2021. https://doi.org/10.1155/202 1/6689337

[5] T. Delbrück, B. Linares-Barranco, E. Culurciello and C. Posch, "Activity-driven, event-based vision sensors," Proceedings of 2010 IEEE International Symposium on Circuits and Systems, 2010, pp. 2426-2429, doi: 10.1109/ISCAS.2010.5537149.

[6] C. Posch, R. Benosman and R. Etienne-Cummings, "Giving machines humanlike eyes," in IEEE Spectrum, vol. 52, no. 12, pp. 44-49, December 2015, doi: 10.1109/MSPEC.2015.7335800.

[7] Y. Suh et al., "A 1280×960 Dynamic Vision Sensor with a 4.95-µm Pixel Pitch and Motion Artifact Minimization," 2020 IEEE International Symposium on Circuits and Systems (ISCAS), 2020, pp. 1-5, doi: 10.1109/ISCAS45731.2020.9180436.

[8] Ryan, Cian & Sullivan, Brian & Elrasad, Amr & Lemley, Joseph & Kielty, Paul & Posch, Christoph & Perot, Etienne. (2020). Real-Time Face & Eye Tracking and Blink Detection using Event Cameras.

[9] Lenz G, Ieng S-H and Benosman R (2020) Event-Based Face Detection and Tracking Using the Dynamics of Eye Blinks. Front. Neurosci. 14:587. doi: 10.3389/fnins.2020.00587 [11] Gao, Shan & Guo, Guangqian & Huang, Hanqiao & Cheng, Xuemei & Chen, C.. (2020). An End-to-End Broad Learning System for EventBased Object Classification. PP. 1-1. 10.1109/ACCESS.2020. 2978109.

[9] F. Mahlknecht, D. Gehrig, J. Nash, F. M. Rockenbauer, B. Morrell, J. Delaune, and D. Scaramuzza, “Exploring event camera-based odometry for planetary robots,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8651–8658, 2022.

[10] P. Etienne, d. T. Pierre, N. Davide, M. Jonathan, and S. Amos, “Learning to detect objects with a 1 megapixel event camera,” Advances in Neural Information Processing Systems, vol. 33, pp. 16 639–16 652, 2020

[11] Li B, Cao H, Qu Z, Hu Y, Wang Z, Liang Z. Event-Based Robotic Grasping Detection With Neuromorphic Vision Sensor and Event-Grasping Dataset. Front Neurorobot. 2020 Oct 8;14:51. doi: 10.3389/fnbot.2020.00051. PMID: 33162883; PMCID: PMC7580650.

[12] D. Weikersdorfer and J. Conradt, "Event-based particle filtering for robot self-localization," 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2012, pp. 866-870, doi: 10.1109/ROBIO.2012.6491077.

[13] T. Taunyazov, W. Sng, B. Lim, H. Hian See, J. Kuan, A. Fatir Ansari, B. Tee, and H. Soh, “Event-driven visual-tactile sensing and learning for Robots,” Robotics: Science and Systems XVI, 2020.

[14] E. Mueggler, N. Baumli, F. Fontana, and D. Scaramuzza, “Towards evasive maneuvers with quadrotors using Dynamic Vision Sensors,” 2015 European Conference on Mobile Robots (ECMR), 2015.

[15] G. Gallego, J. E. A. Lund, E. Mueggler, H. Rebecq, T. Delbruck, and D. Scaramuzza, “Event-based, 6-DOF camera tracking from photometric depth maps,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2402–2412, 2018.

[16] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, “Ultimate Slam? combining events, images, and IMU for robust visual slam in HDR and high-speed scenarios,” *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, 2018.

[17] M. T. H. Fuad, A. A. Fime, D. Sikder, M. A. R. Iftee, J. Rabbi, M. S. Al-Rakhami, A. Gumaei, O. Sen, M. Fuad, and M. N. Islam, “Recent advances in deep learning techniques for face recognition,” IEEE Access, vol. 9, pp. 99 112–99 142, 2021

[18] D. Falanga, S. Kim, and D. Scaramuzza, “How fast is too fast? the role of perception latency in high-speed sense and avoid,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1884–1891, 2019.

[20] Timothée Masquelier and Simon J Thorpe. Unsupervised learning of visual features through spike timing dependent plasticity. PLoS Computational Biology,3(2):247–257, 2007.

[21] Ana I. Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso Garcia, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1–9, 2018.

[22] Timothée Masquelier and Simon J Thorpe. Unsupervised learning of visual features through spike timing dependent plasticity. PLoS Computational Biology, 3(2):247–257, 2007.

[23] Elias Mueggler, Basil Huber, and Davide Scaramuzza. Event-based, 6-dof pose tracking for high-speed maneuvers. 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 2761–2768, 2014.

[24] Bharath Ramesh and Hong Yang. Boosted kernelized correlation filters for event-based face detection. 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW), pages 155–159, 2020.[25]

[26] Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. Proceedings of the IEEE, 102(10):1470–1484, 2014.

[27] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. Eventsto-video: Bringing modern computer vision to event cameras. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–23, 2019.

[28] Sweety Reddy, Silky Goel, and Rahul Nijhawan. Real-time face mask detection using machine learning/ deep feature-based classifiers for face mask recognition. 2021 IEEE Bombay Section Signature Conference (IBSSC), pages 1–6, 2021.

[29] Cian Ryan, Brian O'Sullivan, Amr Elrasad, Aisling Cahill, Joe Lemley, Paul Kielty, Christoph Posch, and Etienne Perot. Real-time face amp; eye tracking and blink detection using event cameras. Neural Networks, 141:87–97, 2021.

[30] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1–10, 2018.

[31] Lea Steffen, Daniel Reichard, Jakob Weinland, Jacques Kaiser, Arne Roennau, and Rüdiger Dillmann. Neuromorphic stereo vision: A survey of bio-inspired sensors and algorithms. Frontiers in Neurorobotics, 13:1–10, 2019.