Pattern Recognition Letters 187 (2025) 42–48

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Neuromorphic face analysis: A survey

Federico Becattini [a], Lorenzo Berlincioni [b,*], Luca Cultrera [b], Alberto Del Bimbo [b]

[a] University of Siena, Siena, Italy
[b] University of Florence, Florence, Italy

ARTICLE INFO

Editor: George Azzopardi

MSC:
41A05
41A10
65D05
65D17

Keywords:
Computer Vision
Video Understanding
Neuromorphic Camera
Biometrics

ABSTRACT

Neuromorphic sensors, also known as event cameras, are a class of imaging devices mimicking the function of biological visual systems. Unlike traditional frame-based cameras, which capture fixed images at discrete intervals, neuromorphic sensors continuously generate events that represent changes in light intensity or motion in the visual field with high temporal resolution and low latency. These properties have proven to be interesting in modeling human faces, both from an effectiveness and a privacy-preserving point of view. Neuromorphic face analysis however is still a raw and unstructured field of research, with several attempts at addressing different tasks with no clear standard or benchmark. This survey paper presents a comprehensive overview of capabilities, challenges and emerging applications in the domain of neuromorphic face analysis, to outline promising directions and open issues. After discussing the fundamental working principles of neuromorphic vision and presenting an in-depth overview of the related research, we explore the current state of available data, data representations, emerging challenges, and limitations that require further investigation. This paper aims to highlight the recent progress in this evolving field to provide researchers an all-encompassing analysis of the state of the art along with its problems and shortcomings.

1. Introduction

Face analysis for decades has been one of the most studied topics in computer vision. Some tasks involving faces can even be considered to be solved, as they can be performed effectively in unconstrained scenarios with off-the-shelf tools. A wide plethora of methods falls under the umbrella of face analysis: landmark detection, age estimation, lip reading, eye tracking, 3D reconstruction, just to name a few. An interesting application of face analysis is the theoretical possibility of estimating human emotions and feelings just by observing facial micro-expressions [1]. In fact, micro-movements of the face, have been mapped directly into emotions and it is known from psychology studies that such movements can be involuntary and almost impossible to hide [2]. To this day, several computer vision applications that strive to estimate emotions by analyzing faces have been proposed, yet such micro-movements can happen at an extremely fast rate [3], that is likely not going to be fully observable with a traditional RGB camera. On the other hand, faces are sensitive biometric data. Analyzing faces has thus raised privacy-related concerns, that have also been addressed in the recent AI Act by the European Commission.[1]

These issues have recently led to an increasing interest in the usage of neuromorphic cameras, as they have shown promising results from different points of view, including effectiveness, latency, power consumption and privacy-preservation. Unlike conventional cameras that capture entire frames at fixed intervals, neuromorphic cameras operate on a fundamentally different principle, mimicking the asynchronous and event-driven nature of biological vision. Events are generated only when pixel-level changes in luminance exceed a predefined threshold. This approach enables the efficient use of computational resources, as only relevant information is transmitted and processed. The absence of a fixed frame rate means that these cameras can capture and process events with microsecond precision, a capability that is especially advantageous in dynamic and fast-paced environments. This novel paradigm has the potential to open up new approaches for various applications, ranging from robotics and autonomous vehicles to surveillance.

In this paper, we propose an overview of the relatively recent field of research involving event cameras and face analysis, which we dub Neuromorphic Face Analysis. Our goal is to provide a compendium for computer vision researchers in the field of face analysis, discussing opportunities and challenges, as well as taking stock of what is possible with event cameras up to this day. Not many works exist on the subject, yet they address several topics of primary importance and we

* Corresponding author.
E-mail addresses: federico.becattini@unisi.it (F. Becattini), lorenzo.berlincioni@unifi.it (L. Berlincioni), luca.cultrera@unifi.it (L. Cultrera), alberto.delbimbo@unifi.it (A. Del Bimbo).

[1] https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf.

https://doi.org/10.1016/j.patrec.2024.11.009
Received 23 April 2024; Received in revised form 3 September 2024; Accepted 9 November 2024
Available online 19 November 2024

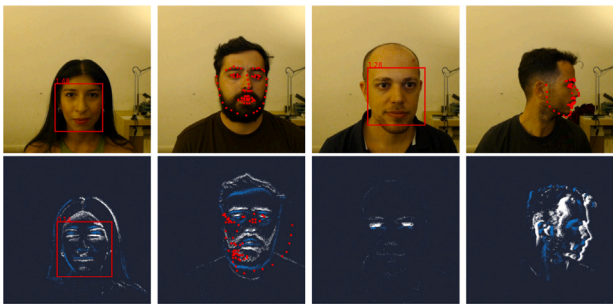

Fig. 1. Failure cases of computer vision models over event data. *Top Row*: Samples of face detection and landmark estimation on RGB frames. *Bottom Row*: Samples of face detection and landmark estimation on the corresponding Event frames.

firmly believe that neuromorphic face analysis will quickly raise the bar compared to traditional frame-based face analysis.

2. Neuromorphic face analysis: Advantages and challenges

Event cameras exhibit unparalleled advantages in face analysis applications, offering a paradigm shift in capturing and interpreting facial dynamics. Their low-latency operation and high temporal resolution ensure real-time responsiveness to facial expressions and lip or eye movements, which are critical for applications like human–computer interaction and security systems. Furthermore, the wide dynamic range of event cameras allows for accurate representation of facial features in challenging lighting conditions, offering improved performance compared to conventional cameras that may struggle with overexposure or underexposure. However, these advantages come with challenges. The asynchronous nature of event data, beneficial for real-time responsiveness, becomes a challenge, as developing algorithms tailored to interpret sporadic events poses a unique computational hurdle. Processing data produced by these sensors requires innovative approaches in computer vision to precisely recognize and understand facial dynamics.

