Neuromorphic Event-based Facial Expression Recognition

Lorenzo Berlincioni*[1] Luca Cultrera*[1] Chiara Albisani[1] Lisa Cresti[1] Andrea Leonardo[1]
Sara Picchioni[1] Federico Becattini[2] Alberto Del Bimbo[1]
[1]University of Florence, MICC {name.surname}@unifi.it
[2]University of Siena {name.surname}@unisi.it

Abstract

Recently, event cameras have shown large applicability in several computer vision fields especially concerning tasks that require high temporal resolution. In this work, we investigate the usage of such kind of data for emotion recognition by presenting NEFER, a dataset for Neuromorphic Event-based Facial Expression Recognition. NEFER is composed of paired RGB and event videos representing human faces labeled with the respective emotions and also annotated with face bounding boxes and facial landmarks. We detail the data acquisition process as well as providing a baseline method for RGB and event data. The collected data captures subtle micro-expressions, which are hard to spot with RGB data, yet emerge in the event domain. We report a double recognition accuracy for the event-based approach, proving the effectiveness of a neuromorphic approach for analyzing fast and hardly detectable expressions and the emotions they conceal.

1. Introduction

Facial expression recognition is important for a large variety of applications [3, 20, 34]. Different kinds of sensors have been used to analyze faces such as depth cameras [9] or sensors with high framerate such as high-speed structured light sensors [67] and extremely fast RGB cameras [43]. In particular, the necessity for elevated framerates stems from the fact that emotions are often conveyed by micro-expressions, which can manifest in short timespans up to 1/25 of a second [12]. Recently, an exploratory approach has studied the capability of neuromorphic sensors, i.e. event cameras, to capture facial expressions [2]. It suggested better recognition rates for event-based approaches compared to RGB.

Event cameras are bio-inspired sensors that, instead of generating streams of synchronous frames, produce asynchronous events for single pixels where illumination changes occur. An advantage is the extremely high rate of events, with temporal resolutions that reach the microsecond. However, due to a lack of data, emotion recognition through event-based videos is still a problem not widely addressed in the literature. In order to cope with the aforementioned lack of data, several attempts have been made to generate synthetic event-based datasets [23, 28, 46]. The authors of [2] framed the recognition setting as a facial reaction recognition system, aiming at understanding whether an expression is positive or negative when using an interactive recommendation system. The authors however, collected a dataset using a VGA event camera, thus with limited spatial resolution. We believe that this poses a strong limit for facial-expression analysis applications since micro-expressions can be very localized in space as much as in time. Reactions are also labeled just as positive, neutral and negative, without providing details about emotions. A few additional works have addressed similar problems, yet

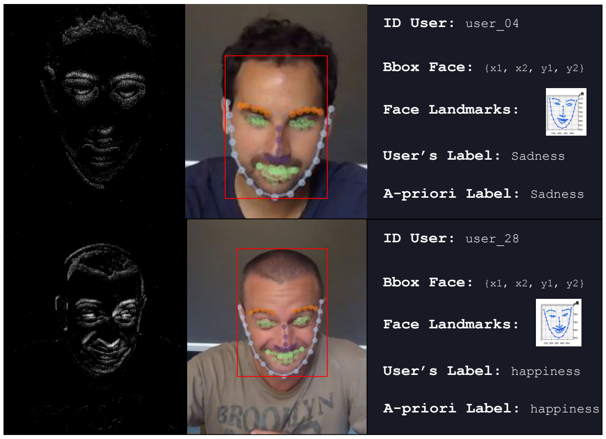

Figure 1. NEFER is a dataset for Neuromorphic Event-based Facial Expression Recognition. We collect paired Event streams and RGB videos, providing for both modalities face bounding boxes, facial landmarks and emotion labels. Emotion labels are provided in two versions: using an a-priori assignment based on the visual stimulus shown to the user and based on actual user feelings.

*Equal contribution

focusing only on face detection and tracking alone, without analyzing expressions or emotions [33, 51].

In our work we present NEFER (Neuromorphic Event-based Facial Expression Recognition) [1] , the first release of an RGB and event dataset for emotion recognition. The dataset is fully labelled with bounding boxes from face detection and facial landmark.

As traditional annotation methods are not ideal for event-based data, we chose a hybrid approach. This involved using the ESIM [46] simulator to obtain aligned RGB images and event streams, which we could then analyze using supervision signals obtained directly from RGB vision methods for face detection and face landmark estimation. We also provide a simple baseline to underline the difficulty of the task and the capabilities of an event-based model with reference to an RGB counterpart. To the best of our knowledge, we are the first to publicly release an event camera facial expression recognition dataset. To summarize, the main contributions of this paper are:

- We propose a dataset for emotion recognition recorded with an high resolution event camera. The dataset consists of more than 600 RGB and events-based videos from more than 30 individuals of different genders and ages.
- We provide labels for face and landmark detection in both RGB and event data.
- For each sample in the dataset we provide two different kind of labels: a-priori labels (pre-defined emotion assignment) and user labels (emotion felt by the user).
- We provide a baseline model to foster future research in the field, underlying the potential of event-based analysis compared to standard RGB approaches.

2. Related Works

The event camera is a neuromorphic sensor that is based on a novel bio-inspired vision paradigm [11, 44]. In contrast to traditional vision systems, it does not produce a synchronous sequence of frames, but instead generates an asynchronous stream of events. Each event is characterized by a local change in brightness and can occur at very short time intervals (in the order of microseconds) with very low latency [36]. Moreover, unlike traditional vision systems, the event camera does not produce any output if there is no change in brightness, thereby conserving resources. To summarize, utilizing a neuromorphic sensor results in reduced motion blur, high temporal resolution, and high dynamic range (up to 140 dB). Additionally, it enables a reduction in bandwidth consumption [15, 16].

