frontiers | Frontiers in Signal Processing

TYPE Review
PUBLISHED 29 August 2025
DOI 10.3389/frsip.2025.1585242

OPEN ACCESS

EDITED BY
Frederic Dufaux,
Université Paris-Saclay, France

REVIEWED BY
Serhan Cosar,
The Open University, United Kingdom
Vijetha U,
St Joseph Engineering College Mangalore, India

*CORRESPONDENCE
Mira Adra,
mira.adra@eurecom.fr

†These authors have contributed equally to this work

RECEIVED 28 February 2025
ACCEPTED 07 August 2025
PUBLISHED 29 August 2025

CITATION
Adra M, Melcarne S, Mirabet-Herranz N and Dugelay J-L (2025) Event-based solutions for human-centered applications: a comprehensive review.
Front. Signal Process. 5:1585242.
doi: 10.3389/frsip.2025.1585242

Event-based solutions for human-centered applications: a comprehensive review

Mira Adra[1*†], Simone Melcarne[2†], Nelida Mirabet-Herranz[2†] and Jean-Luc Dugelay[2]

[1]GTD International, Ramonville-Saint-Agne, France, [2]Department of Digital Security, EURECOM, Biot, France

Event cameras, often referred to as dynamic vision sensors, are groundbreaking sensors capable of capturing changes in light intensity asynchronously, offering exceptional temporal resolution and energy efficiency. These attributes make them particularly suited for human-centered applications, as they capture both the most intricate details of facial expressions and the complex motion dynamics of the human body. Despite growing interest, research in human-centered applications of event cameras remains scattered, with no comprehensive overview encompassing both body and face tasks. This survey bridges that gap by being the first to unify these domains, presenting an extensive review of advancements, challenges, and opportunities. We also examine less-explored areas, including event compression techniques and simulation frameworks, which are essential for the broader adoption of event cameras. This survey is designed to serve as a foundational reference that helps both new and experienced researchers understand the current state of the field and identify promising directions for future work in human-centered event camera applications. A summary of this survey can be found at https://github.com/nmirabeth/event_human.

KEYWORDS

neuromorphic sensors, event-based vision, human-centered applications, event compression, privacy preserving, body analysis, face analysis

1 Introduction

Human-centered applications have long been at the forefront of computer vision, driven by the need to understand and analyze human activities in diverse contexts. Such applications span critical domains, including, among others, surveillance, where the objective is to monitor human behavior in real-time for public safety; biometric authentication, which can leverage unique individual features for secure identification; interactive systems that enable seamless human-computer interactions through gesture or expression recognition; and behavioral analysis, which provide insights into physical activity and social behaviors.

Although this research area encompasses a wide range of tasks, it can be broadly divided into two main directions of focus: body and face analysis. To this day, several traditional computer vision techniques have been developed to address both categories (Viola and Jones, 2001; Zhang et al., 2016; Ma et al., 2023; Sun et al., 2019; Yan et al., 2018). Such widely recognized methods primarily rely on conventional RGB cameras and perform frame-based analysis, without considering or incorporating any other data types. However, these standard solutions are hindered by fundamental limitations. On the one hand, they are

constrained by temporal resolution, meaning that they often fail to capture fast and subtle movements that, for example, distinguish micro-expressions (Becattini et al., 2025; Yan et al., 2013) or may happen in rapid gait changes (Wang et al., 2019). On the other hand, they are prone to motion blur in dynamic scenarios and struggle with challenging lighting conditions, such as high contrast or low illumination (Cazzato and Bono, 2024); finally, these approaches are resource-intensive as they require significant memory and processing power to handle high frame-rate video streams (Gallego et al., 2022).

These challenges have recently fueled the interest in using neuromorphic (often referred to as dynamic vision sensors or event-based) cameras, which can offer a transformative solution. Event cameras are often called neuromorphic or bio-inspired sensors because they are modeled after the retina's sensory neurons, mimicking how photoreceptors respond to changes in light intensity rather than capturing static frames. Like the retina, event cameras operate asynchronously, with each pixel independently detecting brightness changes, much like sensory neurons and retinal ganglion cells in the visual system. This design is inspired by the transient pathways in biological vision, which specialize in detecting motion and dynamic changes in the environment. Just as the retina has cells dedicated to processing movement and contrast to help us perceive motion, event cameras replicate this functionality by focusing only on changes in brightness over time (Steffen et al., 2019; Posch et al., 2014). Unlike the common belief that our eyes 'see' everything continuously, the reality is that biological mechanisms such as microsaccades and the transient pathway ensure continuous perception by creating subtle changes in light input similar to the functioning of an event camera. Without these mechanisms, our vision would fade. Moreover, event cameras are also considered neuromorphic because they integrate seamlessly into neuromorphic computing frameworks, like spiking neural networks (SNNs), which mimic the way biological neurons transmit information as spikes, allowing for efficient, brain-like data processing. This design enables a series of unique characteristics that make them particularly well-suited for the aforementioned tasks.

This survey provides an overview of the current situation and the progress made in using event-based cameras specifically for human-centered applications, identifying key developments, existing challenges, and potential research directions to guide researchers at all levels. A preprint version of this work is available as Adra et al. (2025).

1.1 Related surveys

Event-based cameras are being leveraged in a growing number of applications. Given their impact on both academia and industry, several surveys and reviews have been published in recent years, playing an important role in research as they summarize the state of the art, identify gaps, and propose directions for future investigation. Some researchers have aimed at providing comprehensive overviews of the emerging field of event-based vision, describing in detail the physical sensor design and the technical specifications. In this direction, one of the earliest papers is represented by Etienne-Cummings and der Spiegel (1996), a survey published more than 2 decades ago, which traces the history of neuromorphic sensor development. Since then, various surveys have been published over the years (Kramer and Indiveri., 1998; Indiveri., 2008; Liu and Delbruck., 2010), exploring topics such as hardware developments and the design of very-large-scale integration (VLSI) neuromorphic circuits for processing signals from event-based cameras. A more recent and exhaustive review is given by Gallego et al. (2022), which focused on event-based vision systems operating principles, underlying algorithms, and a wide range of applications addressed, mainly including robotics and perception. Similarly, Chakravarthi et al. (2024) investigated the latest innovations in event camera technology, examining models, datasets, and diverse applications across various domains, highlighting their impact on research and development. In addition, Cazzato and Bono (2024) proposed an application-driven survey, illustrating various outcomes across different application fields and exploring the issue of dataset availability.

In a different line, other studies have mainly focused on a specific topic, showing how event-based methods have evolved to tackle the challenges within that particular field. For example, in the targeted context of human-related analysis, Verschae and Bugueno-Cordova (2023) focused on event-driven gesture and facial expression recognition and compared different algorithms and benchmarks for the purpose of performance evaluation. Becattini et al. (2025) discussed neuromorphic solutions for face analysis, which included detection, recognition, and emotion analysis and compared these with traditional approaches. Eye motion analysis with event cameras, and their potential for applications such as gaze estimation or blink detection, were explored by Iddrisu et al. (2024a).

1.2 Scope and value of our survey

The scope of this survey is specifically focused on human-centered applications of event cameras. These include applications addressing humans as a whole—such as gesture and action recognition, human tracking, and pose estimation—as well as applications focused on facial analysis, including face detection, emotion recognition, and face recognition. Unlike existing surveys, which often concentrate solely on face applications (Becattini et al., 2025) or narrowly on a subset of human actions (Verschae and Bugueno-Cordova, 2023), this survey aims to provide a comprehensive overview of all human-centered event-based applications. Our motivation stems from the evolution of event camera research trends, which have expanded beyond traditional robotics and high-speed tracking applications to demonstrate significant advantages in downstream human-centered tasks as highlighted in Figure 1. It is important to note that Figure 1 was created based on a focused methodology based on papers with clear and explicit relevance to robotics and human-centered applications.

What makes our survey particularly valuable is its uniqueness. To the best of our knowledge, this is the first survey to thoroughly target human-centered applications of event cameras, covering both body- and face-oriented use cases in a unified framework. Secondly, we want to emphasize the authors' contribution to the publications included in this survey, as our findings at Eurecom have contributed to advances in the field of neuromorphic computation across various human-centered applications for both body and face.

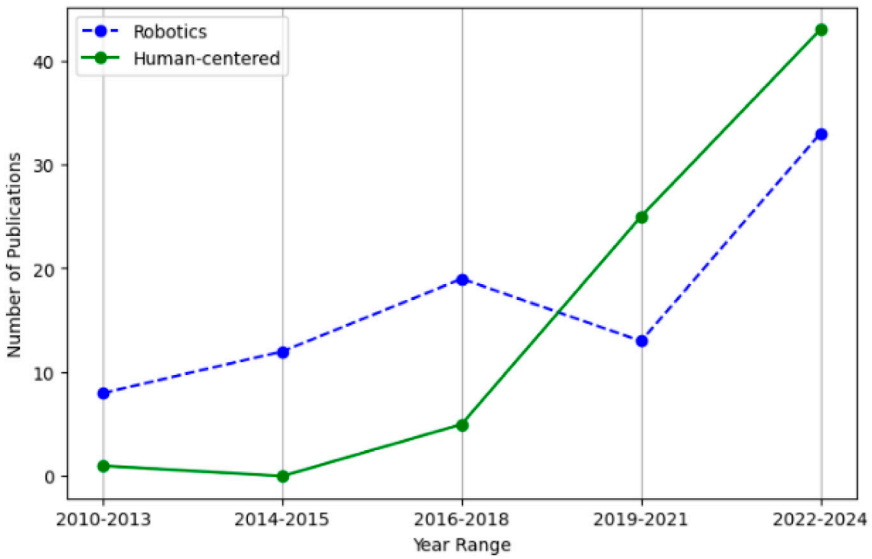

FIGURE 1
Evolution of research focus: Comparing the number of publications on robotics versus human-centered applications of event cameras.

1.3 Structure and coverage of our survey

This survey is organized as follows: in Section 2, we introduce the foundational concepts of event cameras, briefly explaining how they work, listing their main features and discussing the advantages and disadvantages of using this type of sensor in the specific context of human-centered applications; Section 3 explores the description of various strategies for representing event data, underlining the specificity that each of them has and for simulating event data starting from RGB videos using popular methods in the literature. The following part presents a discussion on state-of-the-art datasets designed for experimental validation before highlighting some key techniques for event compression that enable efficient data handling; Section 4 provides an overview of the current state of the literature in the field of human-centered event-based applications, categorizing them into body- and face-related tasks; finally, in Section 5, we conclude with an analysis of current trends and offer future perspectives for the development and integration of event-based vision systems in real-world applications.

We want to underline that, in this survey, we use the terms event camera, neuromorphic vision sensor, and Dynamic Vision Sensor (DVS) interchangeably for simplicity. Technically, neuromorphic sensor denotes a broader class of bio-inspired devices, event camera refers to a specific subset that emit per-pixel events based on changes in brightness, and DVS identifies a specific family of such cameras developed by iniVation. However, this interchangeable use is common in the related literature and surveys, and we follow the same convention here.

2 Foundations of event-based vision systems

Event cameras are bio-inspired vision sensors that represent a fundamental breakthrough compared to traditional frame-based imaging. Rather than capturing entire frames at regular intervals, these sensors operate in an *asynchronous* manner, meaning that each pixel independently triggers an event only when a change in brightness exceeds a specific threshold (Cazzato and Bono., 2024; Gallego et al., 2022; Zheng et al., 2024). As a direct consequence, if no changes are detected, no data is generated, significantly reducing bandwidth usage.

The properties of these cameras align with the dynamic and unpredictable characteristics of human activities, making them particularly suited for capturing fast and irregular actions, such as facial expressions or body movements (Adra et al., 2024b; Eddine and Dugelay., 2022). However, they also introduce unique challenges that require a deep rethinking of conventional vision methodologies.

In the following subsections, we provide the reader with a general overview of the functioning of event cameras, highlighting the advantages to be leveraged and the disadvantages to be managed.

2.1 Operating principle and key features

A new *event* e_k is triggered at a certain pixel $\mathbf{u} = (u_k, v_k)$ whenever the change in brightness $\Delta L(\mathbf{u}, t_k)$, calculated between the current time t_k and the time of the last event t_{k-1}, exceeds a predefined contrast threshold $T > 0$. Each event encodes 1) the pixel's identity $\mathbf{u} = (u_k, v_k)$ that indicate the location of the change, 2) a timestamp t_k, capturing the precise time the event occurred, and 3) the polarity $p_k \in \{-1, +1\}$, specifying whether the brightness increased or decreased; as a result, an event is represented as a tuple, as shown in Equations 1, 2:

$$e_k = \{\mathbf{u}, t_k, p_k\}, \tag{1}$$

$$with\ \Delta L(\mathbf{u}, t_k) = L(\mathbf{u}, t_k) - L(\mathbf{u}, t_{k-1}) \geq p_k T \tag{2}$$

In this sense, what a neuromorphic camera produces is nothing more than a spatio-temporally localized stream of events which can be formally described as shown in Equation 3:

$$\mathcal{E} = \{e_k\}_{k=1}^{K} where\ k \in K, \quad (3)$$

with K representing the total number of events occurring within the entire recording time interval. This data-driven design makes the output rate dependent on scene dynamics, since faster motion or more significant brightness variations lead to a higher event rate.