It is crucial to note also that existing methods commonly employed for face analysis, such as face detectors and landmark detectors designed for traditional frame-based cameras, may not be directly suitable for event cameras. In Fig. 1 we show the outputs of a commonly used object detector and landmark detector (respectively DLIB [4] and Face Alignment [5]) trained on RGB frames from traditional cameras. We tested such models on video sequences captured with a paired set of event and RGB cameras. The object detector can properly locate targets only when the face motion is enough to make it visible (odd columns of Fig. 1). However, the outputs have an extremely low confidence due to the domain shift. The landmark detector instead fails to produce meaningful outputs, even if the face is fully visible (even columns of Fig. 1). The asynchronous nature of event data and the lack of continuous frames present a mismatch with the assumptions underlying traditional face analysis methods. Furthermore, the limited availability of standardized datasets specifically designed for training event-based face analysis models is a notable impediment. Even how events are represented can pose a challenge, since models trained with a given representation are likely to lose effectiveness when such representation is changed.

Data Representation An important challenge when dealing with event data is the fact that learning end-to-end models from raw events would require processing a huge amount of information, as millions of events can be generated every second. This poses a significant difference between the neuromorphic and RGB domains. To bridge this gap, researchers often engineer frame-like representations so that event data can be fed to a computer vision model, such as a convolutional neural network.

At the time, there is no clear standard and each encoding strategy must be tuned with specific hyperparameters. One of the most important is the accumulation time Δt, which controls the temporal granularity at which the information is processed. Based on the task, different accumulation times can yield extremely different results. For instance, a face detector would require a sufficiently large Δt in order to capture enough information for the face to be visible; on the contrary, modeling facial micro-expressions requires a sufficiently small accumulation time, as fast movements might be lost among other unrelated events. The most common representation strategies involve quantizing events into spatio-temporal histograms. However, events can be processed also as raw individual events [15]. For this family of approaches minimal to none pre-processing is done. The data keeps its original format or is quantized in extremely small accumulation times to simulate a continuous signal. This strategy often relies on the use of Spiking Neural Networks [14,15]. These raw event-based approaches can be more efficient in terms of computational resources and latency, as they directly process the event stream without transforming it into a different representation. However, they may require specialized architecture and can be less accurate for certain tasks due to noise and sparsity in the data.

Other works have adopted an intermediate strategy, i.e. do not use a frame-based representation yet increase the accumulation interval Δt, obtaining a voxel of quantized events in the form of a 3D tensor [16,25]. This method balances between raw and frame-based approaches, providing a compromise in terms of computational load and accuracy. By grouping events over slightly longer time intervals, these representations can reduce the noise and enhance the signal of interest, offering a suitable trade-off for real-time applications where both speed and precision are crucial. To treat the task as a more standard computer vision problem, it is common to transform the raw data to a 2D grid (an *image*). This allows the use of models originally developed for processing RGB data by, for example, accumulating events in a spatio-temporal neighborhood [23], treating sequences of binary events as binary digits of a decimal number [27,37] or building a motion history image by applying an exponential decay over the neighborhood of active events [38]. Frame-based representations tend to perform well for tasks such as object detection and recognition, where existing computer vision architectures can be directly applied. However, these methods can introduce latency and may not fully exploit the temporal resolution advantage of event cameras, making them less efficient for real-time processing. In conclusion, the choice of data representation significantly impacts the model's performance and efficiency. Raw event processing offers speed but may sacrifice some accuracy and require specialized systems, intermediate 3D voxel representations provide a balanced approach, and frame-based methods leverage conventional models but might lose some of the temporal advantages inherent to event data. Each method's suitability largely depends on the specific application requirements, such as the need for real-time processing, accuracy, and available computational resources.

Event Cameras as a Privacy Preservation Layer A notable advantage of neuromorphic vision is that several works have attributed to the usage of an event camera the benefit of working under the preservation of privacy [39–42]. Unlike traditional RGB cameras, which capture full-resolution images containing detailed visual information that can easily reveal personal identities, event cameras capture only changes in the scene, recording events that correspond to changes in light intensity. This inherent characteristic means that event streams provide an additional layer of privacy, as they are harder to interpret and less likely to contain sensitive information, particularly when both the camera and the scene are static. In such scenarios, no signal is generated, further reducing the risk of privacy invasion. In particular, [39,42] leverage event cameras for action recognition purposes, while protecting the identity of the analyzed individuals, unlike traditional cameras that could potentially expose personal identities. This idea is used by [42] to work in sensitive environments such as high schools. Nonetheless, whereas it is certainly true that intensity information is discarded in event data, which adds a level of privacy, sensitive information might

Table 1
State of the art divided into intermediate modules (columns) and applications (rows).