Despite the fact that event cameras have not been on the market for an extended period, and their large-scale use is still somewhat limited, there are examples of their application in fields such as robotics and computer vision that can be found in the literature [11, 16]. In fact, in these contexts, the benefits offered by event cameras can be fully leveraged. In [45] the authors propose an event-based descriptor for event camera data and show its results in some vision problems such as object classification, tracking, detection and feature matching. Also [30, 35, 40] propose event-based approaches for object detection and recognition. Event cameras are widely used in literature for tracking [48, 53, 70]. In [69] they propose a trasformer-based architecture to fuse temporal and spatial information encoded in the events for single object tracking. Neuromorphic sensors are extensively utilized in surveillance [37, 49, 56] due to their distinctive characteristics and low power consumption. In fact, one of the most desirable properties of these sensors in surveillance is their ability to transmit information solely when changes occur. [6] propose a neuromorphic vision-based system for autonomous vehicles. Event cameras are also used in a wide range of scenarios in robotics and computer vision such as video super-resolution [22, 27], depth and optical flow prediction [17], monocular and stereo depth estimation [19, 59], SLAM [26] and visual odometry [38, 62, 72], and human pose estimation [7, 52].

Of particular interest for this work, is the fact that neuromorphic sensors have a compelling application scenario in face detection and emotion recognition. The distinct characteristics and properties of event cameras enable them to capture even the subtlest variations and microexpressions in human emotions at remarkably high temporal resolution and with minimal latency. Nevertheless, this aspect has not received widespread attention in the literature. In general, facial images possess crucial features that can serve several biometric applications [39] and therefore face and landmark detection through deep learning algorithms is a problem widely addressed in the literature [57, 64, 65]. Training deep learning models to perform well in face and landmark detection tasks requires a large amount of data: [68] proposes a dataset composed by 0.5M images from 10,575 individuals, [21] use more than 100k individuals to generate approximately 10M of images, and [29] includes 1M images from more the 690k individuals. The usage of synthetic face images has also been explored [18], whereas [63] instead proposed a dataset for masked face recognition, as a response to safety mandates during the covid pandemic. Several other datasets have also been published addressing the study of faces [5, 24, 61].

As for the event-based domain, despite all this interest in the topic, not many datasets can be found in the literature. In fact in [50], to make up for the lack of data, they

[1] The dataset is available here: https://github.com/miccunifi/NEFER

Dataset	Videos	Users	Resolution	Bounding Boxes	Landmarks	Emotions
Savran *et al.* [51]	108	**30**	304 × 204	✗	✗	✗
Lenz *et al.* [33]	48	10	640 × 480	✗	✗	✗
Becattini *et al.* [2]	455	25	640 × 480	✗	✗	✗
NEFER	**609**	29	**1280 × 720**	✓	✓	✓

Table 1. Comparison of event-based face datasets. To the best of our knowledge, NEFER is the first dataset to provide bounding boxes, facial landmarks and emotion labels, as well as an HD resolution. NEFER is also the larger face dataset up to date.

use a synthetic event-based dataset starting from [32]. Several attempts have been made in the literature to generate simulated data for event cameras. In [46] the authors propose ESIM, an event camera simulator with the ability to accurately and efficiently simulate events, while also offering the flexibility to simulate any camera trajectory within a 3D scene of any nature. [28] extends [46] with the goal to reduce the gap between simulation and real sensors by directly mapping noise distributions from real pixels. [23] instead, proposes a tool for generating synthetic event data and demonstrates its effectiveness in two computer vision object recognition and detection tasks. Also [71] and [42] propose event camera simulators. On the one hand, [71] introduces a multiple event simulator method suitable to be used in real-time robotics applications. [42], on the other hand, suggest an approach to emulate the behavior of an attention-based camera sensor. In conclusion, to the best of our knowledge, only three datasets containing facial images captured using a real event camera are present in the literature [2, 33, 51]. In [51] the problem of face pose alignment is analyzed. The authors provide a dataset consisting of 108 videos of extreme head rotations with varying motion intensity, totaling just over 10 minutes of frames acquired. In [33] on the other hand, the authors collected data with an event camera for eye blink detection. The dataset consists of 48 videos (total duration of about 13 minutes). The authors of [2] instead collected a dataset of 455 videos of facial reactions where the recorded users react to garment images. Reactions are classified in three classes: positive, neutral and negative. However, both [51], [33] and [2] use low resolution event cameras with resolution of 304×240px or 640 × 480px. We are of the opinion that this presents a significant constraint for facial expression analysis applications as micro-expressions can be highly localized both spatially and temporally.

In this paper, we introduce NEFER (Neuromorphic Event-based Facial Expression Recognition), a dataset composed of paired RGB and event data for emotion recognition. We collected the dataset with high-resolution RGB and event cameras, providing also facial bounding box and landmark annotations in addition to emotion labels following the frequently used Ekman's emotion classification [13]. As far as our knowledge extends we are the first to publicly release an event camera-based facial expression recognition dataset. A comparison with existing datasets is presented in Tab. 1.