Several key features distinguish these sensors: *High Temporal Resolution*, since events are time-stamped with microsecond (μs) precision and allow the detection of rapid actions or subtle gestures, without motion blur (Gallego et al., 2022); this ability at sensing very fast motion is invaluable for tasks requiring fine-grained temporal analysis, such as gait analysis or blink detection.

Low Latency, as events are generated and transmitted as soon as brightness changes occur ensuring real-time responsiveness without the need to wait for a global frame exposure time; this is an essential requirement for applications like human tracking. For example, the DVS128 camera outputs events at a rate of 1 million events per second (Meps) (Lichtsteiner et al., 2008) while the Samsung DVS-Gen4 has a higher bandwidth of 1,066 Meps.

High Dynamic Range (HDR), since event cameras can capture scenes with a vast range of lighting conditions, from dark environments to bright daylight. These sensors boast a dynamic range that reaches 140 dB, far surpassing the 60 dB typically seen in high-quality frame-based cameras (Gallego et al., 2022). A range of 140 dB indicates the ability to handle brightness differences of up to 10,000,000:1, compared to only 1,000:1 for 60 dB. By minimizing saturation and preserving fine details, event cameras support applications like face detection or pose estimation under challenging lighting conditions.

Low Power Consumption, as event cameras significantly reduce the amount of data produced by only capturing changes in the scene, rather than full images as conventional frame-based cameras do. This feature makes event cameras ideal for long-term monitoring systems or wearable devices where power constraints are critical. For example, Barchid et al. (2023) demonstrated that Spiking-FER, when combined with event data, is 47.42× to 65.39× more energy-efficient than comparable artificial neural networks, highlighting the energy-saving potential of event-based systems and their suitability for low-power applications on edge devices.

2.2 Privacy preservation

Another important characteristic that has brought event cameras to the spotlight is their potential for preserving user's privacy. Since they capture only dynamic scene changes, raw event data are inherently challenging to interpret compared to conventional RGB imagery. This feature adds a level of privacy by design, making event streams less likely to reveal sensitive identity information (Becattini et al., 2025; Al-Obaidi., 2020; Delilovic and Salaj., 2021; Dong et al., 2023; Han et al., 2023). However, the assumption that event data are inherently privacy-preserving has been challenged by advancements in deep learning-based event-to-image reconstruction techniques (Rebecq et al., 2019), which can recover intensity map images from event streams and expose personal identity information. This has led to increased efforts to enhance the privacy of event-based data. In this direction, Du et al. (2021) proposed a 2D chaotic mapping-based algorithm that scrambles event positions and flips polarities, combined with a dynamic key-updating mechanism, ensuring data security while maintaining high efficiency on resource-constrained devices. Similarly, Zhang et al. (2024a) introduced an encryption framework to secure event streams during transmission, effectively preventing a direct application of computer vision models on the encrypted data. In the same line of research, Ahmad et al. (2023), Ahmad et al. (2024) formulated an anonymization strategy that randomizes event streams, making them unintelligible to human observers and demonstrating strong resilience against image reconstruction attacks, inversion, and adversarial learning attempts, while still retaining the information necessary for downstream tasks like person re-identification or human pose estimation. Bendig et al. (2024) designed a novel pipeline for anonymizing event camera data by employing a learnable data-dependent noise prediction network combined with adversarial training, which was able to remove personally identifiable features to prevent re-identification.

When discussing privacy in the context of event cameras, it is crucial to consider it from the machine's perspective, as the primary threat often arises from how machines interpret and utilize data. In their natural form, event data are completely unreadable to humans, appearing as sparse, asynchronous events that lack any recognizable visual information. However, when these events are reconstructed into frames, privacy concerns become more apparent for humans but are significantly greater for machines (Du et al., 2021). Reconstructed frames are typically grayscale and of low resolution but retain substantial information due to their high temporal resolution, often reconstructed at rates approaching 5000 FPS. From a human viewpoint, these frames may appear inferior to traditional RGB images, especially in applications like action recognition where facial details are unclear. Yet, for machines, reconstructed frames hold significant value as they exploit motion edges and spatiotemporal patterns rather than visual clarity, leveraging the rich temporal data inherent in event-based recordings. This distinction emphasizes the need to develop privacy measures based on the machine's capacity to extract sensitive information, recognizing that what seems visually obscure to humans may still be highly informative for automated systems.

2.3 Challenges

Despite their advantages, event cameras also present distinct challenges: while their spatially sparse and temporally asynchronous output allows for more efficient data storage, it necessitates the development of specialized algorithms to process and extract meaningful information that can be used in order to successfully downstream learning tasks. Handling events in an effective manner requires either to employ specialized frameworks, *i.e.,* spiking neural networks (Ghosh-Dastidar and Adeli., 2009), or to represent the event data with more conventional formats, typically in the form of frames (see Section 3.2). Traditional computer vision techniques are fundamentally designed for dense and synchronous images and

therefore are not directly compatible with this novel data format (Gallego et al., 2022). When these methods, trained on RGB frames, are tested with input data coming from event cameras, they predictably struggle to perform well. The lack of a continuous flow of frames leads to a discrepancy with the underlying working assumptions that these methods are based on, resulting in low confidence or inaccurate results, as demonstrated in the study conducted by Becattini et al. (2025) in the specific context of face detection and landmark prediction. Furthermore, event-based sensors might exhibit inherent noise and non-idealities due to hardware constraints or environmental conditions that obscures data and complicates the interpretation. In this sense, the use of robust preprocessing techniques is essential to ensure a reliable performance.

3 Event camera design and processing

In this section, we provide a comprehensive overview of event data, beginning with its representation techniques to facilitate efficient processing and compatibility with existing architectures. We then summarize and categorize the available datasets into real and synthetic, as well as body- and face-focused datasets. To address the scarcity of datasets, we further discuss event data simulators as a vital tool for generating synthetic data. Finally, we introduce event data compression, emphasizing its importance for real-time, human-centered applications of event cameras.

3.1 Camera models and selection criteria

The first Dynamic Vision Sensor (DVS) was introduced in 2008 by iniVation as the DVS128, offering a resolution of 128 × 128 pixels Lichtsteiner et al. (2008). Since then, several event camera models have been developed, primarily by companies such as iniVation[1], Prophesee,2023[2], Samsung (Suh et al., 2020), and CelePixel (Chen and Guo, 2019). IniVation's DAVIS series, such as the DAVIS240 and DAVIS346, combines event-based and frame-based sensing, offering resolutions up to 346 × 260 pixels and dynamic ranges of 120 dB, making them versatile for mixed sensing tasks. Prophesee, (2023) Gen3 and Gen4 cameras, with resolutions as high as 1,280 × 720 pixels and dynamic ranges exceeding 120 dB, are well-suited for applications requiring high spatial detail. CelePixel's CeleX cameras provide features such as grayscale output and IMU integration, while Samsung's DVS-GEN3 and DVS-GEN4 stand out with bandwidth capacities up to 1,066 Meps (million events per second) for high-speed applications.

Gallego et al. (2022) provided a comprehensive comparison of commercial and prototype event cameras serving as a critical reference for researchers to match camera capabilities with their specific application requirements. When selecting an event camera for human-centered applications, specific criteria play a critical role and depend heavily on the task. For action recognition and human tracking, high temporal resolution and low latency (e.g., iniVation's DAVIS240 at 12μs) are critical to capture fast motion dynamics. For facial analysis or anonymization, a higher spatial resolution is often more important to capture fine-grained details, as seen in the CeleX-IV (768 × 640 pixels) or Prophesee Gen4 CD (1,280 × 720 pixels). Applications requiring operation in challenging lighting conditions, such as outdoor crowd density estimation, benefit from models with a high dynamic range (e.g., 120 dB in DAVIS346 or 143 dB in CeleX-IV). Additionally, power consumption is significant for wearable or mobile systems, where models like iniVation's DAVIS240 (5-14 mW) are advantageous.

3.2 Data representation

As presented in the previous section, event cameras operate fundamentally differently from traditional frame-based cameras, resulting in asynchronous event streams encoding changes in intensity at each pixel with microsecond precision. These event streams, while rich in spatiotemporal information, require specialized processing techniques to extract meaningful features. Over time, various representations have emerged, each tailored to address specific challenges and applications. These representations can be categorized into the following groups based on their methodological approach and functional focus:

1. Foundational representations: These include the earlier approaches such as Event Count (Zhu et al., 2018a), Event Histogram, also referred to as event intensity frame, (Liu and Delbrück., 2018), Temporal Binary Representation (Innocenti et al., 2021), Time Surface (Lagorce et al., 2017), and Memory Surface (Pradhan et al., 2019) which prioritize simplicity and provide a quick way to interpret event data
2. Structural representations: These methods leverage advanced processing techniques to represent the spatial and temporal relationships of events. This includes the graph representation first proposed by Bi et al. (2019) and Bi et al. (2020) and then utilized as well by Schaefer et al. (2022) and Deng et al. (2022), which leverages graph theory to process event data both spatially and temporally. Similarly, Tavanaei et al. (2019) proposed the Spiking tensor representation that tries to mimic the brain neurons as much as possible and represents event data as binary tensors. Moreover, Voxel Grid Representation - first proposed in Zhu et al. (2018b) - provides a more detailed approach by discretizing the event stream into 3D spatiotemporal grids. This representation is actually used to train complex networks such as video-based transformers and image reconstruction models.
3. Reconstructed Frame Representations: In attempts to bridge the gap between event-based and frame-based frameworks, it also became popular in research to rely on representations like E2VID Frames (Rebecq et al., 2019) which allow us to mimic video frames and leverage the power of traditional Convolutional Neural Networks (CNNs) and even achieve better results in downstream applications compared to directly using event data.
4. Fused representations: Recently proposed by Gao et al. (2023), the Learnable Multi-Fused Representation (LMFR) integrates

1 https://inivation.com

2 https://www.prophesee.ai

TABLE 1 Event data representations and their details.

Name	Details
Event count	Event data is aggregated by counting the number of events that occur at each pixel within a fixed time interval. This approach provides a straightforward summary of activity, often used as a baseline representation
Event histogram	Similar to the event count, but instead of a single time interval, events are grouped and counted in temporal bins, creating a distribution of event activity that captures variations with more levels of detail
Time surface/surface of active events	Represents data as a continuous map where each pixel value corresponds to the most recent timestamp of an event at that location. This highlights recent activity and is often used to track motion or identify edges
Memory surface	Event data are represented as a temporal map where each pixel's value indicates the time elapsed since the last event occurred at that location within a fixed time window. This approach encodes temporal information by retaining a "memory" of inactivity, making it useful for identifying patterns, and tracking regions with recent or ongoing motion
Voxel grid	Event data is sliced temporally into small time intervals, creating a sequence of event slices. These slices are then stacked into a 3D grid, where each voxel represents the activity in a spatial region during a specific time window. This allows for preserving both spatial and temporal resolution
Spike tensor	Represents data as binary tensors indicating the occurrence of spikes in specific spatiotemporal locations. The tensor is separated into two channels for positive and negative polarities
Graph	Represents data as a graph, where events are treated as nodes in a graph with polarity as the node feature. Then, edges are created between nodes to represent spatiotemporal relationships, often used for tasks like pattern recognition
E2VID Frame	Represents data as reconstructed frames by using neural networks to convert the sparse event stream into intensity frames. This allows event data to be used with traditional frame-based computer vision methods
Temporal binary representation	Events are first stacked together into intermediate binary representations where each pixel can be considered as a binary string. These frames are then grouped into a single frame by applying binary to decimal conversion. Most popular in face analysis applications

multiple event representations, such as Time Surface, Event Frames, and Event Count, into a single embedding in a learnable manner. By leveraging their complementary features, LMFR enhances performance in complex tasks.

Many other representations exist, but the most commonly used benchmarks are detailed in Table 1. Moreover, to provide a better understanding and facilitate comparison, we visualize a selection of these representations in Figure 2a-g using event data from the Gait3 dataset, specifically for a person walking from left to right. These visualizations highlight the diversity in how event data can be processed and interpreted for different applications. Note that for the graph representation in Figure 2f, while it is typically a 3D structure (like a point cloud) with events as nodes, for visualization purposes, we project it onto the temporal axis and represent it as a 2D structure.