	Face detection	Eye blink detection	Pupil/Eye detection	Landmark detection	Drowsiness detection	End-to-End
Face detection	–	Lenz et al. [6] Ryan et al. [7]	Ryan et al. [7]	–	–	Bissarinova et al. [8] Ryan et al. [7] Barua et al. [9]
Identity recognition	Moreira et al. [10]	Chen et al. [11]	–	–	–	–
Lip Reading	Kanamaru et al. [12] Tan et al. [13] Li et al. [14]	–	–	Yu et al. [15]	–	Bulzomi et al. [16] Yoo et al. [17] Rios-Navarro et al. [18]
Voice activity detection	Savran et al. [19]	–	–	Savran [20]	–	–
Driver monitoring systems	Liu et al. [21] Ryan et al. [7] Ryan et al. [22]	Ryan et al. [7]	Ryan et al. [7]	Liu et al. [21]	Kielty et al. [23] Chen et al. [24]	Yang et al. [25] Shariff et al. [26]
Emotion/Expr. recognition	Becattini et al. [27] Berlincioni et al. [28]	–	–	Berlincioni et al. [28]	–	Guo and Huang [29] Becattini et al. [30]
Gaze analysis	Ryan et al. [22]	–	Angelopoulos et al. [31]	–	–	Banerjee et al. [32]
VR/AR	Kang et al. [33]	–	Kang et al. [33]	Kang et al. [33]	–	–
Face pose alignment	Ryan et al. [22]	–	–	–	–	Savran [34] Savran [35] Savran and Bartolozzi [36]

still be recoverable from events [43]. The authors of [44] have thus addressed the problem of encrypting events so to be safely transmitted over an untrusted channel. The proposed encryption strategy should prevent the usage of a computer vision model directly on the protected data. Adopting an opposite paradigm, [45] proposed an event scrambling method that makes the stream impossible to interpret for the human eye but retaining the possibility of applying computer vision models effectively, showing that it is even possible to perform person re-identification on protected data. Finally, [40,41] studied the advantages of processing events on the edge of a federated network to exploit the low power consumption of event cameras to enable an efficient computation directly, without the need to transmit the data to a centralized core. [40] also stresses the fact that event cameras provide an additional layer of privacy due to the lack of intensity information, making it a preferable choice over traditional RGB cameras in scenarios where privacy is a concern.

3. Present research

Neuromorphic Face Analysis has been applied to study several subtopics in the recent literature. We identified a collection of low-level modules, that are commonly used as an intermediate step to address different applications. Low-level modules solve simpler face related tasks, such as face detection, landmark detection, pupil detection, etc. The outputs of such modules can be leveraged for downstream applications, spanning from identity recognition to driving monitoring systems. The only notable exception is face detection, which we consider to be both a low-level module and an application, as it is commonly framed in both declinations. We summarize in Table 1 the most common modules and applications that have been addressed in the literature with an event camera so far, explicitly referring to the corresponding works. As highlighted in Table 1, we identified nine macro areas in which faces are analyzed with an event camera. In the following, we provide a more in-depth analysis of these lines of research.

Face Detection As discussed in Section 2, face detection from event streams is not trivial. However, it is arguably the most important application in neuromorphic face analysis as it enables most face-related tasks. Some works in the literature, have developed neuromorphic face detectors [6–9]. [9] addresses the problem of face detection from event streams using a patch-based model that analyzes small crops of intensity changes to determine whether a face is present. The method was able to achieve good performance on a benchmark dataset of event streams, demonstrating the feasibility of face detection from event data. Starting from this pioneering work, [6] introduced an alternative approach to face detection by utilizing eye blinks as a distinctive cue to identify the presence of a face within a scene. Different approaches instead attempts to solve the task in a similar way to standard solutions in the RGB domain. [7] introduced a gated recurrent YOLO (GR-YOLO) architecture for multi-face and eye detection. To attempt to bridge the gap between synthetic and real events, [8] collected an annotated neuromorphic face detection dataset, along with the comparison of multiple baseline methods for face and landmark detection. These studies have employed metrics such as Euclidean Distance, Mean Square Error loss, Mean Average Precision, Normalized Mean Error, and Receiver Operating Characteristic curve to evaluate the accuracy and effectiveness of face detection models. Other works have followed similar strategies to develop face detection modules finalized to solve a downstream task, such as identity recognition [10], emotion recognition [28], gaze analysis [22], among others. We will discuss in-depth these methods in the following paragraphs.

Identity Recognition Identity recognition is a fundamental task in traditional computer vision, as it can be applied for surveillance and security. Neuromorphic vision offers a valuable asset in this direction, as its incredibly low latency can provide a rich characterization of biometric traits. In fact, [11] presented a biometric authentication system using eye blinks as a unique and distinctive feature for verifying individuals.

To quantify how much identity-related information is carried by facial movements and how such information can be extracted by an automatic neuromorphic system, [10] proposed a different event-based approach for face identity recognition: the events are first aggregated and then normalized and grouped into *face tokens*. These tokens represent the facial activity in a specific time window and are analyzed with a spatio-temporal 3D Convolutional Neural Network (3DCNN). In [45] a similar approach is used for full body re-identification. A siamese ResNet-50 backbone is used to extract features from two event-frame representations, a standard one and its corresponding polar transformation, followed by a Global Average Pooling layer. The features from the two backbones are concatenated and used for identification matching. The authors also presented NVSFD, a dataset tailored for speech-induced facial dynamics. Despite these findings, this remains

quite an unexplored field of research, as to the best of our knowledge no other work on the subject exists yet.

Metrics such as Accuracy, False Positive Rate, and False Negative Rate, have been used to evaluate the performance of identity recognition systems in differentiating individuals based on their unique facial dynamics.