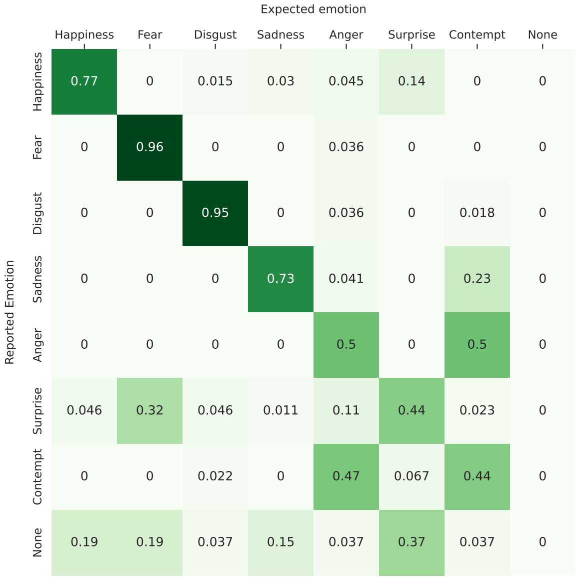

Figure 2. Confusion matrix of the two sets of labels from the NEFER dataset. The expected emotions are a-priori labels assigned based on the content of the visual stimulus shown to the users. The reported emotions instead are emotions declared by the users after observing the videos.

3. NEFER: Neuromorphic Event-based Facial Expression Recognition

The purpose of NEFER is to capture genuine micro-expressions associated to specific emotions with both an event camera and a standard RGB camera. We considered the 7 primary emotions defined by Ekman [13], namely *Disgust*, *Contempt*, *Happiness*, *Fear*, *Anger*, *Surprise* and *Sadness*, since these have been identified as independent from culture, history and personality and are performed in a similar way by everyone.

Emotion	Video Descriptions		
Disgust	Spyder in a man's mouth	Crushed Pimple on Cheek	Man Eating a Larva
Contempt	Cops Killing Protestant	Dog Being Abandoned	Dog Being Mugged
Happiness	Dogs playing	Laughing Child	Old Man Dancing with Boys
Fear	Suddenly Appearing Ghost	Hidden Clown Attacking Camera	Giant Snake Attacking Camera
Anger	Man Attacking Companion	Boy destroying Brother's PC	Professor Assaulted by Student's Parents
Surprise	Baseball Coming Towards Camera	Girl with Unexpected Makeup	Presentation Concluding with a Cat
Sadness	Death of Mufasa in the Lion King	Death of Elly in Up	Boy who has to Undergo an Operation

Table 2. Videos shown to participants and relative emotion label. Emotions follow the 7 basic emotions defined by Ekman [13].

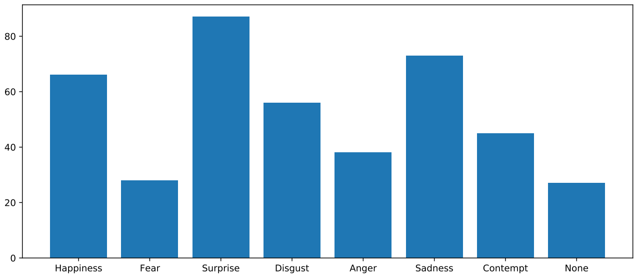

Figure 3. Class distribution with user labels.

3.1. Setting and protocol

In order to obtain realistic and non-simulated expressions, we asked a set of volunteers to maintain a neutral facial expression while watching a selection of videos. A reward has been offered to the participants to encourage a proper behavior during the test (high-stakes situation). The volunteers that took part in the creation of NEFER are both males and females of age ranging between 24 and 52 years, for a total of 29 users.

We showed to each user 21 different videos, 3 for each of Ekman's basic emotions. The videos have been selected from online streaming platforms (e.g. YouTube). Each video was trimmed to the same length of 7s to keep the recording sessions as short as possible so not to induce unwanted expressions due to, for instance, boredom. This choice also simplifies training schemes with deep learning frameworks which process data in mini-batches of the same size. Tab. 2 lists the videos that have been shown to the users with the correspondent emotion label. The overall procedure for the data acquisition and video selection was inspired by previously collected dataset from the state of the art [10, 66].

For the recording we used two capturing devices: a GOPRO Hero+ action camera, recording videos at 60FPS and 1920 × 1080px resolution, and a Prophesee Evaluation Kit HD, recording event videos at a resolution of 1280 × 720px. The cameras have been mounted on a fixed recording rig in a room lit with natural light. We specifically avoided any presence of artificial light to avoid background noise that could alter the event-based recordings. Users are also isolated from other people which could generate distractions.

Users have been asked to sit in front of the screen at approximately 60cm from the cameras. The RGB and event streams have been programmatically synchronized in order to capture two videos of the same duration and content. After viewing each video, we asked the volunteers to provide a personal evaluation of the observed footage. In particular, we asked two questions: (i) select among the 7 basic emotions, plus a *"None"* option, the most suitable one to describe the emotions stemmed from viewing the video; (ii) the intensity, on a 1 to 5 scale, of such emotion. We used the collected answers to create two alternative versions of the annotations, one considering the labeling of the user and one following our a-priori video-emotion assignment. In Fig. 2 a confusion matrix is presented showing the differences between the two label versions. The two versions mostly differ in the fact that following user labelings we have the additional neutral emotion and a slight unbalance in the sample distribution as shown in Fig. 3. Overall, recording sessions lasted 18 minutes on average. Fig. 4 shows a few samples from the dataset.