While we have explored the different event data representations, it is equally important to evaluate their respective advantages and limitations to understand their suitability for various applications. First, Foundational representations, such as Event Count and Time Surface, are fast to compute and highly efficient, and can serve as lightweight baselines; however, their simplicity often results in significant loss of spatial information. In contrast, structural representations, such as graph and spiking tensor approaches, capture spatio-temporal relationships effectively. Graph representations maintain connections between event nodes, making them effective for spatially complex tasks, though they can become computationally expensive in scenarios with dense movements or high event rates, often requiring filtering (Adra et al., 2024a). Spiking tensors, on the other hand, align naturally with the neuromorphic nature of event cameras and are compatible with Spiking Neural Networks. However, SNNs face challenges like gradient optimization issues and remain less mature compared to CNNs (Eshraghian et al., 2023). Similarly, reconstructed representations, such as intensity frames from E2VID, leverage the high temporal resolution of event cameras and leverage the power of the use of well-established CNN architectures; however, they often lose key spatial details due to their grayscale and low-resolution frames. Moreover, these frames can reintroduce redundancy through the huge amounts of frames generated and the static background information, which contradicts one of the main benefits of event-based systems. Finally, despite enhancing performance across tasks, the fused representation which combines multiple approaches is considered computationally intensive and requires significant resources to optimize effectively (Fan et al., 2025). Ultimately, the choice of representation depends on the specific application requirements, balancing computational efficiency, spatio-temporal complexity, and compatibility with existing architectures.

3.3 Neural network architectures for event data

In the literature, several prominent neural network architectures are utilized to process event data and downstream learning tasks, each utilizing unique paradigms. In this section, we outline the most commonly used approaches, briefly describing the foundational concepts.

(1) Spiking Neural Networks: Event data are naturally compatible with SNNs, as they also operate based on an event-driven strategy. SNNs use discrete spikes rather than continuous activations, and perform well in handling spatio-temporal data while offering remarkable energy efficiency. Several works demonstrated how SNNs can better model the asynchronous nature of event data (Liu et al., 2021; Barchid et al., 2023), while others further refined

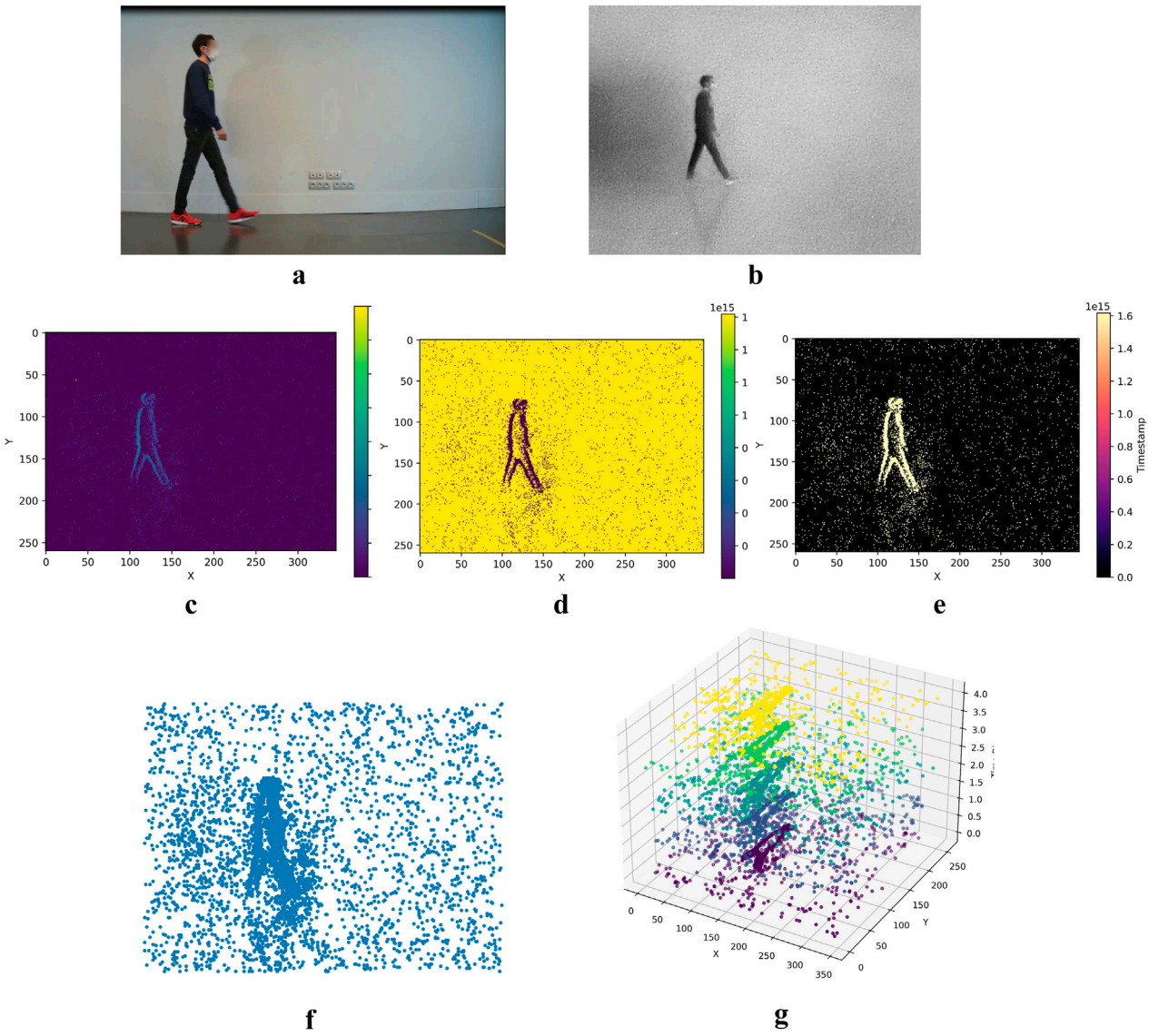

FIGURE 2
Visualizations of various event data representations derived from the Gait3 dataset, specifically illustrating a man walking from left to right.From 2 **(a–g)**, respectively: RGB reference frame, E2VID Reconstruction, Event Count, Memory Surface, Time Surface, Graph, Voxel Grid.

their application to real-world problems (Bulzomi et al., 2023; Vicente-Sola et al., 2025). These networks are considered neuromorphic because, like event cameras, they are inspired by how the brain works. They are a good fit for processing event data asynchronously and without requiring frame reconstruction or heavy preprocessing. They are also energy efficient and suitable for low-power applications like embedded systems. However, training them is challenging due to the non-differentiable nature of spikes, which complicates the use of standard backpropagation. Surrogate gradient methods have been introduced to address this issue, yet training remains less efficient and less mature compared to traditional deep learning models. As a result, their performance often lags behind state-of-the-art frame-based networks, and tools for large-scale deployment are still limited.

(2) Graph Neural Networks (GNNs): GNNs are specialized artificial neural networks designed to process and analyze input data as graphs, *i.e.,* structures that represent the relations (edges) between a collection of entities (nodes). GNNs directly operate on the spatio-temporal graph structure from the raw events, where the nodes can represent the event pixels and the edges the spatio-temporal dependencies between the nodes (Wang et al., 2021; Gao et al., 2024). One advantage of these networks is that they model event data as nodes and edges, making them well-suited for tasks like pose estimation or action recognition. This allows GNNs to capture complex spatio-temporal relationships and represent irregular and sparse structures efficiently. However, graph construction is often complex and computationally expensive, and these models are harder to scale and optimize than CNNs.

(3) Convolutional Neural Networks: CNNs are particularly suited for grid-structured data, *i.e.,* images, and aim to perform local feature extraction through convolution operations. To harness the capabilities of CNNs for event data, these data are either encoded

TABLE 2 Papers presented in this survey, classified by the type of AI architecture used for their models.

SNN	Graph NN	CNN	Transformers	Not AI-based
Liu et al. (2021) Barchid et al. (2023) Ren et al. (2023b) Bulzomi et al. (2023) Tao et al. (2024) Vicente-Sola et al. (2025)	Wang et al. (2021) Eisl et al. (2023) Fu and Yan. (2023) Gao et al. (2024)	Li et al. (2019) Wang et al. (2019) Sokolova and Konushin. (2019) Ryan et al. (2021) Banerjee et al. (2022) Becattini et al. (2022) Moreira et al. (2022) Plizzari et al. (2022) Ryan et al. (2023) Gao et al. (2023) Rios-Navarro et al. (2023) Bissarinova et al. (2023) Berlincioni et al. (2023) Goyal et al. (2023) Kanamaru et al. (2023) Xiao et al. (2024) Kohyama et al. (2024) Adra et al. (2024b) Iddrisu et al. (2024b)	Xu et al. (2020) de Blegiers et al. (2023) Zou et al. (2023) Cultrera et al. (2024)	Barua et al. (2016) Savran et al. (2018) Lenz et al. (2020) Chen et al. (2020b) Angelopoulos et al. (2020) Eddine and Dugelay. (2022) Ren et al. (2023a) Guo and Huang. (2023) Savran (2023) Himmi et al. (2024)

into frame-based representations or processed by incorporating spatio-temporal convolutional layer in the architecture. The effectiveness of using CNNs in event data processing has been well-established (Li et al., 2019; Banerjee et al., 2022; Becattini et al., 2022), allowing future studies to build upon these assumptions and further enhance event feature extraction (Kanamaru et al., 2023; Goyal et al., 2023). Actually, using event frame representations has allowed us to leverage all the advancements in research dedicated towards building well-established CNN architectures with pre-trained models trained on benchmark datasets. However, this conversion from events to frames might compromise some of the key advantages of event data such as high temporal resolution and sparsity. Moreover, most CNNs are inherently designed to process RGB data and might reply on texture and color information, which may not be present in event data. This might require heavy fine-tuning to perform well on event data.

(4) Transformers: Transformers moved away from traditional Recurrent Neural Networks (RNNs) and CNNs structures and revolutionized sequence-based data processing with their self-attention mechanism, which enables the allocation of higher weights to more significant information in the data. Depending on the tokenization strategy, events can be either processed in the form of frames, or in their raw format. Transformers allow for enhanced spatiotemporal feature extraction, effectively capturing the fine-grained dynamics of these types of data (Cultrera et al. (2024); Zou et al. (2023). These networks are one of the better performing as they capture the spatiotemporal dependencies between raw event data using attention mechanisms. They can directly operate on raw event streams or voxel representations, making them suitable for capturing both local and global patterns in human activity. This allows them to often achieve state-of-the-art results specially in action recognition or motion prediction tasks. However, they are very resource-intensive in terms computational complexity, large memory usage, and long training times, making them difficult to use in low-resource scenarios or in real-time embedded systems.

Each one of these architectures offers specific advantages, and the choice of which one to use is based on the specific characteristics of the event data and the requirements of the application. Table 2 provides a comprehensive classification of key works that use these architectures, showcasing the diversity of methodologies and their applications.

3.4 Event simulators

As highlighted in Figure 1, event cameras are increasingly adopted in new domains. Initially applied in driver monitoring and robotics, their use has expanded to human motion analysis, including gait and action recognition, and more recently to face biometrics, capturing subtle facial movements. However, the widespread adoption of event cameras is hindered by the lack of publicly available datasets. This limitation has driven the development of event camera simulators, which convert conventional RGB video data into synthetic event streams by replicating the characteristics of event data as accurately as possible.

The first notable simulator, ESIM, was introduced by Rebecq et al. (2018), providing a foundational framework for generating synthetic event streams. Gehrig et al. (2020) build upon this work and developed the Vid2E simulator by adding an upsampling step to RGB videos, enhancing the accuracy and the ability of models trained with synthetic data to generalize for real data. Hu et al. (2021) later introduced V2E, a versatile simulator capable of producing raw event streams alongside grayscale frames and corresponding text files, broadening its applicability. Lin et al. (2022) proposed the DVS-Voltmeter, which improves synthetic event data quality by modeling the behavior of the DVS sensor using a unified approach that incorporates its circuit properties. Most recently, Zhang et al. (2024b) proposed the V2CE simulator, which stands out as the most precise event data simulator to date. Numerous other simulators have been developed (Prophesee, 2023); Joubert et al. (2021); Han et al. (2024); Mueggler et al. (2017); SimulatorC. (2023); however, to the best of our knowledge, the ones discussed here are the most widely adopted and publicly available tools in the research community for generating synthetic event data.

The main advantage of event simulators is [removed repetition] that they provide researchers with a means to bypass the need for expensive hardware and complex data collection during early experimentation for validating hypotheses, and also, to train simple networks for preprocessing tasks, like detecting regions of interest in real event data (Becattini et al., 2022; Barchid et al., 2023). In addition, simulators overcome one of the main challenges of working with real event data: synchronization between RGB and event data, particularly in applications where the training data requires both modality pairs (Berlincioni et al., 2023).

However, despite their advantages, event simulators have notable drawbacks. First, they often fail to fully replicate the complexity and noise of real-world event data, which makes it harder for models trained on synthetic datasets to generalize effectively to real-world scenarios (Gehrig et al., 2020; Bi et al., 2020). Second, the quality of synthetic event streams heavily depends on the input RGB videos; if the videos lack high resolution and frame rate, the resulting event streams may miss critical details and temporal accuracy. Most critically, simulators inherently lose the high temporal resolution of event cameras, as they convert frame-based video inputs into events, reducing temporal precision from microseconds to frames per second, which can significantly affect tasks requiring fine-grained temporal information (Rebecq et al., 2018; Hu et al., 2021).

