IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 July 2023, accepted 15 October 2023, date of publication 27 October 2023, date of current version 3 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3328220

TOPICAL REVIEW

Event-Based Gesture and Facial Expression Recognition: A Comparative Analysis

RODRIGO VERSCHAE, (Member, IEEE), AND IGNACIO BUGUENO-CORDOVA, (Member, IEEE)
Institute of Engineering Sciences, Universidad de O'Higgins, Rancagua 2841959, Chile

Corresponding author: Rodrigo Verschae (rodrigo@verschae.org)

This work was supported in part by the Programa de Equipamiento Científico y Tecnológico, ANID, Chile (FONDEQUIP) Project under Grant EQM170041.

ABSTRACT Event-based cameras are novel vision sensors that respond to local variations in intensity, generating asynchronous pixels, referred to as events, with low latency, high temporal resolution, and high dynamic range. These events contain information related to the spatio-temporal dynamics of a scene. Given the temporal nature of the asynchronous event stream, several authors have contributed to recognising deformable objects in motion, specifically gestures. However, another category of deformable objects, such as facial expressions, has yet to be adequately addressed. In this paper, we present a comprehensive review of two topics of interest in novel event-based cameras: gesture and facial expression recognition. For both tasks, we evaluate two existing state-of-the-art learning models, and also we use a third model that learns from temporal and spatial correlations of events. To this end, we evaluate a wide range of classification models across multiple scenarios, analysing: the time/event cut-off window of the sample, the number of samples per class for each database, the spatial resolution of the databases, amongst other factors. In the case of gesture recognition, we utilise existing databases, while in the case of facial expression recognition we have synthetically generated two completely new databases (based on two state-of-the-art image databases): e-CK+ and e-MMI, with promising results for the future of this area. Finally, we provide our contributions to the community, specifically the databases developed and used for this study.

INDEX TERMS Event-based cameras, bio-inspired vision, asynchronous sensors, gesture recognition, facial expression recognition.

MULTIMEDIA MATERIAL

The project's code is available on the following page:
https://github.com/uoh-rislab/event-based_gesture_and_facial_expression_recognition

I. INTRODUCTION

Traditionally, computer vision research has focused on developing methods for frame-based cameras, thus working with images or videos. While recent advancements in deep learning, active learning, and reinforcement learning have led to significant progress in computer vision applications, these techniques cannot fully address the limitations of traditional cameras, such as high latency, motion blur, and low dynamic range. To overcome these limitations, researchers have begun to explore the use of event-based neuromorphic vision sensors, also known as event cameras [1]. These devices are characterised by their response to local intensity changes in a scene, generating an asynchronous stream of events that capture the time stamp and polarity of the brightness change of each pixel. A complete survey on event-based cameras can be found in [2]. For event-based vision tasks, deep learning [3], [4] and other classification techniques have been successfully applied to object and gesture recognition problems, achieving 90% or better classification accuracy when using event-based cameras (e.g. [5], [6]). However, event-based facial expression recognition remains a largely unexplored territory, where research has mainly focused on face detection [7], [8], [9] and eye blinking/tracking [8], [10], [11]. This may be due to the scarcity of neuromorphic databases on facial

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea Bottino.

expressions and the complexity of the problem. Nevertheless, this represents a unique and exciting opportunity for research in event-based facial analysis, specifically focusing on event-based facial expression recognition (as in computer vision [12], [13], [14]). Consequently, this study aims to perform comparative analyses of two interesting topics: event-based gesture and facial expression recognition. For this purpose, we use two event-based gesture benchmark datasets and generate two new facial expression datasets (by applying event emulation methods). Then, we evaluate the performance of the existing classification methods on each task (using the respective datasets) and compare the results by varying two key parameters: the time size and the number of events in a sample of events.