4. Video Annotation Through Simulated Events

The wide range of off-the-shelf functionalities for RGB-based computation is not available for event-based data. This includes modules that nowadays are common building blocks in computer vision pipelines such as face detectors and landmark estimators. In addition, it is necessary to preprocess the raw data of the neuromorphic sensor in order to use it with frame-based computational tools. Bridging this gap is not trivial, since due to the asynchronous nature of the domain, the usual annotation process for many different tasks becomes cumbersome and expensive. Even generating relatively simple annotations such as facial bounding boxes, which are reliably obtainable with RGB data, would require lots of manual annotation.

To provide additional annotations for event-based data we exploit RGB data and an event camera simulator,

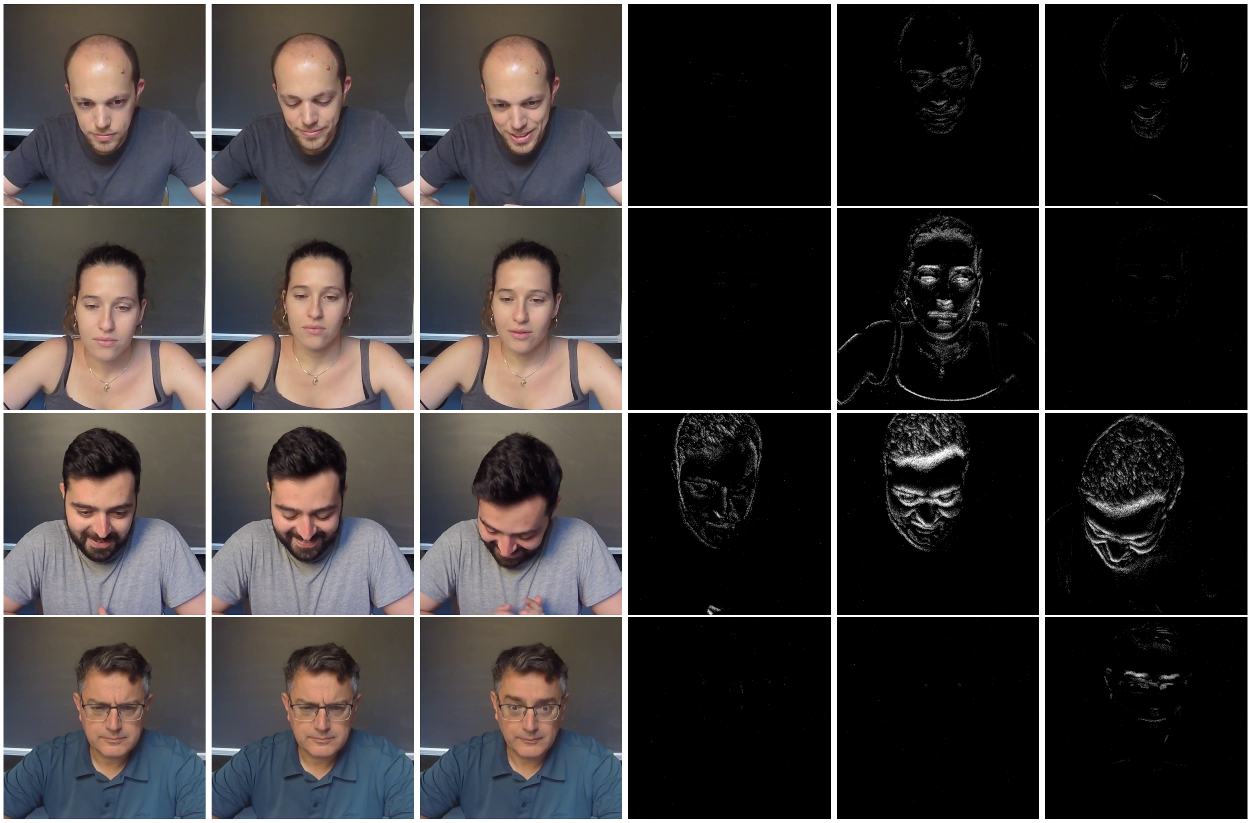

Figure 4. Four samples from the NEFER dataset. First row: happiness; Second row: fear; third row: disgust; fourth row: surprise. Subtle movements are almost invisible with RGB but are emphasized in event frames.

ESIM [46]. Through the use of the ESIM simulator we convert the RGB videos into physically accurate simulated event streams. We then run a face detector and facial landmark estimator on the RGB frames, which is easily done with tools such as FaceAlignment [4]. We train a face detector (Yolov2) [47] and a landmark estimator [4] on simulated data and test it on real event streams. This approach provides satisfactory results on most frames, decimating the annotation time. The final annotations are manually refined and validated using CVAT [60].

4.1. ESIM

ESIM [46] is an event-based camera simulator that can generate a synthetic event-based stream from its RGB video counterpart in a physically realistic way. The images are rendered by the simulator at a high frame rate, interpolating pixel brightness along the camera trajectory using an adaptive sampling technique, which is adapting the frame rate based on a prediction of the previous signals. We feed to the simulator all the RGB frames to generate a synthetic event-based version of each stream. In this way, we are able to associate the bounding boxes provided by face alignment on RGB frames with event data. The simulator-generated outputs are encoded using an exponential time surface [31]. Note the synthetic event-based videos obtained from the RGB data are used only as a mean for training models to quickly collect annotations. These are not pixel-wise aligned with the real event streams and we do not treat them as part of the final dataset, which only comprises real event data.