3.5 Datasets

The development and evaluation of event-based systems for human-centered applications heavily rely on publicly available datasets. In this survey, we categorize the datasets into two main groups based on their focus: body application datasets, which are designed for tasks such as action recognition, gait analysis, and human tracking, and face application datasets, which target applications like face detection, facial expression recognition, and anonymization. Each dataset is characterized by its unique properties, which we summarize in Tables 3, 4.

3.6 Real datasets

Table 3 summarizes the human-centered real event datasets, focusing on body and face applications respectively. Each section of the table provides detailed information about the datasets, including the year of publication, the dataset name, the number of videos, the number of participants, and other information that could be relative to the corresponding application.

For body applications, key datasets include *Action Dataset TUM* for action recognition, *DVS128 Gesture* for gesture recognition, and *DVS128-Gait-Day* for gait analysis. Table 3 also highlights newer datasets like *DailyDVS-200* and *THU-MV-E-ACT-50*, which include multimodal data and a large number of classes. As for Face applications, notable datasets include *DVS-Lip* for lip reading and *NEFER* for Micro-Expression Recognition (MER). The more recent *VETEX* dataset combines multimodal data, including RGB and event streams, to enhance facial analysis tasks.

3.7 Simulated datasets

Building on the challenges associated with real datasets—such as synchronization issues, noise, and limited diversity—we summarize in Table 4 the available synthetic datasets, categorizing them by their focus on body or face-related tasks. These datasets are generated in controlled environments, offering precise annotations and diverse scenarios that complement real-world data.

Unlike body applications, there are fewer publicly available synthetic datasets for face analysis. The currently available datasets are typically generated from the same benchmark RGB datasets, such as *e-CK+* and *e-MMI*, and have been simulated multiple times by different researchers using tools like V2E, often with different parameters (Verschae and Bugueno-Cordova, 2023; Barchid et al., 2023). As a result, the generated event data cannot be published as standalone datasets but are typically shared as simulation code. In cases like *NEFER* dataset, Berlincioni et al. (2023) utilized an unpublished simulated event dataset to train a face detector for creating synchronized bounding boxes, enabling event-based face detection on real data, even though the primary goal of the work was micro-expression recognition where a real dataset was collected. This has resulted in limited diversity in synthetic datasets for face-related tasks compared to body applications.

3.8 Data compression

Despite event cameras being energy efficient, one key challenge is the significant data volume they generate, especially in real-time applications like robotics and video surveillance, where embedded systems require efficient storage and processing. While the asynchronous nature of event data reduces redundancy compared to traditional video streams, the sheer volume of events captured during high-speed motion or complex scenes remains a bottleneck. Several works have proposed compression techniques to address this challenge, leveraging both traditional and deep-learning-based methods (Sezavar et al., 2024; Wang et al., 2023).

Approaches to compress event data generally fall into multiple categories. Traditional methods include transforming events into frame-like representations for compatibility with standard video coding techniques (Schiopu and Bilcu, 2022). Other methods, such as Spike coding, directly leverages the sparse and asynchronous nature of event streams to encode only significant changes, effectively reducing data size while preserving critical temporal information (Sengupta and Kasabov, 2017; Bi et al., 2018). More recent research explores geometric-based structures as proposed by Martini et al. (2022) where they introduce a point cloud-based compression method capable of both lossy and lossless operations, achieving efficient data reduction. Huang et al. (2023) worked on point-cloud compression, further demonstrating that compression at high ratios maintains performance for tasks such as object detection and image reconstruction. They were able to achieve a compression ratio of 5 with lossless point cloud coding and with zero accuracy degradation on recognition tasks. Some large companies such as Google also contributed to that domain. In particular, Google designed *Draco* which further extended previous methods by supporting additional attributes such as polarity, making it well-suited for event point clouds and stands out for its faster processing

TABLE 3 Real event-based datasets for human-centered applications.

Body datasets								
Year	Authors	Name	# Videos	# People	Modalities	Application	# Classes	DVS Resolution
2017	Amir et al. (2017)	DVS Gesture	1,342	29	EV	Action Recognition	11	128 × 128
2019	Miao et al. (2019)	Action Dataset TUM	291	15	EV	Action Recognition	10	346 × 260
2019	Calabrese et al. (2019)	DHP19	2,244	17	EV	Pose Estimation	-	346 × 260
2019	Wang et al. (2019)	DVS128-Gait-Day	4,000	20	EV	Gait Recognition	-	128 × 128
2019	Wang et al. (2019)	DVS128-Gait-Night	4,000	20	EV	Gait Recognition	-	128 × 128
2021	Liu et al. (2021)	DailyAction-DVS	1,440	15	EV	Action Recognition	12	346 × 260
2022	Eddine and Dugelay. (2022)	Gait3	168	56	RGB - EV - TH	Gait Recognition	-	346 × 260
2023	Gao et al. (2023)	THU-E-ACT-50	10,500	105	EV	Action Recognition	50	346 × 260
2023	Gao et al. (2023)	THU-E-ACT-50-CHL	2,330	18	EV	Action Recognition	50	346 × 260
2024	Gao et al. (2024)	THU-MV-E-ACT-50	31,500	105	EV	Action Recognition	50	346 × 260
2025	Wang et al. (2025)	DailyDVS-200	22,000	46	RGB - EV	Action Recognition	200	320 × 240
Face datasets								
Year	Authors	Name	# Videos	# People	Modalities	Application	DVS Resolution	
2016	Barua et al. (2016)	-	-	30	EV	Face Detection	128 × 128	
2019	Li et al. (2019)	-	34,000	34	EV-audio	Lip Reading	340 × 280	
2020	Angelopoulos et al. (2020)	-	24	24	EV	Eye gaze tracking	346 × 260	
2020	Chen et al. (2020a)	EDDD	260	26	EV	Drowsiness	346 × 260	
2020	Lenz et al. (2020)	-	48	10	EV	Face Detection	304 × 240	
2020	Chen et al. (2020b)	NeuroBiometric	180	45	EV	Authentication	346 × 260	
2022	Banerjee et al. (2022)	-	3,360	6	RGB - EV	Eye gaze tracking	346 × 260	
2022	Becattini et al. (2022)	-	455	25	RGB - EV	MER	640 × 480	
2022	Tan et al. (2022)	DVS-Lip	19,871	40	EV	Lip Reading	346 × 260	
2022	Moreira et al. (2022)	NVSFD	436	40	EV	Identity Recognition	1280 × 800	
2023	Bissarinova et al. (2023)	FES	~4,000	73	EV	Face Detection	480 × 360	
2023	Berlincioni et al. (2023)	NEFER	609	29	RGB - EV	MER	1280 × 720	
2023	Kanamaru et al. (2023)	-	1,500	20	EV	Lip Reading	340 × 280	
2024	Adra et al. (2024b)	VETEX	2,506	30	RGB - EV - TH	MER	346 × 260	

compared to other methods. Recent advancements also explore deep-learning-based solutions. Nguyen et al. (2021) proposed VoxelDNN, a model that captures the geometric structure of event data through convolutional networks, achieving both high compression efficiency and preservation of critical information.

The growing importance of event data compression is further highlighted by recent initiatives such as the JPEG XE standard, developed by the JPEG committee (Brites and Ascenso., 2024). JPEG XE focuses on creating a standardized framework for efficiently representing event-based vision data, ensuring interoperability

TABLE 4 Synthetic datasets for human-centered applications.

Body datasets						
Year	Authors	Name	# Videos	# People	Application	# Classes
2019	Wang et al. (2019)	EV-CASIA-B	8,184	124	Gait Recognition	-
2020	Bi et al. (2020)	HMDB51-DVS	6,766	-	Action Recognition	51
2020	Bi et al. (2020)	UCF101-DVS	13,320	-	Action Recognition	101
2022	Plizzari et al. (2022)	N-EPIC-Kitchens	64	-	Action Recognition	8
2023	Zou et al. (2023)	SynEventHPD	9,197	47	Pose Estimation	-
2023	Goyal et al. (2023)	eH36m	748	7	Pose Estimation	-

Face datasets					
Year	Authors	Name	# Videos	# People	Application
2022	Moreira et al. (2022)	SynFED	6,536	30	Identity Recognition
2023	Barchid et al. (2023)	ADFES	198	22	Face Expression Recognition
2023	Barchid et al. (2023)	Oulu-CASIA	480	80	Face Expression Recognition
2023	Barchid et al. (2023), Verschae and Bugueno-Cordova. (2023)	e-CK+	327	93	Face Expression Recognition
2023	Barchid et al. (2023), Verschae and Bugueno-Cordova. (2023)	e-MMI	2,900+	75	Face Expression Recognition
2023	Ryan et al. (2023)	-	-	5	Multitask Facial Analysis
2024	Tan et al. (2024)	DVS-LRW100	107,664	-	Lip Reading

between sensing, storage, and processing systems. This initiative reflects the increasing interest in event cameras within industry and research, as standardization efforts like these are critical for facilitating broader adoption. By targeting machine vision applications, this initiative addresses the unique challenges of event cameras, such as their sparse and asynchronous nature, while emphasizing their potential for real-world application.

The techniques presented above, collectively enable event cameras to manage large-scale data effectively, facilitating their integration into real-time systems while maintaining the benefits of event-based vision.

4 Applications

In this section, we discuss the human-centered state-of-the-art applications of event data, divided into two macro areas: body and face. Table 5 summarizes the applications addressed in the literature using event cameras, referencing the corresponding state-of-the-art works. It is important to note that some papers tackle more than one application, and thus a reference may appear in multiple categories. In each subsection, we provide a more in-depth analysis of these research areas, explaining the relevant models for each application.

4.1 Body

In this subsection, we detail the applications of event-based data that require information from the full body of a person: gait recognition, action recognition, human tracking, and pose estimation.

4.2 Gait recognition

One of the first human-centered applications of event-based camera, explored the feasibility of utilizing data obtained with this new sensor to address the classic problem of gait recognition. Gait recognition is a biometric technique aimed at identifying individuals based on their unique walking patterns. By mainly capturing motion with high temporal resolution and sparse data representation, researchers could effectively analyze and distinguish walking patterns to determine human identities.

The first work on event-based gait recognition was presented in 2019 by Wang et al. (2019). Due to the noisy and asynchronous nature of events, traditional vision-based gait recognition algorithms were unsuitable for such data. To address this challenge, they proposed a novel approach called EV-Gait. This method leverages motion consistency to effectively reduce noise in event streams and employs a deep neural network to recognize gait patterns from the asynchronous and sparse event data, making it specifically tailored to the capabilities and challenges of this technology.

Over time, various architectures have been proposed to tackle the task of event-based gait recognition. An early work by Sokolova and Konushin (2019) introduced a pipeline composed of five consecutive steps: visualization of the event stream, human figure detection, optical flow estimation, human pose estimation, and finally, gait recognition based on neural features. This approach achieved performance comparable to conventional methods using color videos. Another approach by Tao et al. (2024) utilized SNNs to process event data, introducing a domain-specific Locomotion-Invariant Representation (LIR). LIR replaced the static Cartesian

TABLE 5 The table presents applications of event cameras for human data along with an exhaustive selection of relevant works for each category. The applications are categorized into two main areas: face and body.

Body			
Human tracking	Gait recognition	Action recognition	Pose estimation
Eisl et al. (2023) Xu et al. (2020)	Wang et al. (2019) Sokolova and Konushin. (2019) Wang et al. (2021) Eddine and Dugelay. (2022) Fu and Yan. (2023) Tao et al. (2024)	Liu et al. (2021) Plizzari et al. (2022) Ren et al. (2023a) Ren et al. (2023b) de Blegiers et al. (2023) Gao et al. (2023) Gao et al. (2024) Vicente-Sola et al. (2025) Wang et al. (2025)	Sokolova and Konushin. (2019) Zou et al. (2023) Goyal et al. (2023) Kohyama et al. (2024)
Face			
Face detection	Identity recognition	Lip reading	Eye blinking & gaze
Barua et al. (2016) Lenz et al. (2020) Ryan et al. (2021) Bissarinova et al. (2023) Ryan et al. (2023) Himmi et al. (2024) Iddrisu et al. (2024b)	Chen et al. (2020b) Moreira et al. (2022)	Savran et al. (2018) Li et al. (2019) Rios-Navarro et al. (2023) Savran (2023) Kanamaru et al. (2023) Bulzomi et al. (2023)	Lenz et al. (2020) Chen et al. (2020b) Angelopoulos et al. (2020) Ryan et al. (2021) Banerjee et al. (2022) Ryan et al. (2023) Iddrisu et al. (2024b)

coordinates of the raw event camera data with a floating polar coordinate system centered on the motion axis, improving the representation's adaptability to dynamic scenarios. Further innovations in Fu and Yan (2023) include the use of hypergraph neural networks for gait recognition. This method employed an event flow downsampling module to reduce data volume without compromising discriminability, an event feature extraction module to convert events into graph nodes, and a spatiotemporal hypergraph convolution module to construct a hypergraph, extract spatiotemporal features, and obtain pedestrian gait features.