Face Pose Alignment A challenging application in the context of human–robot interaction concerns estimating face poses. This task can also serve as an important pre-processing step for face analysis tasks, as it determines the orientation and position of facial features. [36] addresses the alignment task through a regression cascade of tree ensembles. Efficiency is also taken into account, as the alignment is initialized only when a pose change is detected to minimize unnecessary processing. The authors also recorded a dataset with human subjects exhibiting large head rotations, varying movement speeds, speaking intervals, and multi-human annotations. The same authors also extended this approach in [34,35], by enhancing it with a multi-timescale event encoding strategy. In [34] multiple timescales are used to improve the efficiency and quality of face pose alignment. This is achieved by adaptively adjusting the processing rate based on facial movement intensity. The effect is that the method generates sparse pose-events, reducing computational demands. Further analysis was carried out in [35], where two different timing strategies for face pose alignment are used: constant time frame and constant event count frame. This work addressed the problem of determining the appropriate accumulation time intervals for the face alignment problem. Face pose alignment has also been addressed by [22] for driver monitoring system applications. The authors propose a multi-task network for pose, gaze and occlusion detection, using a shared convolutional backbone and a multi-branch head to predict the three quantities. Metrics used for assessing face pose alignment include the Normalized Mean Error, Precision Error, and E-rate (average positive rate). These metrics measure the accuracy and alignment quality of the detected face pose with reference to a standard model.

LipReading & Voice Activity Detection Given the capacity of event cameras to model high temporal resolution signals, several works have leveraged neuromorphic strategies to analyze mouth-related tasks, such as lip reading and Voice Activity Detection (VAD). VAD is a method for identifying and isolating periods of speech within an audio stream; differently, lip-reading is the process of understanding spoken language by observing the movement of a person's lips. A pioneering approach focusing on the task of Voice Activity Detection using both audio and video was proposed in [19]. The pipeline starts by jointly locating and detecting lip activity using a probabilistic estimation technique after applying spatio-temporal filtering. In a similar vein, [20] proposed an event intensity-based approach by constructing a fully convolutional network for effective neuromorphic VAD. This work presents purely vision based VAD, breaking away from traditional audio-centric approaches. Also lip reading approaches have declined the task as a video or an audio-video approach. Among the video-based approaches, [13] proposed MSTP, a Multi-grained Spatio-Temporal features Perceived network composed of two branches processing low-rate and high-rate event frames: the low-rate branch accumulates events for a longer time-span, thus capturing complete spatial features; the high-rate branch focuses on fine-grained temporal features. The two branches are connected throughout the network using message flow modules merging the features. A 20K samples lip-reading dataset is also released in this work.

The idea of separating neuromorphic information in independent streams at different resolutions has been also followed in [17]. The authors propose RN-Net, an architecture specifically designed to process asynchronous temporal data. It employs simple convolution layers seamlessly integrated with dynamic temporal encoding reservoirs to effectively detect spatio-temporal features at both local and global levels. Similarly, [12,16] have explored the use of event-based cameras for lip reading. Differently from prior work, [16] proposed a spiking neural network architecture for this task, demonstrating an improvement in accuracy. [12] instead leveraged a hybrid approach that combines event-based and frame-based camera data and used a Temporal Convolutional Network to recognize sounds. They also proposed a Japanese utterances dataset.

Differently from VAD, the idea of combining audio with event cameras for lip reading is more recent. Among these works, [15] introduced a spiking neural network for audio-visual speech recognition based on lip reading. Their innovative use of liquid state machines and a soft fusion method inspired by the attention mechanism showcases the effectiveness of a combined audio-visual approach also for the lip reading task.

Interestingly, the neuromorphic paradigm has been adopted for both the video and audio modalities. A few works have in fact used neuromorphic audio sensors along with neuromorphic cameras [14,18]. [14] presented a multi-modal fusion deep network for event-based lip reading using spiking sensors, combining Dynamic Video Sensors (DVS) and Dynamic Audio Sensors (DAS). The fusion model enhances lip reading performance by integrating visual and auditory information, reflecting the multimodal nature of the approach. In [18] the LIPSFUS dataset is introduced, which comprises both visual (lip movement) and auditory (word utterance) information. The data has been captured by a Neuromorphic Auditory Sensor (NAS) and a DVS camera and is synchronized with the same timing source. For lip reading, Accuracy is the primary metric used to evaluate the recognition rate of spoken words based on lip movements. In the context of VAD, metrics such as True Positive Rate, True Negative Rate, and Area under the Curve are employed to assess the system's ability to correctly identify speech from noise.

Facial Expression & Emotion Recognition One of the most challenging, yet suitable domain for event-based approaches, is the analysis of facial expressions and their underlying emotions. The difficulty of these tasks is due to the extremely high temporal resolution at which facial micro-movements happen. This poses a natural limit for standard RGB approaches, which makes the usage of neuromorphic vision sensors an interesting solution. In this context, [27] proposed a method for event based expression modeling analyzing temporal patterns of brightness changes to identify micro-expressions. The study demonstrated that, by using event cameras, it is possible to understand human reactions solely by observing facial expressions and showing improved performances wrt RGB approaches. This idea was then extended in [28], where a dataset for Neuromorphic Event-based Facial Expression Recognition (NEFER) was introduced, containing pairs *(event-data, RGB)* of sequences along with expected and self reported emotions from the subjects. They also proposed a baseline method using a 3D convolutional network for emotion recognition.