4.2. Face Detection

Using the synthetic data from the simulator, we generated an annotated dataset in the event spectrum to train a face detector. First, we generated face annotation for RGB frames using FaceAlignment [4], an open-source tool for face analysis[2]. We then bound the face labels with the corresponding synthetic event frames obtained with ESIM. This allowed us to train a YOLOv2 [47] on the synthetic version of NEFER. We found the detector to have good generalization capabilities from synthetic to real event data, which yielded high-quality annotations at a slight cost of manual validation using CVAT [60].

[2] https://github.com/1adrianb/face-alignment

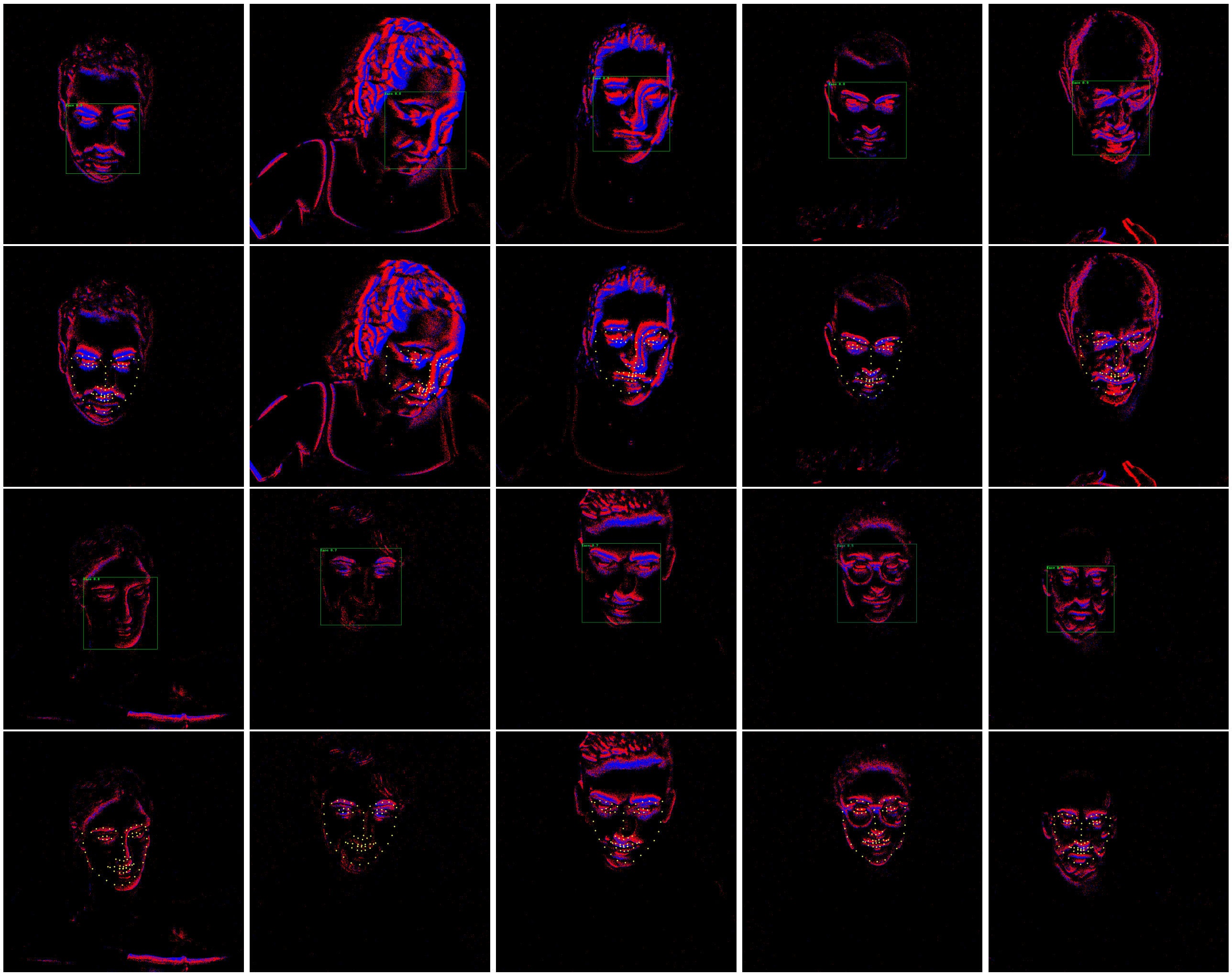

Figure 5. Examples of detected faces and estimated landmarks on real event videos of NEFER. Better viewed in color on a PC screen. Bounding boxes are shown in green, landmarks are shown in yellow.

4.3. Landmark Detection

The facial landmark detection is performed by an Xception [8] architecture trained on the synthetic data from ESIM to regress the position of 68 landmarks of the face. Similarly to face detection, we obtained the ground truth labels from the RGB videos by using FaceAlignment [4]. The Xception architecture is composed of three stages, all of them employing depthwise separable convolutions along skip connections, resulting in a faster convergence training [8]. The final linear layer outputs the 136 normalized numbers representing the coordinates of the standard 68 facial landmarks. The model is optimized using Adam with a learning rate of 8×10^{-4} for 10 epochs over 30K frame samples with the use of standard augmentation techniques (random changes in brightness, contrast, rotation, translation, and crop).

5. Baseline Method

We provide a simple baseline for the dataset. This baseline architecture is based on a 3D convolutional network C3D [58]. It has been chosen as it has been a long-standing, simple, standard approach for video-based action and activity recognition tasks [1, 14, 41, 58]. The C3D model is implemented using 5 3D convolutional blocks, all with kernel size 3 and padding 1, followed by a 3D max-pooling of size 2 and stride 2. This chain of sequential blocks reduces the input stacked sequence of images down to a 72 channels feature map, which is then flattened and fed to two fully connected layers of size 512 and 64 before a final classification layer. ReLU activations are present between all layers. The model architecture is depicted in Fig. 6.