The present paper is structured as follows: Section I introduces the problem under study; Section II provides an overview of the state-of-the-art in event-based gestures and facial recognition, including available datasets, tools, and a review of existing classification methods; Section III presents the comparative analysis of event-based gestures and facial expressions recognition; Section IV addresses the use of two event-based gesture benchmark datasets and the development of two new event-based facial expression dataset; Section V presents the achieved results; and Section VI concludes and projects this work.

II. STATE OF THE ART

This section provides a comprehensive overview of the current state-of-the-art event-based gesture and facial expression recognition. Specifically, we begin by reviewing the fundamentals of event representations and prior research in this field, followed by an examination of the current state-of-the-art in event-based object recognition, with the different methods that have been proposed. This includes reviewing deep learning-based methods, particularly effective for specific tasks, and examining relevant datasets. Finally, we examine novel methods for generating event-based datasets, such as simulators and emulators. These tools are essential for creating realistic and diverse datasets that can be used to train and test event-based recognition systems.

A. EVENT REPRESENTATIONS

Event cameras are electronic devices that asynchronously respond to changes in the received logarithmic brightness signal $L(u_k, t_k) \doteq \log(I(u_k, t_k))$, where u_k corresponds to the sensor pixel, t_k to the associated time, L to the luminosity, and I to the brightness signal [2]. Then, an event is triggered at pixel $u_k = (x_k, y_k)^T$ and at time t_k as soon as the brightness change since the last event at the pixel reaches a threshold $\pm C$ (with $C > 0$):

$$L(u_k, t_k) - L(u_k, t_k - \Delta t_k) \geq p_k C, \tag{1}$$

where $p_k \in \{-1, +1\}$ is the sign of the brightness change and Δt_k is the time since the last event at u_k. Then, in a given time interval, an event camera produces a sequence of events, $\mathcal{E}(t_N) = \{e_k\}_{k=1}^{N} = \{(x_k, y_k, t_k, p_k)\}_{k=1}^{N}$ with microsecond resolution, where t_N denotes the last timestamp in the event stream $\mathcal{E}$. An important issue in event-based cameras is extracting and representing meaningful information from the generated event data. As mentioned above, the most basic representation in event-based cameras is an individual event, which contains information about the pixel's position, timing, and polarity that triggered the event. A second level of representation is an event packet, which corresponds to a spatio-temporal neighbourhood of events. Several methods have been proposed for representing event data, including accumulating events in the image plane. This representation can be used for computer vision tasks like object detection, tracking, and recognition. Likewise, many other learning-based approaches work by first preprocessing events, converting them into dense images or dense tensors –which is a convenient representation for image-based models–, and then using those dense images in convolutional neural networks. Existing representations include [2]: a) Event frame (image) or 2D histogram, b) Time surface, c) Voxel Grid, d) 3D point set, e) Point sets on the image plane, f) Motion-compensated image, g) Reconstructed images. Thus, it should be emphasised that event representation performs a key role in event-based vision tasks, as it describes the spatial and temporal distribution of events. Combined with the high temporal resolution of events, the representation captures detailed temporal distributions over scenes -independently of the vision task- which is impossible with traditional frame-based cameras. Therefore, it is essential to carefully consider the event representation used to extract meaningful information from event data effectively.

B. EVENT-BASED OBJECT RECOGNITION

Gestures and facial expressions are often used as similar terms. However, the taxonomy of deformable and non-deformable objects shows that these categories describe different types of spaces. To clarify this, a taxonomy of objects is described below (as presented in Fig. 1) to distinguish the different types of objects clearly.