A more fine-grained analysis was recently proposed in [29] where two main components are used for micro-expression recognition: an event-based feature extraction module and a global–local feature fusion network. The first module extracts a local count image from an up-sampled video using SloMo [46]. The global–local feature fusion network combines the local count image with global dense optical flow to obtain deeper features, showing that is necessary for accurate micro-expression recognition. The problem of facial expression recognition has also been declined as a valence-arousal estimation problem in [47]. The authors of [30] focus on recognizing a particular set of facial micro expressions called Action Units [2], the combination of which can be mapped to the displayed emotions. The authors propose several architectures, including ResNet+LSTM, ResNet+Transformer and Inception3D and propose a cross-modal loss to exploit RGB data to obtain 3D supervision. The dataset presented in [30] comprised a large number of paired RGB and Event videos with micro expression and landmark annotations. To evaluate the models described above, commonly used metrics include Accuracy, F1-score, Precision, and Underweighted Average Recall.

Gaze Analysis and AR/VR Another intriguing application of event cameras is gaze and pupil detection and tracking. Event cameras are well-suited for this task due to their high temporal resolution, which allows them to accurately track the movements of the eye. This application is closely related to augmented and virtual reality scenarios, as gaze is a natural way of interacting with head-mounted devices.

The challenge behind gaze analysis lies in the fact that eye saccades can happen abruptly and very quickly. [31] presented a hybrid frame-event near-eye gaze tracking system that delivers high update rates up to 10,000 Hz. Their system maintained an accuracy comparable to high-end desktop-mounted commercial trackers using an event camera for both high-frequency events and low-frequency frames to provide a more efficient and accurate representation of eye motion. The authors introduced a model-based eye tracking algorithm that operates at the event rate and an online 2D pupil fitting method that updates a parametric model every one or few events. The system also employed a polynomial regressor to estimate the point of gaze from the parametric pupil model in real-time.

Differently, [32] presented an event-based gaze detection system using retinomorphic events recorded on a DVS camera. The authors developed a new and compact event-based dataset for gaze detection under various lighting conditions. They also proposed a novel event encoding technique that encodes event logs into six-channel images and then used a Convolutional Neural Network for gaze prediction. The authors evaluated their method with multiple metrics and found that it achieved high accuracy in gaze prediction. Finally, [33] developed a remote pupil-tracking technique using event cameras for head-up displays (HUDs). The presented pipeline includes a frame-based accumulator, followed by eye-nose region detection, eye-nose shape alignment and a tracking checker. The metrics used in these studies include Intersection over Union for assessing the overlap between predicted and actual gaze areas, center error for pupil tracking evaluation, and accuracy for gaze mapping. Other commonly used metrics are Average Angle and Average Distance. Together, these metrics evaluate both spatial precision and tracking accuracy in dynamic environments.

Driver Monitoring System Event cameras have proven to be attractive also for developing Driver Monitoring Systems (DMS) capable of detecting fatigue, distractions, or altered states of the driver. Basic modules such as face detectors, landmark estimators or pupil trackers find a natural applicability in the development of such applications. The most basic problem is to localize the driver in the frame. [21] proposed an event-based face detection framework. The proposed method involves constructing event representation, incorporating a shift feature pyramid network and shift context modules that process temporal information at different scales. Similarly, [7,22] also detected the faces of drivers. [22] developed a face detection module jointly estimating head pose, eye gaze and facial occlusions in real-time. The framework is trained on synthetic event-stream data from conventional video datasets and validated on real event camera data. [7] instead predicted eye blinks and pupil detections in addition to facial bounding boxes.

A different take on the problem is found in [23,24], who proposed to directly infer the drowsiness state of the driver. [23] use event-based data to analyze mouth movements in search of yawning behaviors that provide a complementary indicator of tiredness. They also proposed a dataset of pairs of RGB and simulated event camera sequences (using [48]). Instead of only detecting yawns, [24] proposed to also recognize eye blinks and mouth movements. Additionally, they provided the EDDD dataset, dedicated to event-based drowsiness driving detection. Finally, [26] analyzed faces of drivers to estimate distraction. The authors developed a sparse-ResNet, that extracts features efficiently from event data and classifies the driver's distraction. Furthermore, two synthetic event datasets, Drive&Act and DMD, were created to train and evaluate the proposed model. Interestingly, the same task has also been addressed by looking at both faces and bodies in [25]. The paper proposed a real-time driver distraction and action recognition. The performance of the proposed system was evaluated on a large-scale simulated event dataset and a self-recorded real event dataset with a DAVIS346 event camera. Additionally, the authors implemented transfer learning experiments on real event data, demonstrating promising generalization capabilities. Metrics such as F1-score, Precision-Recall, Accuracy, Area under ROC curve and Equal Error Rate are used to evaluate the system's effectiveness in correctly identifying driver behaviors and states.

4. Datasets

At the time of writing, there are no well-established benchmarks and the lack of event camera data poses a challenge to the development of new face analysis models. However, several works that analyze faces with neuromorphic cameras collected footage for carrying out their experiments. Table 2 provides an overview of the datasets that have been collected in the past years for neuromorphic face analysis. While these datasets have contributed valuable insights, they are relatively small in size with a lack of diversity, which limits their generalizability and the robustness of the developed models. Future development must focus on acquiring larger and more diverse datasets to enhance the scalability of neuromorphic face analysis models and to support the creation of standardized benchmarks. At the same time, the spatial resolution of these datasets is limited due to hardware limitations, since only in the last few years event cameras have been equipped with FHD sensors. We report in Table 2 also the annotations that are made available along with the recorded videos. These annotations include low-level labels such as bounding boxes of faces (`Face`), lip regions (`Lip`), eye regions (`Eye`) and coordinates of specific facial keypoints (`Landmk.`). Some datasets also include higher-level annotations such as positive/negative facial reactions to visual stimuli (`Binary reactions`), words pronounced by the subjects (`Utterances`) and whether a subject is speaking (`Voice`).