We train the same model separately with RGB-frame-based data and with event data obtained by converting

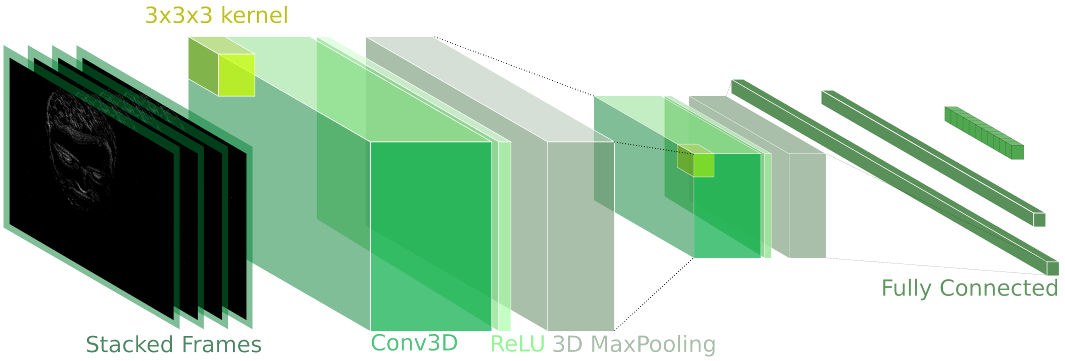

Figure 6. Illustration of our C3D model. The stacked frames form the input to the first of 5 Conv3D + ReLU + 3DMaxPooling blocks. Finally the 3D feature maps are flattened and fed to the final two fully connected layers

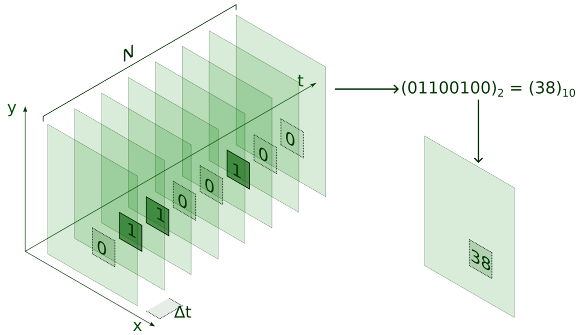

Figure 7. Visual diagram illustrating the TBR encoding aggregating multiple events in a frame.

events into frame-wise representations using Temporal Binary Representation (TBR) [25] (see Sec. 5.1). We detect the face using our pre-trained detector (see Sec. 4.2), and resize the bounding box to a 200 × 200px patch before feeding it as input to the model.

5.1. Temporal Binary Representation

Temporal Binary Representation [25] (TBR) is an aggregation strategy to map the asynchronous events into a stream of synchronous frames that can be then processed by a standard computer vision pipeline. Given a fixed Δt we can build the binary representation b^i of a pixel at (x, y) by checking for an event in such a time interval, $b^i_{x,y} = \mathbb{1}(x, y)$.

We can then collect N consecutive representations and stack them together as $B \in \mathbb{R}^{H \times W \times N}$ forming for each pixel a binary string $[b^0_{x,y}, b^1_{x,y}, ..., b^N_{x,y}]$, as shown in Fig.7. This approach manages to create a frame processable by traditional Computer Vision algorithms with a minimal memory footprint and by retaining temporal information within the value of each pixel.

For our experiments, we used this representation setting $\Delta t = 15$ milliseconds and $N = 8$.

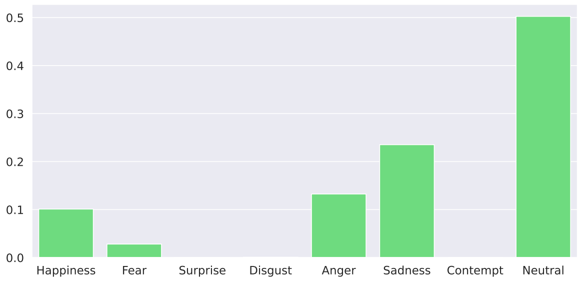

Figure 8. Distribution of predicted labels on frames of the NEFER validation set using Deep Face [54]. Almost 50% of the frames are predicted as neutral, whereas *Surprise*, *Disgust* and *Contempt* are predicted for less than 0.1% of the frames.

6. Experimental Results

We implemented our C3D model using PyTorch and trained it using the Adam optimizer initialized at the default learning rate value of 1×10^{-4} which is then reduced following the scheduling technique presented in [55] with the annealing strategy. As loss, we adopt the Binary Cross-Entropy Loss, regularized with weight decay.

We compare the performances of our model by training it separately first on the RGB videos and then on the event streams, using both the self-reported user annotations and the a-priori expected one as labels for the target emotion. We define a validation split by selecting 20% of the users at random (thus keeping each user either in the training set or in the validation set to avoid unwanted biases), for a total of 126 videos.

We found that the RGB model results in poor accuracy, obtaining an average of 14.37% using the user labels and 14.60% using the expected ones. The event-based model

Data	A-Priori Labels	%	Reported Labels	%
RGB	14.60	-	14.37	-
TBR Event	22.95	+57.2%	**30.95**	+115.4%

Table 3. Absolute accuracy and relative performances of our baseline model over the different data domains and using both labelling versions of NEFER.

instead showed much better performances, reaching an accuracy of 22.95% with the user labels and of 30.95% using the expected ones. We report these experimental results in Tab. 3. This confirms that neuromorphic cameras are well suited for analyzing faces and that event footage carries valuable information for identifying subtle micro-expressions that are not easily detectable with RGB data.