Comparative works have also emerged in the literature. In 2022, Eddine and Dugelay (2022) conducted experiments using a baseline algorithm based on gait energy images adapted to event-camera output. They compared this approach to results from RGB and thermal videos using the same algorithm, demonstrating a distinct advantage for event-based data. Wang et al. (2021) investigated different representations of event streams for deep neural network classifiers. They proposed novel event-based gait recognition approaches using two distinct representations: graph-based and image-like. These methods leveraged graph convolutional networks and convolutional neural networks, respectively, showcasing the versatility of event-based data for gait recognition.

4.3 Action recognition

Action recognition is a major research focus in computer vision due to its importance in applications such as security and human-computer interaction (Adra et al., 2024a). Research in this field has advanced with the use of bio-inspired event sensors which capture only the activity in their field of view and automatically differentiate the foreground from the background, making them ideal for recognizing human actions.

Liu et al. (2021) made an early attempt to apply motion information to event-based action recognition by extracting motion features from events, progressing from local to global perception. On the other hand, Ren et al. (2023b) introduced SpikePoint, a novel end-to-end point-based SNN architecture that processes event data as cloud data and converts them into spikes using rate coding. More recently, in 2025, Vicente-Sola et al. (2025) demonstrated that spiking neurons can enable temporal feature extraction in feed-forward neural networks without requiring recurrent synapses, and how recurrent SNNs can achieve performance comparable to LSTMs with fewer parameters, validating their approach in action recognition.

Beyond SNNs, other architectures have been explored to create more lightweight models. de Blegiers et al. (2023) proposed a video transformer-based framework that acquires spatial embeddings per event-frame and utilizes a temporal self-attention mechanism. This approach separates spatial and temporal operations, making the video transformer more computationally efficient than other video transformers. Ren et al. (2023a) proposed a point cloud-based method for action recognition using event data, featuring a hierarchical structure that distinguishes local and global features. Their model is lightweight, thanks to the application of tensor decomposition to compress the data.

In more recent works, Gao et al. (2023) introduced EV-ACT, an event-based action recognition framework that uses a slow-fast network to fuse motion and appearance-related features. One of their key contributions is the Learnable Multi-Fused Representation, which integrates multiple event representations, such as time surfaces, event frames, and event count, into a single embedding. In an extension of their work, Gao et al. (2024) proposed HyperMV, a multi-view event-based action recognition framework utilizing hypergraphs and a hypergraph neural network to capture relationships across viewpoint and temporal features.

Additionally, Plizzari et al. (2022) proposed two new strategies; directly processing event-camera data with traditional video-processing architectures and using event data to extract optical flow information. They also compared the performance of different pairings of event, RGB, and optical flow. Another comparative study was conducted by Wang et al. (2025), where in addition to introducing their benchmark database, *DailyDVS-200*, they evaluated it using 12 event-based architectures for action recognition.

For this application in particular, we believe that there are sufficient resources and a well-established benchmark dataset that allows for a fair evaluation across the different network types discussed in Section 3.3. Therefore, we trained each of these networks on the *DVS Gesture* dataset (mentioned in Table 3) and compared their performance in terms of accuracy, training time, and number of parameters to assess both performance and complexity. The results are shown in Table 6. Despite achieving the best performance, event-based transformers are also the most complex and require almost 30 times more training time than the SNN or Graph CNN. The SNN model, on the other hand, is the lightest model in terms of parameters and training time but slightly less accurate than both the transformer and the 3D CNN. Interestingly, the 3D CNN stands as a middle ground, offering high accuracy with a level of complexity that is higher than SNNs but still significantly lower than transformers. This highlights the importance of selecting the right trade-off between performance and complexity based on the specific requirements of our target application.

4.4 Pose estimation

Human Pose Estimation refers to the identification of key body joints in a human and plays a vital role in many human-centered tasks (Rafi et al., 2020). In fields like robotics, IoT, and smart home applications, pose estimation is the initial step that supports subsequent processes such as action recognition, posture analysis, and emotion and intent detection (Goyal et al., 2023).

In 2019, Sokolova and Konushin (2019) attempted the first pose estimation using event-based human data. Although their primary goal was human gait recognition, they also addressed several auxiliary challenges, such as moving object detection and human pose estimation in event-based video sequences. Their model focused on detecting areas of interest and subsequently computing optical flow to estimate the positions of key pose points. In more recent approaches, Goyal et al. (2023) presented a system for high-frequency 2D human pose estimation for a single person. The core of their approach is the use of a lightweight, image-like event representation that resolves the issue of static body parts disappearing and allows pre-training on widely available frame-based datasets with high-accuracy ground truth, followed by fine-tuning on native event-camera datasets.

Zou et al. (2023) introduced the first end-to-end method for 3D human pose tracking using only event data, leveraging Spiking Neural Networks. In 2024, Kohyama et al. (2024) proposed a method that exclusively uses event data to create 3D voxel representations by moving an event camera around a stationary body Kohyama et al. (2024). This method reconstructs human pose and mesh through attenuated rays while fitting statistical body models to preserve high-frequency details.

4.5 Human tracking

In recent years, Mitrokhin et al. (2018) and Ramesh et al. (2020) have proposed several approaches for event-based object tracking, primarily focusing on tracking objects with simple shapes. Building on this, a new research direction has emerged, addressing the relatively novel problem of tracking 3D human inputs solely based on event streams from an event camera, thereby completely eliminating the need for additional dense input images. In 2023, Eisl et al. (2023) presented a novel framework for tracking humans using a single event camera, comprising three main components. First, a Graph Neural Network was trained to identify a person within the stream of events. To preserve the sparse nature of the event data and leverage its high temporal resolution, batches of events are represented as spatio-temporal graphs. Next, the person was localized in a weakly-supervised manner via Class Activation Maps to their graph-based classification model, eliminating the need for ground truth human positions during training followed by a Kalman filter for tracking.

Existing works in pose tracking either require the presence of additional grayscale images to establish a reliable initial pose as it is the case in Xu et al. (2020) or disregard temporal dependencies altogether by collapsing segments of event streams into static event frames like in Rudnev et al. (2021). Zou et al. (2023) introduced a dedicated end-to-end sparse deep learning approach for event-based 3D human pose tracking where the task is achieved without any reliance on frame-based images. Their method is based on a Spiking Neural Network, with the incorporation of a Spike-Element-Wise ResNet and a novel Spiking Spatiotemporal Transformer.

4.6 Face

In this subsection, we analyze the use of event-based data for tasks that involve solely the face of an individual. Those applications are face detection, identity recognition, lip-reading, eye blinking and gaze analysis and microexpression and emotion recognition.

4.7 Face detection

An early application of event-based facial data was face detection, a task that involves identifying and locating human faces within an image or video stream. Face detection serves as a foundational step for various facial applications, including identity recognition, soft biometric estimation, and behavior analysis.

In 2016, Barua et al. (2016) developed a pioneering face detection model based on translating event streams into large-scale images using a patch-based approach. Their method involved learning a sparse dictionary of patches to reconstruct both simulated and real event data, even in noisy conditions. Their event-based face detection framework achieved results comparable to the traditional Viola-Jones face detector (Viola and Jones, 2001). Bissarinova et al. (2023) proposed an

TABLE 6 Action recognition performance comparison on the DvsGesture dataset.

Model	Authors	Accuracy	Training time	# Parameters
Graph CNN	Wang et al. (2021)	81.94%	36.8 min	7.6M params
Spiking Neural Network	Fang et al. (2021)	93.4%	**24.1 min**	130.4K params
Event-based Transformer	de Blegiers et al. (2023)	**97.91%**	785 min	113.4M params
3D CNN (using reconstructed frames)	Tran et al. (2015)	95.26%	270 min	78.04M params

Bold values indicate the best results in each column, i.e., the highest accuracy and the lowest training time.

architecture that utilizes events accumulated over time and incorporates past event information for effective face detection. They presented 12 models trained on their dataset to predict bounding boxes and facial landmark coordinates. Additionally, they showcased real-time face detection capabilities using event-based cameras and their models. More recently, Himmi et al. (2024) defined the concept of multispectral events, capturing data across multiple spectral bands to enhance event-based face detection. They demonstrated that multispectral events significantly improve face detection performance compared to monochromatic grayscale events, surpassing even conventional multispectral image performance.

As face detection often precedes other facial processing tasks, several studies have combined face detection with additional applications. In 2020, Lenz et al. (2020) introduced the first purely event-based method for face detection, relying on eye-blink detection. They analyzed the temporal signature of eye blinks and employed a Gaussian tracker to statistically measure pixel activity in the event stream. In 2021, Ryan et al. (2021) proposed GR-YOLO, a novel neural network for face and eye detection using event cameras, specifically in driver monitoring systems. Their architecture, based on YOLOv3-tiny, incorporated a fully convolutional gated recurrent unit layer. By 2023, Ryan et al. (2023) extended this work by introducing a two-stage event-based multi-task facial analytics framework. The first stage used a CNN to locate and track faces and eyes, while the second stage employed another CNN to estimate head pose, eye gaze, and occlusions within a multi-task learning setup. Building on previous work, Iddrisu et al. (2024b) utilized a Temporal Binary Representation of event data and trained a GR-YOLO model, comparing its performance to YOLOv8 for face and eye detection tasks.

4.8 Identity recognition

Identity recognition via face images is a biometric technology that identifies or verifies individuals based on their distinct facial features. This task is critical for numerous applications requiring reliable verification or identification, as the face is a unique and easily accessible trait crucial for enhancing security systems. So far in the literature, identity recognition from event data has been performed with the help of other auxiliary tasks such as eye blink characterization or facial dynamics derived from speech.

In 2021, Chen et al. (2020b) proposed the first neuromorphic, event-based biometric authentication system. Their method for identity recognition relied on eye blink characterization. They defined a set of biometric features describing the motion, speed, energy, and frequency signals of eye blinks, leveraging the microsecond temporal resolution of event densities. Using these features, they trained both an ensemble model and a non-ensemble model with their NeuroBiometric dataset for biometric authentication. In a subsequent work, Moreira et al. (2022) explored the potential of event sensors for identity recognition through a novel facial characteristic: facial dynamics derived from speech. They also validated the contribution of facial motion to human face identity categorization. Their approach involved aggregating events into frames, normalizing them, and grouping them into so-called "face tokens," which were then processed by a spatio-temporal 3D CNN to extract insights about the individual's identity.

4.9 Lip reading

Voice Activity Detection (VAD) is a technique used to identify and isolate segments of speech within an audio stream. Event cameras, with their high temporal resolution and ability to capture micro-movements, are particularly beneficial for this task. By accurately detecting subtle mouth movements, event data can enhance the precision of VAD, as well as related applications like lip-reading, where understanding spoken language relies on analyzing lip motions.

Savran et al. (2018) explored for the first time voice activity detection (VAD) using event data. In their VAD pipeline, they leveraged event-based facial data by adding an initial module in their pipeline where lip activity was filtered spatio-temporally and then detected jointly through probabilistic estimation. In a later work, Savran (2023) continued their research proposing an event intensity-based method for VAD by designing a fully convolutional network to segment vocally active durations efficiently. In their approach, the raw event sequence was first processed to ensure that voice-related temporal information was preserved in a low-dimensional representation. Subsequently, a fully convolutional VAD network was constructed to carry out the detection task. In 2023, Kanamaru et al. (2023) presented an event camera-based lip-reading method for isolated single-sound recognition. Their pipeline included imaging from event data, face and facial feature detection, and recognition using a Temporal Convolutional Network (TCN). Their findings demonstrated that event-based cameras achieved higher lip-reading accuracy than traditional frame-based cameras. Furthermore, the authors showed that combining two modalities, the frame-based camera and the event-based camera, yielded higher accuracy than using either modality alone. In the same year, Bulzomi et al. (2023)

proposed the first SNN model for event-based lip reading, achieving competitive results compared to state-of-the-art artificial neural networks.

An innovative approach by Li et al. (2019) combined video and audio data for the first time. The authors introduced a lip-reading deep neural network that fused the asynchronous spiking outputs of two bio-inspired silicon multimodal sensors: the Dynamic Vision Sensor and the Dynamic Audio Sensor. Their classification process, based on event-based features generated from the spikes of these sensors, was tested on the GRID visual-audio lipreading dataset. Similarly, Rios-Navarro et al. (2023). Utilized CNNs to process visual and auditory information in their self-collected dataset, which involved participants speaking a set of words. The visual information was derived from lip movements captured by event cameras as the subjects articulated words. The event activity was converted into histograms, which a CNN further processed.

4.10 Eye blinking and gaze analysis

Eye movement has been extensively studied in the biometrics community due to its potential for applications in authentication, gaze tracking, and behavioral analysis. Indeed, when addressing eye-blinking characterization from event-based data, researchers often solve this problem as an auxiliary task for other major objectives. Lenz et al. (2020) implemented a low-power human eye-blink detection method designed to exploit the high temporal precision provided by event-based cameras. Similarly, Chen et al. (2020b) developed an authentication system based on eye blinks captured with an event camera, achieving high accuracy with computationally simple processes.