Event data annotation is not straightforward. Instead of annotating the frame, the labels must be related to a set of spatio-temporal events. Also, since an event camera does not produce any signal in the absence of motion, identifying the entities to annotate is harder. For instance, one may want to annotate a static face, even when it is not visible. In addition, events are often accumulated to be processed with a frame-based model. The accumulation time and the encoding strategy directly affect the quality or the granularity of the annotations. As neuromorphic vision research advances, it will be crucial to develop standardized annotation protocols and benchmarks to ensure consistency and comparability across studies, ultimately facilitating the broader adoption and application of these technologies.

Synthethic Approaches It must be noticed that all dataset in Table 2 are made of real event data, i.e. recorded with an actual neuromorphic sensor. However, a large crop of literature [7,15,22,23,29,36,47] works with synthetic events, obtained by converting RGB videos into neuromorphic streams. Also some of the methods in Table 2 [9,24,26,27] leverage synthetic events to perform additional experiments or train on more data. Some works, such as [27] also convert RGB videos to event streams to trivially annotate events by transferring any available label attached to the original data.

Different event camera simulators have been proposed in the literature, namely, ESIM [49] and V2E [48]. These simulators are capable of producing neuromorphic counterparts from RGB videos. To this end, they first perform a temporal upsampling of RGB frames, with a rate that adapts to the video content and its estimated visual dynamics (the more the video changes, the more frames are added). Then, synthetic events are generated by analyzing the differences between adjacent frames.

While these tools have proven to be beneficial for many tasks they also present drawbacks. As discussed in [47], from which we report also here Fig. 2, the RGB to Event simulators are sensible to the compression of the source video, to the point that the block coding artifacts, barely visible in the original footage, are exaggerated and turned into block-sized *macro*-events. This hinders the applicability of simulators and underlines the domain shift between real and simulated events.

Table 2
Several dataset focused on human faces expression and detection annotated with their respective attributes.

Dataset	Video	Users	Resolution	Public	Annotations
Savran and Bartolozzi [36]	108	30	304 × 204	X	Eye; Lip
Lenz et al. [6]	48	10	640 × 480	✓	Eye blink
Becattini et al. [27]	455	25	640 × 480	X	Binary Reactions
Berlincioni et al. [28]	609	29	1280 × 720	✓	Face; Landmk.; Emotion
Bissarinova et al. [8]	3889	73	408 × 360	✓	Face; Landmk.
Tan et al. [13]	200	40	346 × 260	✓	Lip; Utterances
Chen et al. [24]	260	26	346 × 260	✓	Voice; Eye blink; Yawn
Banerjee et al. [32]	3360	6	1920 × 1080	✓	Pupil Coordinates
Angelopoulos et al. [31]	24	24	346 × 260	✓	Gaze
Rios-Navarro et al. [18]	22	110	128 × 128	✓	Utterances
Barua et al. [9]	–	30	128 × 128	X	Face
Moreira et al. [10]	436	40	–	✓	Identity
Kanamaru et al. [12]	1500	20	1280 × 800	X	Face; Landmk.; Lip
Ryan et al. [22]	–	5	1280 × 720	X	Head pose
Chen et al. [11]	180	45	346 × 260	✓	Eye blink
Savran et al. [19]	360	18	304 × 240	X	Face; Lip; Voice
Becattini et al. [30]	3148	64	1280 × 720	✓	Action Units; Landmk.

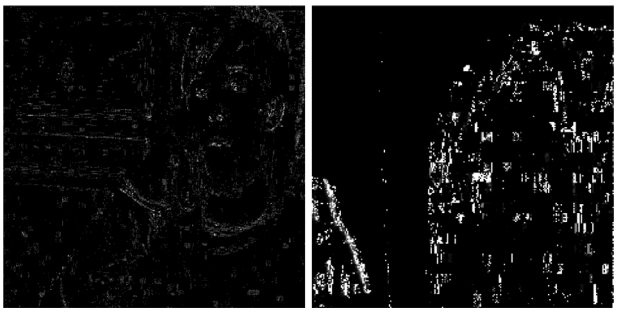

Fig. 2. Effect of compression on simulator [48].
Source: Courtesy of [47].

5. Conclusions and future developments

Neuromorphic Face Analysis is a field of research at a very early stage. Nonetheless, several works have underlined the effectiveness of neuromorphic cameras for capturing facial features, offering several benefits compared to traditional computer vision thanks to its low-latency and high dynamic range. In this paper, we have discussed the state of the research and the related unsolved challenges. The lack of annotated data, especially for equal tasks in RGB, still represents a big issue. Many works have bypassed this problem by relying on synthetic data obtained with a simulator. However, whereas this could be a suitable solution for many problems, we believe that real events should be leveraged to exploit the full capacity of neuromorphic sensors for analyzing faces as micro-movements are lost when converting low-framerate RGB videos. The lack of data in the field of expression and emotion comes as a surprise as observing the extremely fast movements that convey emotions is a perfect application for a sensor with such a low latency as an event camera.

To ensure further progress, there is a pressing need to prioritize the acquisition of extensive and diverse real-world datasets and establish standardized benchmarks. Studying emotions would also allow the development of applications in healthcare, human–computer interaction and could also play a crucial role in advancing augmented reality technologies. Beyond these areas, potential future applications include security and surveillance systems, where real-time monitoring and recognition of faces could enhance safety and threat detection. Neuromorphic face analysis could also be beneficial in remote learning and online education platforms, where real-time emotional feedback could help tailor teaching methods to student needs, improving educational outcomes. In the realm of social robotics, event cameras could enable robots to better understand human emotions and respond accordingly, enhancing human–robot interactions. Additionally, neuromorphic sensors could be utilized in customer service settings, where real-time emotion recognition can help agents provide personalized support, increasing customer satisfaction and improving service quality. The energy-efficient nature of neuromorphic sensors makes them also well-suited for edge computing applications, allowing the integration of these sensors into small, resource-constrained devices, contributing to the growth of decentralized and efficient computing.