Interestingly, we observed that our baseline model, just as the human a-priori assumptions, tends to confuse classes that share similar expressions, such as *fear* with *surprise* or *anger* with *contempt* even when trained on the self-reported emotions.

Finally, we perform a control experiment by running a frame-based pre-trained state of the art emotion recognition framework on the RGB data. As a model we adopt Deepface [54], a recent facial attribute analysis framework. The model uses the same categories as we do, following Ekman's emotion classification, with the only exception of the *Contempt* category, which is missing in Deepface. As shown in Fig.8, we note its tendency towards classifying most of the frames with the neutral class *None*. This underlines the difficulty of the task in the setting that we propose: most frames do not carry a very polarized expression and most emotion cues happen very quickly, in a way that it is difficult to grasp them with RGB cameras. We argue that to fully comprehend the underlying emotions of humans from a vision-based point of view, event cameras will play an important role in the near future due to their ability to capture fine-grained micro-expressions and micro-movements of the face.

7. Conclusions and Future Work

In this paper, we presented a first release of NEFER, a dataset for expression recognition based on event camera data. This dataset is composed of paired visual spectrum images and event camera streams. For every sequence of frames, both the expected emotion and the self reported one by the user are given. Every frame has multiple annotations, namely the user face bounding box and the respective facial landmarks that we collected by leveraging models trained on synthetic data obtained using a simulator. Finally, we presented and discussed a 3D convolutional baseline, trained on both version of our dataset, which achieved improved results on event camera data with respect to the RGB frame based data.

We consider this a starting point for a future larger collection of data in the event camera domain for similar high-time resolution tasks. The large interest given by the computer vision community towards understanding facial expressions and emotions proves the importance of the task, yet the neuromorphic community and the traditional RGB vision one still have several gaps to be bridged. We believe that pursuing this line of research will bring attention to an emerging field, bringing together the best of both worlds and providing multiple modalities to approach problems that, based on experimental results, appear to be better addressed in the event domain rather than in the RGB domain alone.

Acknowledgements

This work was supported by the European Commission under European Horizon 2020 Programme, grant number 951911—AI4Media.

References

[1] T-c3d: Temporal convolutional 3d network for real-time action recognition. 32. 6

[2] Federico Becattini, Federico Palai, and Alberto Del Bimbo. Understanding human reactions looking at facial micro-expressions with an event camera. *IEEE Transactions on Industrial Informatics*, 2022. 1, 3

[3] Federico Becattini, Xuemeng Song, Claudio Baecchi, Shi-Ting Fang, Claudio Ferrari, Liqiang Nie, and Alberto Del Bimbo. Plm-ipe: A pixel-landmark mutual enhanced framework for implicit preference estimation. In *ACM Multimedia Asia*, pages 1–5. 2021. 1

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 5, 6

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2

[6] Guang Chen, Fa Wang, Weijun Li, Lin Hong, Jörg Conradt, Jieneng Chen, Zhenyan Zhang, Yiwen Lu, and Alois Knoll. Neuroiv: Neuromorphic vision meets intelligent vehicle towards safe driving with a new database and baseline evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):1171–1183, 2022. 2

[7] Jiaan Chen, Hao Shi, Yaozu Ye, Kailun Yang, Lei Sun, and Kaiwei Wang. Efficient human pose estimation via 3d event point cloud. *arXiv preprint arXiv:2206.04511*, 2022. 2

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 6

[9] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal,

and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016. 1

[10] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing*, 9(1):116–129, 2016. 4

[11] Tobi Delbruckl. Neuromorophic vision sensing and processing. In *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, pages 7–14, 2016. 2

[12] Paul Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company, 2009. 1

[13] P Ekmann. Universal facial expressions in emotion. *Studia Psychologica*, 15(2):140, 1973. 3, 4

[14] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 445–450, New York, NY, USA, 2016. Association for Computing Machinery. 6

[15] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, Hirotsugu Takahashi, Hayato Wakabayashi, Yusuke Oike, and Christoph Posch. 5.10 a 1280×720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86µm pixels, 1.066geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, pages 112–114, 2020. 2

[16] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 2

[17] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12280–12289, 2019. 2

[18] Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–234, 2018. 2

[19] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021. 2

[20] Hajer Guerdelli, Claudio Ferrari, Walid Barhoumi, Haythem Ghazouani, and Stefano Berretti. Macro-and micro-expressions facial datasets: A survey. *Sensors*, 22(4):1524, 2022. 1

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 2

[22] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4882–4891, 2021. 2

[23] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. v2e: From video frames to realistic dvs events. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1312–1321, 2021. 1, 3

[24] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 2

[25] Simone Undri Innocenti, Federico Becattini, Federico Pernici, and Alberto Del Bimbo. Temporal binary representation for event-based action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10426–10432. IEEE, 2021. 7

[26] Jianhao Jiao, Huaiyang Huang, Liang Li, Zhijian He, Yilong Zhu, and Ming Liu. Comparing representations in tracking for event camera-based slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1369–1376, 2021. 2

[27] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Turning frequency to resolution: Video super-resolution via event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7772–7781, 2021. 2

[28] Damien Joubert, Alexandre Marcireau, Nicholas O. Ralph, A. Jolley, André van Schaik, and Gregory Cohen. Event camera simulator improvements via characterized parameters. *Frontiers in Neuroscience*, 15, 2021. 1, 3

[29] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016. 2

[30] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2146–2156, 2021.