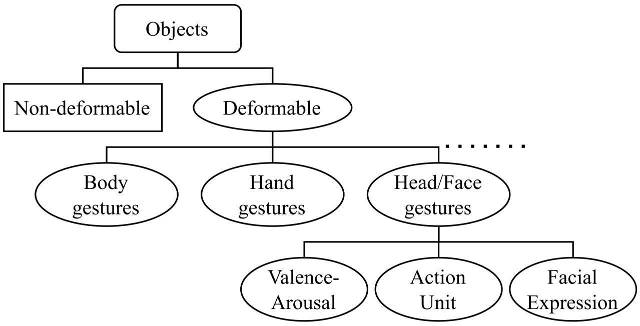


FIGURE 1. Object recognition taxonomy in computer vision: from non-deformable to deformable objects. The left branch represents non-deformable objects characterised by a constant shape and volume. The right branch represents deformable objects, characterised by a variable shape and volume, including body gestures, hand gestures, head and face gestures.

In computer vision, objects can be classified into non-deformable and deformable objects. Non-deformable objects maintain their shape and structure (such as a chair or a table), while deformable objects change their shape and structure, such as a human body or a face. Deformable objects can be further divided into several subcategories. Body gestures, for example, refer to the movement of the entire body, such as a wave or a dance move. Hand gestures, on the other side, refer to the movement of the hands and fingers, such as pointing or grasping. Facial gestures, meanwhile, refer to the movement of the face and head, such as nodding or shaking. Facial gestures can be further classified into several subcategories, such as Valence-Arousal, Action Unit, and Facial Expression. Valence-Arousal refers to the emotional state of the individual, with valence indicating the positivity or negativity of the emotion and arousal indicating the intensity of the emotion. Action Unit, on the other hand, refers to individual muscle movements on the face, such as raising the eyebrows or smiling. Lastly, Facial Expression combines Action Units to convey a specific emotion, such as happiness or sadness. Frame-based deformable object recognition is an active area of research focusing on recognising objects that can change their shape or appearance over time. Event cameras in this task offer several advantages over traditional cameras, such as high dynamic range and lower event latency. However, processing data from event cameras also involves significant challenges, such as handling large amounts of data and developing new algorithms to extract meaningful information from asynchronous event streams. Therefore, the following subsection reviews the state-of-the-art of event-based object recognition, providing the foundations for developing the comparative analysis proposed in this article.

1) EVENT-BASED NON-DEFORMABLE OBJECT RECOGNITION

In event-based pattern recognition, the study of non-deformable object recognition is one of the most extensively researched areas. Early contributions made use of traditional Machine Learning methods for real-time recognition, such as CNN [15] and Random Forest [16], as well as advanced methods that take advantage of the spatial and temporal dispersion of events [17]. More recent research has focused on the use of state-of-the-art methods, such as SNN [18], Graph-based [19], Visual Transformer [20]) and the development of specialised hardware [21].

At the same time, the availability of large and well-known object databases in computer vision –such as Caltech101 [22] and MNIST [23], has led to the development of neuromorphic counterparts –N-Caltech101 [24] and N-MNIST [24], respectively–. These neuromorphic datasets are generated using event cameras and capture the spatiotemporal characteristics of the objects, making them well-suited for event-based recognition tasks. Other examples of neuromorphic datasets include MNIST-DVS [25], Poker-DVS [25], N-CARS [26], N-ImageNet [27], among others. These datasets provide a valuable resource for developing and validating new event-based non-deformable object recognition methods.

2) EVENT-BASED GESTURE RECOGNITION

Along with non-deformable object recognition, gesture recognition is another important area in event-based applications. In this field, the development of recognition methods based on state-of-the-art computer vision has laid the foundations for new and novel contributions that have aimed to exploit the temporal dispersion of events. Examples include: End-to-End Learning of Representations using Event Spike Tensor (EST) [5], Event-based Asynchronous Sparse Convolutional Networks (Asynet) [6], Spike Layer Error Reassignment in time (SLAYER) [28], Deep Continuous Local Learning (DECOLLE) [29], Time-Ordered Recent Event (TORE) [30], RG-CNN [31]. These contributions use Deep Learning-based architectures and emphasise the design of new event representations capable of capturing gestures' spatial and temporal dispersion, enabling these recognition systems to achieve high-performance metrics.