A different area of focus is gaze and eye tracking. In 2022, Angelopoulos et al. (2020), defined a pipeline for gaze tracking that combined frames recorded at a fixed sampling rate with asynchronous events capturing eye motion at high speed. Their method outputs a gaze point derived from an estimate of the pupil, forming an almost continuous tubular structure that outlined the pupil's movement. Banerjee et al. (2022) proposed a novel event-encoding technique that converted motion event logs into six-channel images. They then designed a CNN to predict gaze using the encoded events from the event camera. In another study, Iddrisu et al. (2024b) employed an event simulator to convert RGB videos into event-based data. Their approach involved accumulating events into binary frames and aggregating these frames into a single one to enhance the density and quality of the simulated data. They subsequently compared the performance of different state-of-the-art models using the generated event data.

In 2021, Ryan et al. (2021) leveraged event-based data to create a low-energy consumption model for simultaneously detecting and tracking faces and eyes, specifically for driver monitoring applications. They developed a customized fully convolutional neural network for this purpose. Later, in 2023, Ryan et al. (2023) extended their work for the same application by designing a multitask neural network for real-time facial analysis. This new model simultaneously estimated head pose, eye gaze, and facial occlusions. It was trained on synthetic data and evaluated in real-world scenarios.

4.11 Micro-expressions

Facial Emotion Recognition (FER) is a technology that analyzes facial expressions from static images and videos to infer a person's emotional state. Recent advancements in the FER domain have focused on estimating microexpressions, subtle and rapid facial movements often performed involuntarily, due to their strong connection with emotions as defined by the Facial Action Coding System.

Becattini et al. (2022) pioneered the application of event cameras for FER using synthetic event data. Leveraging an event-camera simulator, they generated synthetic event streams and transferred face bounding boxes onto the data. Cropped face sequences were then processed by a CNN, followed by a long short-term memory network to account for the temporal dimension. In 2023, Barchid et al. (2023) introduced "Spiking-FER," a deep convolutional SNN inspired by ResNet18, achieving superior performance compared to traditional visible-domain methods. Berlincioni et al. (2023) classified microexpressions into three categories neutral, positive, and negative using a baseline 3D-CNN. Similarly, Guo and Huang (2023) proposed a lightweight approach utilizing a global-local event feature fusion network, which merged local count images with global dense optical flow to extract deeper features for FER.

In 2024, three studies further advanced the use of event data for microexpression estimation. Xiao et al. (2024) developed a system with two key components: the Event-Enhanced Motion Extractor, which amplified subtle movements, and the Event-Guided Attention module, which focused on crucial facial regions for microexpression analysis. Cultrera et al. (2024) introduced the first video transformer model for action unit classification from event streams, significantly improving accuracy. Finally, Adra et al. (2024b) conducted experiments on their novel dataset revealing that thermal and event-based modalities outperformed visible-spectrum cameras for microexpression recognition. Although thermal images provided the best performance under varying illumination conditions, event data also demonstrated strong capabilities, as its high temporal resolution proved more effective at capturing small facial movements than traditional RGB cameras.

4.12 Discussion

In this section, we have extensively reviewed the various applications of data obtained from neuromorphic cameras in human-centered contexts. Neuromorphic human analysis is a relatively new field of research. Nonetheless, several studies have highlighted the effectiveness of neuromorphic cameras for a variety of applications related to both the human body and face, offering notable advantages compared to traditional computer vision techniques. For instance, as demonstrated in Table 7, neuromorphic cameras show significant improvements in tasks such as action and microexpression on recognition when compared to RGB-based methods. However, for other tasks, the reported improvements are marginal or even negligible. For example, in gait recognition, the observed gains are minimal, and

TABLE 7 This table presents a summary of the works included in this survey that compare their event-based networks with RGB-trained models. Works are classified by their target application and the authors, year, and any reported improvement of event-based methods over RGB, if applicable are reported.

	Authors	Findings	Improvements of event
Gait recognition	Wang et al. (2019)	For viewing angles 72, 90 and 108 EV-Gait performs better than RGB based approaches	3% increase in accuracy
	Sokolova and Konushin. (2019)	Similar perfomances reported for Event-based and RGB approaches	-
	Wang et al. (2021)	For viewing angle 90° EV-Gait-Graph performs better than RGB based approaches	0.5% increase in accuracy
	Eddine and Dugelay. (2022)	Advantage of event data over RGB and thermal for gait recognition	2% increase in accuracy
	Tao et al. (2024)	They report the advantage of event over RGB across all different rotation angles for gait recognition	Up to 14% increase in accuracy
Action recognition	Plizzari et al. (2022)	Event data can surpass RGB for action recognition in unseen scenarios on test data	4% increase in accuracy
	de Blegiers et al. (2023)	Event surpass RGB action recognition models in different setups	Up to 14% increase in accuracy
Pose estimation	Goyal et al. (2023)	Pose estimation from event data surpasses RGB data	Up to 5% increase in accuracy
	Kohyama et al. (2024)	Event does not suffer from motion blur as RGB does for 3D-based pose estimation	Error (in mm) is divided by 5 in certain scenarios
Face detection	Barua et al. (2016)	Comparable results to Viola-Jones face detector	-
	Ryan et al. (2023)	Traditional RGB models perform better on RGB images than on their simulated event data counterpart	-
Lip reading	Kanamaru et al. (2023)	They combined event and RGB modalities for lip reading	-
Microexpression and emotion recognition	Becattini et al. (2022)	Event data overperforms RGB for detecting three types of expressions: Positive, Neutral, Negative	Up to 9% increase in accuracy
	Berlincioni et al. (2023)	Event overperforms RGB in the prediction of seven different emotions	Up to 15% increase in accuracy
	Xiao et al. (2024)	Event and RGB are merged as input to the network	1% increase in accuracy
	Cultrera et al. (2024)	For the estimation of some action units event data delivers better performance	For 6 out of 24 action units event data is more accurate
	Adra et al. (2024b)	Event data gives more information than RGB for microexpression recognition	Up to 13% increase in accuracy

for applications like face detection and lip reading, the performances of neuromorphic cameras are often comparable to those achieved with RGB-based approaches. This suggests that, while event data holds promise, its benefits are not yet universally realized across all applications. Additionally, a critical limitation in the current state of research is the lack of standardized benchmark datasets. There is a tendency for researchers to report results on newly created databases, often without direct comparison to existing datasets, making it challenging to objectively evaluate and compare progress across studies.

Moreover, over the past decade, computer vision has made remarkable advancements in their AI-based architectures such as CNNs and vision transformers. These models are highly optimized to extract detailed and meaningful information from RGB data, resulting in state-of-the-art performances across a wide range of tasks while other modalities such as event data have not received the same level of attention in model development. Current neural networks are not inherently designed to fully leverage the unique characteristics of these modalities, which limits their potential. While event-based sensors may provide additional, task-specific information that could be more useful than RGB data in certain scenarios, the lack of tailored architectures results in RGB data often outperforming these modalities. This is evident in the treatment of event data, where, as noted in Table 2, processing frequently involves converting event streams into representations that mimic the structure of RGB frames to enable their use with pre-existing CNN architectures.

5 Conclusion and future perspectives

Human-centered applications are one of the foundations of computer vision research, addressing challenges and

opportunities in diverse areas such as surveillance, biometric authentication, autonomous driving, and behavioral analysis. While traditional frame-based methods using RGB cameras have achieved remarkable advancements, their limitations in temporal resolution, motion blur, and low-light conditions have increased the interest in neuromorphic cameras. These sensors represent a paradigm shift, capturing asynchronous pixel intensity changes that provide high temporal resolution, robustness in challenging environments, and reduced computational demands.

This survey offers a comprehensive overview of the progress and potential of event-based cameras for human-centered applications. By categorizing advancements in human body- and face-related tasks, we highlight the progress made in recent years, emphasizing the strengths and innovations of architectures leveraging event-based data. Presenting the current state of the art, identifying challenges, and suggesting future directions, this survey aims to guide researchers in exploiting the potential of event-based cameras for human-centered applications.

By analyzing the properties of event data, we observed how it is uniquely suited for machines due to characteristics such as high temporal resolution and low latency. These attributes provide AI models with richer and more precise information compared to traditional RGB data. However, event data is less intuitive and interpretable for humans, making it better aligned with AI capabilities than with human understanding. Moreover, in our state-of-the-art review, we identified a significant drawback in many of the models presented: researchers often focus on demonstrating the suitability of event data for specific applications and few works conduct thorough comparisons with RGB-based methods. This lack of direct performance comparisons highlights the early developmental stage of event-based models, which in some cases have yet to reach the maturity required for widespread adoption. Additionally, we want to remark that the future of acquisition sensors remains uncertain, particularly as generative AI continues to advance enabling the creation of highly accurate synthetic images that can be used to train high-precision networks without compromising individual privacy. Such synthetic datasets have the potential to complement or even replace real-world event-based data in certain scenarios.

However, looking ahead, different promising directions for future research emerge. So far, one of the most critical limitations is the lack of a standardized event-based datasets, particularly those that capture real-world conditions in diverse human-centered scenarios. Several studies in the area of face or body analysis rely on custom or unpublished datasets, making it difficult to benchmark proposed event-based methods. To support fair comparison and reproducibility across studies, future efforts should focus on releasing publicly available datasets, potentially across multiple and synchronized modalities and covering a wide range of possible scenarios. A critical challenge in this direction is the temporal alignment between event data and other sensor streams. Since events are fundamentally asynchronous, precise synchronization is not trivial, and many existing datasets either lack synchronized modalities or provide only vaguely aligned data.

Another key limitation is how event data is currently processed since the majority of existing approaches tend to convert the asynchronous event stream into frame-like representations, primarily to make it compatible with conventional CNN architectures. While this strategy can simplify the usage of the data to downstream learning tasks, it comes at the cost of not fully exploiting the intrinsic properties of event data. Future research should therefore aim to improve event representation and processing methods, along with architectures that natively operate on raw events (e.g., spiking neural networks, graph-based models)

One other interesting point that deserves deeper investigation is the role of privacy in event-based vision. Recent works have demonstrated that it is possible to reconstruct intensity maps from events, potentially revealing sensitive information, including facial identity. This is considered as a privacy attack against event data known as a reconstruction attack. Several studies in the literature have already proposed strategies for anonymizing or encrypting event streams, but we are far from seamlessly integrating them in the majority of learning pipelines. Research efforts should also aim to design privacy-preserving architectures, especially in surveillance-related applications.

Finally, while some studies already tried to explore the combination of event data with other sensory inputs (e.g., integrating events and RGB frames), systematic approaches to multimodal sensor fusion are still rare. Developing systems that integrate event-based and conventional modalities offer greater potential, as these systems could exploit the complementary strengths of event cameras and other mainstream sensors, enhancing the overall performance. In this context, attention-based fusion strategies or encoder-decoder architectures with shared latent spaces are promising options to be explored.

Author contributions

MA: Writing – original draft, Writing – review and editing. SM: Writing – original draft, Writing – review and editing. NM-H: Writing – original draft, Writing – review and editing. J-LD: Conceptualization, Project administration, Supervision, Validation, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research is a part of the HEIMDALL project, funded by the BPI as part of the AAP I-Demo. Additionally, the work was supported by the European Union's Horizon Europe research and innovation program under Grant Agreement No 101094831 for the Converge-Telecommunications and Computer Vision Convergence Tools for Research Infrastructures project.

Acknowledgments

A preprint of this work is available on arXiv (Adra et al., 2025, arXiv:2502.18490).

Conflict of interest

Author MA was employed by company GTD International.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Adra, M., Dugelay, J.-L., and Ichard, C. (2024a). "A comparative study: error analysis and model efficiency in event-based action recognition," in 2024 IEEE Thirteenth International Conference on Image Processing Theory, Tools and Applications (IPTA) (IEEE), 01–06.

Adra, M., Mirabet-Herranz, N., and Dugelay, J.-L. (2024b). "Beyond rgb: tri-modal microexpression recognition with rgb, thermal, and event data," in *2024 twenty-seventh international conference on pattern recognition (ICPR)* (Springer).

Adra, M., Melcarne, S., Mirabet-Herranz, N., and Dugelay, J.-L. (2025). Event-based solutions for human-centered applications: a comprehensive review. arXiv Prepr. *arXiv: 2502*, 18490). doi:10.48550/arXiv.2502.18490

Ahmad, S., Bue, A. D., and Morerio, P. (2023). "Person re-identification without identification via event anonymization," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 12115–12125. doi:10.1109/ICCV51070.2023.01022

Ahmad, S., Morerio, P., and Del Bue, A. (2024). Event anonymization: privacy-preserving person re-identification and pose estimation in event-based vision. *IEEE Access* 12, 66964–66980. doi:10.1109/ACCESS.2024.3399539

Al-Obaidi, S. M. (2020). *Privacy aware human action recognition: an exploration of temporal salience modelling and neuromorphic vision sensing*. UK: University of Sheffield. Ph.D. thesis.

Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Nolfo, C. D., et al. (2017). "A low power, fully event-based gesture recognition system," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7243–7252. doi:10.1109/CVPR.2017.764

Angelopoulos, A. N., Martel, J. N., Kohli, A. P., Conradt, J., and Wetzstein, G. (2020). Event-based near-eye gaze tracking beyond 10,000 hz. *IEEE Trans. Vis. Comput. Graph.* 27, 2577–2586. doi:10.1109/tvcg.2021.3067784

Banerjee, A., Prasad, S. S., Mehta, N. K., Kumar, H., Saurav, S., and Singh, S. (2022). "Gaze detection using encoded retinomorphic events," in *International conference on intelligent human computer interaction* (Springer), 442–453.

Barchid, S., Allaert, B., Aissaoui, A., Mennesson, J., and Djeraba, C. C. (2023). "Spiking-fer: spiking neural network for facial expression recognition with event cameras," in 20th International Conference on Content-based Multimedia Indexing, 1–7. doi:10.1145/3617233.3617235

Barua, S., Miyatani, Y., and Veeraraghavan, A. (2016). "Direct face detection and video reconstruction from event cameras," in 2016 IEEE winter conference on applications of computer vision (WACV) (IEEE), 1–9.

Becattini, F., Palai, F., and Bimbo, A. D. (2022). "Understanding human reactions looking at facial microexpressions with an event camera," in IEEE Transactions on Industrial Informatics, 9112–9121. doi:10.1109/TII.2022.3195063*IEEE Trans. Ind. Inf.* 18.

Becattini, F., Berlincioni, L., Cultrera, L., and Del Bimbo, A. (2025). Neuromorphic face analysis: a survey. *Pattern Recogn. Lett.* 187, 42–48. doi:10.1016/j.patrec.2024.11.009

Bendig, K., Schuster, R., Thiemer, N., Joisten, K., and Stricker, D. (2024). Anonynoise: anonymizing event data with smart noise to outsmart re-identification and preserve privacy. *arXiv preprint arXiv:2411.16440* Accepted at WACV 2025

Berlincioni, L., Cultrera, L., Albisani, C., Cresti, L., Leonardo, A., Picchioni, S., et al. (2023). "Neuromorphic event-based facial expression recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4109–4119. doi:10.1109/cvprw59228.2023.00432

Bi, Z., Dong, S., Tian, Y., and Huang, T. (2018). "Spike coding for dynamic vision sensors," in 2018 Data Compression Conference, 117–126. doi:10.1109/dcc.2018.00020

Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., and Andreopoulos, Y. (2019). "Graph-based object classification for neuromorphic vision sensing," in IEEE International Conference on Computer Vision (ICCV), 491–501. doi:10.1109/iccv.2019.00058

Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., and Andreopoulos, Y. (2020). "Graph-based spatio-temporal feature learning for neuromorphic vision sensing," in IEEE Transactions on Image Processing, 9084–9098. doi:10.1109/tip.2020.3023597*IEEE Trans. Image Process.* 29.

Bissarinova, U., Rakhimzhanova, T., Kenzhebalin, D., and Varol, H. A. (2023). Faces in event streams (fes): an annotated face dataset for event cameras. *Prepr. Authorea*. doi:10.3390/s24051409

Brites, C., and Ascenso, J. (2024). Neuromorphic vision data coding: classifying and reviewing. *arXiv Prepr. arXiv:2405.07050*. doi:10.48550/arXiv.2405.07050

Bulzomi, H., Schweiker, M., Gruel, A., and Martinet, J. (2023). "End-to-end neuromorphic lip-reading," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4101–4108. doi:10.1109/cvprw59228.2023.00431

Calabrese, E., Taverni, G., Easthope, C. A., Skriabine, S., Corradi, F., Longinotti, L., et al. (2019). "Dhp19: dynamic vision sensor 3d human pose dataset," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 1695–1704. doi:10.1109/cvprw.2019.00217

Cazzato, D., and Bono, F. (2024). An application-driven survey on event-based neuromorphic computer vision. *Information* 15, 472. doi:10.3390/info15080472

Chakravarthi, B., Verma, A. A., Daniilidis, K., Fermuller, C., and Yang, Y. (2024). Recent event camera innovations: a survey. *arXiv Prepr. arXiv:2408.13627*. doi:10.48550/arXiv.2408.13627

Chen, S., and Guo, M. (2019). "Live demonstration: celex-v: a 1m pixel multi-mode event-based sensor," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (IEEE), 1682–1683.

Chen, G., Hong, L., Dong, J., Liu, P., Conradt, J., and Knoll, A. (2020a). Eddd: event-Based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor. *IEEE Sensors J.* 20, 6170–6181. doi:10.1109/jsen.2020.2973049

Chen, G., Wang, F., Yuan, X., Li, Z., Liang, Z., and Knoll, A. (2020b). "Neurobiometric: an eye blink based biometric authentication system using an event-based neuromorphic vision sensor," in IEEE/CAA Journal of Automatica Sinica, 206–218. doi:10.1109/jas.2020.1003483*IEEE/CAA J. Autom. Sin.* 8.

Cultrera, L., Becattini, F., Berlincioni, L., Ferrari, C., and Del Bimbo, A. (2024). Spatio-temporal transformers for action unit classification with event cameras. *arXiv Prepr. arXiv:2410.21958*. doi:10.48550/arXiv.2410.21958

de Blegiers, T., Dave, I. R., Yousaf, A., and Shah, M. (2023). "Eventtransact: a video transformer-based framework for event-camera based action recognition," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1–7. doi:10.1109/iros55552.2023.10341740

Delilovic, N., and Salaj, D. (2021). *Bio-inspired neuromorphic ai methods enables privacy respecting security and surveillance*. Belgrade, Serbia: Transactions on Advanced Research.

Deng, Y., Chen, H., Liu, H., and Li, Y. (2022). "A voxel graph CNN for object classification with event cameras," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1162–1171. doi:10.1109/CVPR52688.2022.00124

Dong, Y., Li, Y., Zhao, D., Shen, G., and Zeng, Y. (2023). "Bullying10k: a large-scale neuromorphic dataset towards privacy-preserving bullying recognition," in Thirty-

seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Du, B., Li, W., Wang, Z., Xu, M., Gao, T., Li, J., et al. (2021). Event encryption for neuromorphic vision sensors: framework, algorithm, and evaluation. *Sensors (Basel)* 21, 4320. doi:10.3390/s21134320

Eddine, M. J., and Dugelay, J.-L. (2022). "Gait3: an event-based, visible and thermal database for gait recognition," in 2022 International Conference of the Biometrics Special Interest Group (BIOSIG) (IEEE), 1–5.

Eisl, D., Herzog, F., Dugelay, J.-L., Apvrille, L., and Rigoll, G. (2023). "Introducing a framework for single-human tracking using event-based cameras," in 2023 IEEE International Conference on Image Processing (ICIP) (IEEE), 3269–3273.

Eshraghian, J. K., Ward, M., Neftci, E. O., Wang, X., Lenz, G., Dwivedi, G., et al. (2023). "Training spiking neural networks using lessons from deep learning," in Proceedings of the IEEE, 1016–1054. doi:10.1109/JPROC.2023.3308088*Proc. IEEE* 111.

Etienne-Cummings, R., and der Spiegel, J. V. (1996). Neuromorphic vision sensors. *Sensors Actuators A Phys.* 56, 19–29. doi:10.1016/0924-4247(96)01277-0

Fan, L., Yang, J., Wang, L., Zhang, J., Lian, X., and Shen, H. (2025). Efficient spiking neural network for rgb–event fusion-based object detection. *Electron. (Basel).* 14, 1105. doi:10.3390/electronics14061105

Fang, W., Yu, Z., Chen, Y., Huang, T., Masquelier, T., and Tian, Y. (2021). "Deep residual learning in spiking neural networks," in *Advances in neural information processing systems*. A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan

Fu, L., and Yan, S. (2023). Hypergraph neural network for gait recognition based on event camera. Third Int. Conf. Adv. Algorithms Signal Image Process. (AASIP 2023) *(SPIE)* 12799, 1151–1155. doi:10.1117/12.3006325

Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., et al. (2022). "Event-based vision: a survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 154–180. doi:10.1109/tpami.2020.3008413*IEEE Trans. Pattern Anal. Mach. Intell.* 44.

Gao, Y., Lu, J., Li, S., Ma, N., Du, S., Li, Y., et al. (2023). "Action recognition and benchmark using event cameras," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 14081–14097. doi:10.1109/tpami.2023.3300741*IEEE Trans. Pattern Anal. Mach. Intell.* 45.

Gao, Y., Lu, J., Li, S., Li, Y., and Du, S. (2024). "Hypergraph-based multi-view action recognition using event cameras," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 6610–6622. doi:10.1109/tpami.2024.3382117*IEEE Trans. Pattern Anal. Mach. Intell.* 46.

Gehrig, D., Gehrig, M., Hidalgo-Carrió, J., and Scaramuzza, D. (2020). "Video to events: recycling video datasets for event cameras," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3586–3595.

Ghosh-Dastidar, S., and Adeli, H. (2009). Spiking neural networks. *Int. J. Neural Syst.* 19, 295–308. doi:10.1142/S0129065709002002

Goyal, G., Di Pietro, F., Carissimi, N., Glover, A., and Bartolozzi, C. (2023). "Moveenet: online high-frequency human pose estimation with an event camera," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4024–4033.

Guo, C., and Huang, H. (2023). "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, 17–24. doi:10.1145/3607829.3616446

Han, B., Fu, Q., and Zhang, X. (2023). Towards privacy-preserving federated neuromorphic learning via spiking neuron models. *Electronics* 12, 3984. doi:10.3390/electronics12183984

Han, H., Lyu, J., Li, J., Wei, H., Li, C., Wei, Y., et al. (2024). "Physical-based event camera simulator," in European Conference on Computer Vision, 19–35. doi:10.1007/978-3-031-72995-9_2

Himmi, S., Parret, V., Chhatkuli, A., and Van Gool, L. (2024). "Ms-evs: multispectral event-based vision for deep learning based face detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 616–625.

Hu, Y., Liu, S. C., and Delbruck, T. (2021). "v2e: from video frames to realistic dvs events," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1312–1321. doi:10.1109/cvprw53098.2021.00144

Huang, B., Lazzarotto, D., and Ebrahimi, T. (2023). Evaluation of the impact of lossy compression on event camera-based computer vision tasks. Appl. Digital Image Process. XLVI *(SPIE)* 12674, 89–100. doi:10.1117/12.2676419

Iddrisu, K., Shariff, W., Corcoran, P., O'Connor, N. E., Lemley, J., and Little, S. (2024a). Event camera-based eye motion analysis: a survey. *IEEE Access* 12, 136783–136804. doi:10.1109/access.2024.3462109

Iddrisu, K., Shariff, W., O'Connor, N. E., Lemley, J., and Little, S. (2024b). Evaluating image-based face and eye tracking with event cameras. *ECCV*. doi:10.48550/arXiv.2408.10395

Indiveri, G. (2008). Neuromorphic vlsi models of selective attention: from single chip vision sensors to multi-chip systems. *Sensors* 8, 5352–5375. doi:10.3390/s8095352

Innocenti, S. U., Becattini, F., Pernici, F., and Bimbo, A. D. (2021). "Temporal binary representation for event-based action recognition," in 2020 25th International Conference on Pattern Recognition (ICPR) (IEEE), 10426–10432.

Joubert, D., Marcireau, A., Ralph, N., Jolley, A., Van Schaik, A., and Cohen, G. (2021). Event camera simulator improvements via characterized parameters. *Front. Neurosci.* 15, 702765. doi:10.3389/fnins.2021.702765

Kanamaru, T., Arakane, T., and Saitoh, T. (2023). Isolated single sound lip-reading using a frame-based camera and event-based camera. *Front. Artif. Intell.* 5, 1070964. doi:10.3389/frai.2022.1070964

Kohyama, K., Shiba, S., and Aoki, Y. (2024). "3d human scan with a moving event camera," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5586–5596. doi:10.1109/cvprw63382.2024.00568

Kramer, J., and Indiveri, G. (1998). "Neuromorphic vision sensors and preprocessors in system applications," in Proceedings of the Advanced Focal Plane Arrays and Electronic Cameras II (Zurich, Switzerland: SPIE), 134–146. 3410.

Lagorce, X., Orchard, G., Galluppi, F., Shi, B. E., and Benosman, R. B. (2017). "Hots: a hierarchy of event-based time-surfaces for pattern recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 1346–1359. doi:10.1109/tpami.2016.2574707*IEEE Trans. Pattern Anal. Mach. Intell.* 39.

Lenz, G., Ieng, S.-H., and Benosman, R. (2020). Event-based face detection and tracking using the dynamics of eye blinks. *Front. Neurosci.* 14, 587. doi:10.3389/fnins.2020.00587

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

Lichtsteiner, P., Posch, C., and Delbruck, T. (2008). "A 128 × 128 120 db 15μs latency asynchronous temporal contrast vision sensor," in IEEE Journal of Solid-State Circuits, 566–576. doi:10.1109/jssc.2007.914337*IEEE J. Solid-State Circuits* 43.