[31] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad B. Benosman. Hots: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1346–1359, 2017. 5

[32] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, pages 679–692. Springer, 2012. 3

[33] Gregor Lenz, Sio-Hoi Ieng, and Ryad Benosman. Event-based face detection and tracking using the dynamics of eye blinks. *Frontiers in Neuroscience*, 14:587, 2020. 2, 3

[34] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020. 1

[35] Yijin Li, Han Zhou, Bangbang Yang, Ye Zhang, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Graph-based asynchronous event processing for rapid object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 934–943, 2021. 2

[36] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 2

[37] M. Litzenberger, B. Kohn, A.N. Belbachir, N. Donath, G. Gritsch, H. Garn, C. Posch, and S. Schraml. Estimation of vehicle speed based on asynchronous data from a silicon retina optical sensor. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 653–658, 2006.

[38] Daqi Liu, Alvaro Parra, and Tat-Jun Chin. Spatiotemporal registration for event-based visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2021. 2

[39] Mostafa Mehdipour Ghazi and Hazim Kemal Ekenel. A comprehensive analysis of deep learning based representation for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2016. 2

[40] Anindya Mondal, Jhony H Giraldo, Thierry Bouwmans, Ananda S Chowdhury, et al. Moving object detection for event-based vision using graph spectral clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 876–884, 2021. 2

[41] Alberto Montes, Amaia Salvador, Santiago Pascual, and Xavier Giro-i Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. In *1st NIPS Workshop on Large Scale Computer Vision Systems*, December 2016. 6

[42] Md Jubaer Hossain Pantho, Joel Mandebi Mbongue, Pankaj Bhowmik, and Christophe Bobda. Event camera simulator design for modeling attention-based inference architectures. *Journal of Real-Time Image Processing*, pages 1–12, 2021. 3

[43] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. 2009. 1

[44] 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. 2

[45] Bharath Ramesh, Hong Yang, Garrick Orchard, Ngoc Anh Le Thi, Shihao Zhang, and Cheng Xiang. Dart: distribution aware retinal transform for event-based cameras. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2767–2780, 2019. 2

[46] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. *Conf. on Robotics Learning (CoRL)*, Oct. 2018. 1, 2, 3, 5

[47] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 5

[48] Alpha Renner, Matthew Evanusa, Garrick Orchard, and Yulia Sandamirskaya. Event-based attention and tracking on neuromorphic hardware. In *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 132–132, 2020. 2

[49] J.P. Rodríguez-Gomez, A. Gómez Eguíluz, J.R. Martínez-de Dios, and A. Ollero. Asynchronous event-based clustering and tracking for intrusion monitoring in uas. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8518–8524, 2020. 2

[50] Cian Ryan, Brian O'Sullivan, Amr Elrasad, Aisling Cahill, Joe Lemley, Paul Kielty, Christoph Posch, and Etienne Perot. Real-time face & eye tracking and blink detection using event cameras. *Neural Networks*, 141:87–97, 2021. 2

[51] Arman Savran and Chiara Bartolozzi. Face pose alignment with event cameras. *Sensors*, 20(24), 2020. 2, 3

[52] Gianluca Scarpellini, Pietro Morerio, and Alessio Del Bue. Lifting monocular events to 3d human poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1358–1368, 2021. 2

[53] Hochang Seok and Jongwoo Lim. Robust feature tracking in dvs event stream using bezier mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[54] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 7, 8

[55] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019. 7

[56] Tanin Sultana and Khan A Wahid. Iot-guard: Event-driven fog-based video surveillance system for real-time security management. *IEEE Access*, 7:134881–134894, 2019. 2

[57] Murat Taskiran, Nihan Kahraman, and Cigdem Eroglu Erdem. Face recognition: Past, present and future (a review). *Digital Signal Processing*, 106:102809, 2020. 2

[58] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 6

[59] SM Nadim Uddin, Soikat Hasan Ahmed, and Yong Ju Jung. Unsupervised deep event stereo for depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7489–7504, 2022. 2

[60] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation: A set of

best practices for high quality, economical video labeling. *International journal of computer vision*, 101:184–204, 2013. 5

[61] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018. 2

[62] Yifu Wang, Jiaqi Yang, Xin Peng, Peng Wu, Ling Gao, Kun Huang, Jiaben Chen, and Laurent Kneip. Visual odometry with an event camera using continuous ray warping and volumetric contrast maximization. *Sensors*, 22(15):5687, 2022. 2

[63] Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, et al. Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093*, 2020. 2

[64] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 499–515. Springer, 2016. 2

[65] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127:115–142, 2019. 2

[66] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–7. IEEE, 2013. 4

[67] Yuping Ye, Zhan Song, and Juan Zhao. Facial micro-expression analysis via a high speed structured light sensing system. *Journal of Image and Graphics*, 9(1):15–19, 2021. 1

[68] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2

[69] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8801–8810, 2022. 2

[70] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13043–13052, 2021. 2

[71] Andreas Ziegler, Daniel Teigland, Jonas Tebbe, Thomas Gossard, and Andreas Zell. Real-time event simulation with frame-based cameras, 2022. 3

[72] Yi-Fan Zuo, Jiaqi Yang, Jiaben Chen, Xia Wang, Yifu Wang, and Laurent Kneip. Devo: Depth-event camera visual odometry in challenging conditions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2179–2185. IEEE, 2022. 2