In conclusion, while current research has laid a strong foundation, the acquisition of comprehensive datasets and the creation of standardized benchmarks will be critical for the continued evolution of neuromorphic face analysis. The ongoing research and innovations in this area are likely to shape the next generation of intelligent systems and contribute to the continued advancement of artificial intelligence.

CRediT authorship contribution statement

Federico Becattini: Writing – review & editing, Writing – original draft, Investigation, Data curation, Conceptualization. **Lorenzo Berlincioni:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Luca Cultrera:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Alberto Del Bimbo:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the European Commission under European Horizon 2020 Programme AI4Media, grant number 951911. This work was partially supported by the Piano per lo Sviluppo della Ricerca (PSR 2023) of the University of Siena - project FEATHER: Forecasting and Estimation of Actions and Trajectories for Human–robot intERactions.

Data availability

No data was used for the research described in the article.

References

[1] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, Y.-J. Liu, Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 44 (9) (2021) 5826–5846.

[2] P. Ekman, W.V. Friesen, Facial action coding system, Environ. Psychol. Nonverbal Behav. (1978).

[3] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, X. Fu, How fast are the leaked facial expressions: The duration of micro-expressions, J. Nonverbal Behav. 37 (2013) 217–230.

[4] D.E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.

[5] A. Bulat, G. 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.

[6] G. Lenz, S.-H. Ieng, R. Benosman, Event-based face detection and tracking using the dynamics of eye blinks, Front. Neurosci. 14 (2020) 587.

[7] C. Ryan, B. O'Sullivan, A. Elrasad, A. Cahill, J. Lemley, P. Kielty, C. Posch, E. Perot, Real-time face & eye tracking and blink detection using event cameras, Neural Netw. 141 (2021) 87–97.

[8] U. Bissarinova, T. Rakhimzhanova, D. Kenzhebalin, H.A. Varol, Faces in event streams (FES): An annotated face dataset for event cameras, Authorea Prepr. (2023).

[9] S. Barua, Y. Miyatani, A. Veeraraghavan, Direct face detection and video reconstruction from event cameras, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2016, pp. 1–9.

[10] G. Moreira, A. Graça, B. Silva, P. Martins, J. Batista, Neuromorphic event-based face identity recognition, in: 2022 26th International Conference on Pattern Recognition, ICPR, IEEE, 2022, pp. 922–929.

[11] G. Chen, F. Wang, X. Yuan, Z. Li, Z. Liang, A. Knoll, NeuroBiometric: an eye blink based biometric authentication system using an event-based neuromorphic vision sensor, IEEE/CAA J. Autom. Sin. 8 (1) (2020) 206–218.

[12] T. Kanamaru, T. Arakane, T. Saitoh, Isolated single sound lip-reading using a frame-based camera and event-based camera, Front. Artif. Intell. 5 (2023) 1070964.

[13] G. Tan, Y. Wang, H. Han, Y. Cao, F. Wu, Z.-J. Zha, 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.

[14] X. Li, D. Neil, T. Delbruck, S.-C. Liu, Lip reading deep network exploiting multi-modal spiking visual and auditory sensors, in: 2019 IEEE International Symposium on Circuits and Systems, ISCAS, IEEE, 2019, pp. 1–5.

[15] X. Yu, L. Wang, C. Chen, J. Tie, S. Guo, Multimodal learning of audio-visual speech recognition with liquid state machine, in: International Conference on Neural Information Processing, Springer, 2022, pp. 552–563.

[16] H. Bulzomi, M. Schweiker, A. Gruel, J. Martinet, End-to-end neuromorphic lip-reading, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4100–4107.

[17] S. Yoo, E. Yeu-Jer Lee, Z. Wang, X. Wang, W.D. Lu, RN-Net: Reservoir nodes-enabled neuromorphic vision sensing network, 2023, arXiv e-prints, arXiv–2303.

[18] A. Rios-Navarro, E. Piñero-Fuentes, S. Canas-Moreno, A. Javed, J. Harkin, A. Linares-Barranco, LIPSFUS: A neuromorphic dataset for audio-visual sensory fusion of lip reading, 2023, arXiv preprint arXiv:2304.01080.

[19] A. Savran, R. Tavarone, B. Higy, L. Badino, C. Bartolozzi, Energy and computation efficient audio-visual voice activity detection driven by event-cameras, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, IEEE, 2018, pp. 333–340.

[20] A. Savran, Fully convolutional event-camera voice activity detection based on event intensity, in: 2023 Innovations in Intelligent Systems and Applications Conference, ASYU, IEEE, 2023, pp. 1–6.

[21] P. Liu, G. Chen, Z. Li, D. Clarke, Z. Liu, R. Zhang, A. Knoll, Neurodfd: Towards efficient driver face detection with neuromorphic vision sensor, in: 2022 International Conference on Advanced Robotics and Mechatronics, ICARM, IEEE, 2022, pp. 268–273.

[22] C. Ryan, A. Elrasad, W. Shariff, J. Lemley, P. Kielty, P. Hurney, P. Corcoran, Real-time multi-task facial analytics with event cameras, IEEE Access (2023).