Lin, S., Ma, Y., Guo, Z., and Wen, B. (2022). "Dvs-voltmeter: stochastic process-based event simulator for dynamic vision sensors," in European Conference on Computer Vision (Cham: Springer Nature Switzerland), 578–593.

Liu, S., and Delbruck, T. (2010). Neuromorphic sensory systems. *Curr. Opin. Neurobiol.* 20, 288–295. doi:10.1016/j.conb.2010.03.007

Liu, M., and Delbrück, T. (2018). "Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors," in Proceedings of the British Machine Vision Conference (BMVC) (Newcastle upon Tyne, UK: BMVC).

Liu, Q., Xing, D., Tang, H., Ma, D., and Pan, G. (2021). Event-based action recognition using motion information and spiking neural networks. *IJCAI*, 1743–1749. doi:10.24963/ijcai.2021/240

Ma, F., Sun, B., and Li, S. (2023). "Facial expression recognition with visual transformers and attentional selective fusion," in IEEE Transactions on Affective Computing, 1236–1248. doi:10.1109/taffc.2021.3122146*IEEE Trans. Affect. Comput.* 14.

Martini, M., Adhuran, J., and Khan, N. (2022). Lossless compression of neuromorphic vision sensor data based on point cloud representation. *IEEE Access* 10, 121352–121364. doi:10.1109/access.2022.3222330

Miao, S., Chen, G., Ning, X., Zi, Y., Ren, K., Bing, Z., et al. (2019). Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Front. Neurorobotics* 13, 38. doi:10.3389/fnbot.2019.00038

Mitrokhin, A., Fermüller, C., Parameshwara, C., and Aloimonos, Y. (2018). "Event-based moving object detection and tracking," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE), 1–9.

Moreira, G., Graça, A., Silva, B., Martins, P., and Batista, J. (2022). "Neuromorphic event-based face identity recognition," in 2022 26th International Conference on Pattern Recognition (ICPR) (IEEE), 922–929. doi:10.1109/icpr56361.2022.9956236

Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., and Scaramuzza, D. (2017). The event-camera dataset and simulator: event-based data for pose estimation, visual odometry, and slam. *Int. J. Robotics Res.* 36, 142–149. doi:10.1177/0278364917691115

Nguyen, D. T., Quach, M., Valenzise, G., and Duhamel, P. (2021). "Learning-based lossless compression of 3d point cloud geometry," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), 4220–4224.

Plizzari, C., Planamente, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., et al. (2022). "E2 (go) motion: motion augmented event stream for egocentric action recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 19935–19947.

Posch, C., Serrano-Gotarredona, T., Linares-Barranco, B., and Delbruck, T. (2014). "Retinomorphic event-based vision sensors: bioinspired cameras with spiking output," in Proceedings of the IEEE, 1470–1484. doi:10.1109/jproc.2014.2346153*Proc. IEEE* 102.

Pradhan, B. R., Bethi, Y., Narayanan, S., Chakraborty, A., and Thakur, C. S. (2019). "N-har: a neuromorphic event-based human activity recognition system using memory surfaces," in Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), 1–5. doi:10.1109/iscas.2019.8702581

Prophesee, (2023). Video to event simulator [Computer software]. *Metavision SDK*. Available online at: https://www.prophesee.ai/metavision-sdk-pro/.

Rafi, U., Doering, A., Leibe, B., and Gall, J. (2020). "Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16 (Springer), 36–52.

Ramesh, B., Zhang, S., Yang, H., Ussa, A., Ong, M., Orchard, G., et al. (2020). e-tld: event-based framework for dynamic object tracking. *IEEE Trans. Circuits Syst. Video Technol.* 31, 3996–4006. doi:10.1109/tcsvt.2020.3044287

Rebecq, H., Gehrig, D., and Scaramuzza, D. (2018). "Esim: an open event camera simulator," in Proceedings of The 2nd Conference on Robot Learning. Editors A. Billard, A. Dragan, J. Peters, and J. Morimoto, 969–982.*(PMLR), vol. 87 of* proceedings of machine learning research.

Rebecq, H., Ranftl, R., Koltun, V., and Scaramuzza, D. (2019). High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)* 43, 1964–1980. doi:10.1109/tpami.2019.2963386

Ren, H., Zhou, Y., Fu, H., Huang, Y., Xu, R., and Cheng, B. (2023a). "Ttpoint: a tensorized point cloud network for lightweight action recognition with event cameras," in Proceedings of the 31st ACM International Conference on Multimedia, 8026–8034. doi:10.1145/3581783.3612258

Ren, H., Zhou, Y., Huang, Y., Fu, H., Lin, X., Song, J., et al. (2023b). Spikepoint: an efficient point-based spiking neural network for event cameras action recognition. *arXiv Prepr. arXiv:2310.07189*. doi:10.48550/arXiv.2310.07189

Rios-Navarro, A., Piñero-Fuentes, E., Canas-Moreno, S., Javed, A., Harkin, J., and Linares-Barranco, A. (2023). "Lipsfus: a neuromorphic dataset for audio-visual sensory fusion of lip reading," in 2023 IEEE International Symposium on Circuits and Systems (ISCAS) (IEEE), 1–5.

Rudnev, V., Golyanik, V., Wang, J., Seidel, H.-P., Mueller, F., Elgharib, M., et al. (2021). "Eventhands: real-time neural 3d hand pose estimation from an event stream," in Proceedings of the IEEE/CVF international conference on computer vision, 12385–12395.

Ryan, C., O'Sullivan, B., Elrasad, A., Cahill, A., Lemley, J., Kielty, P., et al. (2021). Real-time face & eye tracking and blink detection using event cameras. *Neural Netw.* 141, 87–97. doi:10.1016/j.neunet.2021.03.019

Ryan, C., Elrasad, A., Shariff, W., Lemley, J., Kielty, P., Hurney, P., et al. (2023). Real-time multi-task facial analytics with event cameras. *IEEE Access* 11, 76964–76976. doi:10.1109/ACCESS.2023.3297500

Savran, A., Tavarone, R., Higy, B., Badino, L., and Bartolozzi, C. (2018). "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), 333–340.

Savran, A. (2023). Fully convolutional event-camera voice activity detection based on event intensity,"*2023 Innovations in Intelligent Systems and Applications Conference (ASYU),* Sivas, Turkiye, 1–6. doi:10.1109/ASYU58738.2023.10296754

Schaefer, S., Gehrig, D., and Scaramuzza, D. (2022). "Aegnn: asynchronous event-based graph neural networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12371–12381. doi:10.1109/CVPR52688.2022.01203

Schiopu, I., and Bilcu, R. C. (2022). Lossless compression of event camera frames. *IEEE Signal Process. Lett.* 29, 1779–1783. doi:10.1109/lsp.2022.3196599

Sengupta, N., and Kasabov, N. (2017). Spike-time encoding as a data compression technique for pattern recognition of temporal data. *Inf. Sci.* 406, 133–145. doi:10.1016/j.ins.2017.04.017

Sezavar, A., Brites, C., and Ascenso, J. (2024). Low complexity learning-based lossless event-based compression. *arXiv Prepr. arXiv:2411.07155*, 85–92. doi:10.1109/ism63611.2024.00018

Simulator, C. (2023). Carla simulator dvs camera.

Sokolova, A., and Konushin, A. (2019). "Human identification by gait from event-based camera," in 2019 16th International Conference on Machine Vision Applications (MVA) (IEEE), 1–6.

Steffen, L., Reichard, D., Weinland, J., Kaiser, J., Ronnau, A., and Dillmann, R. (2019). Neuromorphic stereo vision: a survey of bioinspired sensors and algorithms. *Front. Neurorobotics* 13, 28. doi:10.3389/fnbot.2019.00028

Suh, Y., Choi, S., Ito, M., Kim, J., Lee, Y., Seo, J., et al. (2020). "A 1280× 960 dynamic vision sensor with a 4.95-μm pixel pitch and motion artifact minimization," in 2020 IEEE International Symposium on Circuits and Systems (ISCAS) (IEEE), 1–5.

Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). "Deep high-resolution representation learning for human pose estimation," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5686–5696. doi:10.1109/CVPR.2019.00584

Tan, G., Wang, Y., Han, H., Cao, Y., Wu, F., and Zha, Z. J. (2022). "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, 20094–20103.

Tan, G., Wan, Z., Wang, Y., Cao, Y., and Zha, Z. J. (2024). Tackling event-based lip-reading by exploring multigrained spatiotemporal clues. *IEEE Trans. Neural Netw. Learn. Syst.* 36, 8279–8291. doi:10.1109/tnnls.2024.3440495

Tao, Y., Chang, C.-H., Saïghi, S., and Gao, S. (2024). "Gaitspike: event-based gait recognition with spiking neural network," in 2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS) (IEEE), 357–361.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., and Maida, A. S. (2019). Deep learning in spiking neural networks. *Neural Netw.* 111, 47–63. doi:10.1016/j.neunet.2018.12.002

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3d convolutional networks," in 2015 IEEE International Conference on Computer Vision (ICCV), 4489–4497. doi:10.1109/ICCV.2015.510

Verschae, R., and Bugueno-Cordova, I. (2023). Event-based gesture and facial expression recognition: a comparative analysis. *IEEE Access* 11, 121269–121283. doi:10.1109/access.2023.3328220

Vicente-Sola, A., Manna, D. L., Kirkland, P., Di Caterina, G., and Bihl, T. J. (2025). Spiking neural networks for event-based action recognition: a new task to understand their advantage. *Neurocomputing* 611, 128657. doi:10.1016/j.neucom.2024.128657

Viola, P., and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proc. 2001 IEEE Comput. Soc. Conf. Comput. Vis. pattern Recognit. CVPR 2001 *(Ieee)* 1, I. doi:10.1109/CVPR.2001.990517

Wang, Y., Du, B., Shen, Y., Wu, K., Zhao, G., Sun, J., et al. (2019). "Ev-gait: event-Based robust gait recognition using dynamic vision sensors," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 6358–6367.

Wang, Y., Zhang, X., Shen, Y., Du, B., Zhao, G., Cui, L., et al. (2021). Event-stream representation for human gaits identification using deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 3436–3449. doi:10.1109/TPAMI.2021.3054886

Wang, C., Wang, X., Yan, C., and Ma, K. (2023). Feature representation and compression methods for event-based data. *IEEE Sensors J.* 23, 5109–5123. doi:10.1109/jsen.2023.3237754

Wang, Q., Xu, Z., Lin, Y., Ye, J., Li, H., Zhu, G., et al. (2025). "Dailydvs-200: a comprehensive benchmark dataset for event-based action recognition," in European Conference on Computer Vision (Springer), 55–72.

Xiao, P., Zhang, Y., Kai, D., Peng, Y., Zhang, Z., and Sun, X. (2024). "Estme: event-driven spatio-temporal motion enhancement for micro-expression recognition," in 2024 IEEE International Conference on Multimedia and Expo (ICME) (IEEE), 1–6.

Xu, L., Xu, W., Golyanik, V., Habermann, M., Fang, L., and Theobalt, C. (2020). "Eventcap: monocular 3d capture of high-speed human motions using an event camera," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4968–4978.

Yan, W.-J., Wu, Q., Chen, Y.-H., Liang, J., and Fu, X. (2013). How fast are the leaked facial expressions: the duration of micro-expressions. *J. Nonverbal Behav.* 37, 217–230. doi:10.1007/s10919-013-0159-8

Yan, S., Xiong, Y., and Lin, D. (2018). "Silk fabric dyed with extract of sophora flower bud," in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, 308–315. doi:10.1080/14786419.2017.1359170*Nat. Prod. Res.*32

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* 23, 1499–1503. doi:10.1109/LSP.2016.2603342

Zhang, P., Zhu, S., and Lam, E. Y. (2024a). Event encryption: rethinking privacy exposure for neuromorphic imaging. *Neuromorphic Comput. Eng.* 4, 014002. doi:10.1088/2634-4386/ad207b

Zhang, Z., Cui, S., Chai, K., Yu, H., Dasgupta, S., Mahbub, U., et al. (2024b). "V2ce: video to continuous events simulator," in 2024 IEEE International Conference on Robotics and Automation (ICRA) (IEEE), 12455–12461.

Zheng, X., Liu, Y., Lu, Y., Hua, T., Pan, T., Zhang, W., et al. (2024). Deep learning for event-based vision: a comprehensive survey and benchmarks

Zhu, A. Z., Yuan, L., Chaney, K., and Daniilidis, K. (2018a). "Ev-flownet: self-supervised optical flow estimation for event-based cameras," in Proceedings of Robotics: Science and Systems (RSS), 315–323.

Zhu, Z., Yuan, L., Chaney, K., and Daniilidis, K. (2018b). "Unsupervised event-based optical flow using motion compensation," in Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

Zou, S., Mu, Y., Zuo, X., Wang, S., and Cheng, L. (2023). Event-based human pose tracking by spiking spatiotemporal transformer. *arXiv Prepr. arXiv:2303.09681*.