[23] P. Kielty, M.S. Dilmaghani, C. Ryan, J. Lemley, P. Corcoran, Neuromorphic sensing for yawn detection in driver drowsiness, in: Fifteenth International Conference on Machine Vision, ICMV 2022, Vol. 12701, SPIE, 2023, pp. 287–294.

[24] G. Chen, L. Hong, J. Dong, P. Liu, J. Conradt, A. Knoll, EDDD: Event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor, IEEE Sens. J. 20 (11) (2020) 6170–6181.

[25] C. Yang, P. Liu, G. Chen, Z. Liu, Y. Wu, A. Knoll, Event-based driver distraction detection and action recognition, in: 2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI, IEEE, 2022, pp. 1–7.

[26] W. Shariff, M.S. Dilmaghani, P. Kielty, J. Lemley, M.A. Farooq, F. Khan, P. Corcoran, Neuromorphic driver monitoring systems: A computationally efficient proof-of-concept for driver distraction detection, IEEE Open J. Veh. Technol. (2023).

[27] F. Becattini, F. Palai, A. Del Bimbo, Understanding human reactions looking at facial microexpressions with an event camera, IEEE Trans. Ind. Inform. 18 (12) (2022) 9112–9121.

[28] L. Berlincioni, L. Cultrera, C. Albisani, L. Cresti, A. Leonardo, S. Picchioni, F. Becattini, A. Del Bimbo, Neuromorphic event-based facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4108–4118.

[29] C. Guo, H. Huang, GLEFFN: A global-local event feature fusion network for micro-expression recognition, in: Proceedings of the 3rd Workshop on Facial Micro-Expression: Advanced Techniques for Multi-Modal Facial Expression Analysis, 2023, pp. 17–24.

[30] F. Becattini, L. Cultrera, L. Berlincioni, C. Ferrari, A. Leonardo, A. Del Bimbo, Neuromorphic facial analysis with cross-modal supervision, in: Proceedings of the IEEE/CVF European Conference on Computer Vision, 2024.

[31] A.N. Angelopoulos, J.N. Martel, A.P. Kohli, J. Conradt, G. Wetzstein, Event based, near eye gaze tracking beyond 10,000 hz, 2020, arXiv preprint arXiv:2004.03577.

[32] A. Banerjee, S.S. Prasad, N.K. Mehta, H. Kumar, S. Saurav, S. Singh, Gaze detection using encoded retinomorphic events, in: International Conference on Intelligent Human Computer Interaction, Springer, 2022, pp. 442–453.

[33] D. Kang, Y.K. Lee, J. Jeong, Exploring the potential of event camera imaging for advancing remote pupil-tracking techniques, Appl. Sci. 13 (18) (2023) 10357.

[34] A. Savran, Multi-timescale boosting for efficient and improved event camera face pose alignment, Comput. Vis. Image Underst. 236 (2023) 103817.

[35] A. Savran, Comparison of timing strategies for face pose alignment with event camera, in: 2023 8th International Conference on Computer Science and Engineering, UBMK, IEEE, 2023, pp. 97–101.

[36] A. Savran, C. Bartolozzi, Face pose alignment with event cameras, Sensors 20 (24) (2020) 7079.

[37] S.U. Innocenti, F. Becattini, F. Pernici, A. Del Bimbo, Temporal binary representation for event-based action recognition, in: 2020 25th International Conference on Pattern Recognition, ICPR, IEEE, 2021, pp. 10426–10432.

[38] E. Mueggler, C. Bartolozzi, D. Scaramuzza, Fast event-based corner detection, 2017.

[39] S. Al-obaidi, Privacy aware human action recognition: an exploration of temporal salience modelling and neuromorphic vision sensing (Ph.D. thesis), University of Sheffield, 2020.

[40] N. Delilovic, D. Salaj, Bio-inspired neuromorphic AI methods enables privacy respecting security and surveillance, Trans. Adv. Res. (2021).

[41] B. Han, Q. Fu, X. Zhang, Towards privacy-preserving federated neuromorphic learning via spiking neuron models, Electronics 12 (18) (2023) 3984.

[42] Y. Dong, Y. Li, D. Zhao, G. Shen, Y. Zeng, Bullying10k: A large-scale neuromorphic dataset towards privacy-preserving bullying recognition, in: Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023.

[43] H. Rebecq, R. Ranftl, V. Koltun, D. Scaramuzza, Events-to-video: Bringing modern computer vision to event cameras, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3857–3866.

[44] P. Zhang, S. Zhu, E.Y. Lam, Event encryption: Rethinking privacy exposure for neuromorphic imaging, 2023, arXiv preprint arXiv:2306.03369.

[45] S. Ahmad, G. Scarpellini, P. Morerio, A. Del Bue, Event-driven re-id: A new benchmark and method towards privacy-preserving person re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 459–468.

[46] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, J. Kautz, Super slomo: High quality estimation of multiple intermediate frames for video interpolation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9000–9008.

[47] L. Berlincioni, L. Cultrera, F. Becattini, A.D. Bimbo, Neuromorphic valence and arousal estimation, 2024, arXiv:2401.16058.

[48] Y. Hu, S.C. Liu, T. Delbruck, V2e: From video frames to realistic DVS events, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, IEEE, 2021, URL http://arxiv.org/abs/2006.07722.

[49] H. Rebecq, D. Gehrig, D. Scaramuzza, ESIM: an open event camera simulator, Conf. Robotics Learn. (CoRL) (2018).