In terms of neuromorphic gesture databases, their availability is more limited. However, there are several datasets available in the literature, such as the DVS128 Gesture Dataset [32], IITM DVS128 Gesture Dataset [33], NavGesture [34], and Celex DVS gesture dataset [35]. We highlight the DVS128 Gesture Dataset, a pioneering database with 11 classes of hand and arm gestures (including the category "other"). This dataset is widely used in the literature as a comparative benchmark to evaluate new event-based gesture recognition methods' performance, as seen in Table 1.

TABLE 1. State-of-the-art summary of classification accuracy on the DVS128 Gesture dataset [32].

Model	Ref	10 cl.	11 cl.
Time-surfaces	[34]	96.59	90.62
SNN eRBP	[36]	N/A	92.70
Slayer	[28]	N/A	93.64
CNN	[32]	96.49	94.59
Space-time clouds	[37]	97.08	95.32
DECOLLE	[29]	N/A	95.54
TORE	[30]	N/A	96.2
EvT	[38]	98.46	96.20
RG-CNN	[31]	N/A	97.2
AlexNet - LSTM	[39]	97.5	97.53
Inception3D + Voting	[39]	99.58	**99.62**

Table 1 summarises the different state-of-the-art event-based gesture recognition methods using two different numbers of classes for the DVS128 Gesture dataset. The best-performing model for this task is Inception3D + Voting, which achieves an accuracy of 99.58% and 99.62% for data sets of 10 and 11 classes, respectively. Advanced models, such as EvT, Space-time clouds and AlexNet - LSTM, also obtain high accuracy scores. The performance in Table 1 illustrates that event-based gesture recognition is promising, with several high-performance models currently available. Thus, it is clear that event-based gesture recognition has experienced significant advances in recent years, and the performance of the designed learning methods is continuously improving.

3) EVENT-BASED FACIAL ANALYSIS

Face analysis is a well-studied area in computer vision, with a significant amount of research devoted to the study of facial features, facial recognition, and facial expression recognition [12], [13], [14]. However, the field of event-based facial analysis is relatively unexplored regarding event-based cameras. To date, research in this area has been primarily limited to face detection [7], [8], [9] and eye blinking/tracking [8], [10], [11]. These studies address important issues in face analysis; however, they are not directly applicable to facial expression recognition. Furthermore, a lack of event-based databases for facial expressions and the complexity of the problem make event-based facial expression recognition a challenging task. Despite these challenges, the potential for event-based cameras to capture high-temporal-resolution facial expressions presents an exciting new opportunity for research. Two additional works that mention emotion recognition using event-based cameras are [40], which explores using event-based cameras to identify individuals based on facial dynamics derived from speech, and [41], which proposes modelling expressions with event-based cameras to understand human reactions.

a: EVENT-BASED FACE DETECTION

In event-based face detection, several approaches have been proposed to take advantage of the unique properties of event cameras. In [7], a patch-based model was proposed for event streams acquired from event cameras, which performed direct face detection and highlighted the potential of these sensors for low-power vision applications. Another approach, presented in [8], exploited the high temporal resolution of events to detect the presence of a face in a scene, using eye blinks (characterised by a unique temporal signature over time) and applying a probabilistic framework for face localisation and tracking, in both indoor and outdoor environments. Additionally, [9] proposed an event-based method for learning face representations using kernelized correlation filters within a boosting framework, useful for surveillance applications. These works demonstrate the potential of event cameras for face detection and open up new lines of research.

b: EVENT-BASED EYE BLINKING/TRACKING

In this topic, research has focused on utilising the high temporal resolution of event cameras to detect and track eye movements. In [8], a correlation of the acquired local face activity with a generic temporal model of eye blinks is made, using the blinks to correct for drift and tracking errors in faces. Another work, [10], presents a hybrid event-based near-eye gaze tracking system that offers update rates beyond 10,000 Hz. This system integrates an online 2D pupil fitting method that updates a parametric model for a few or every event, using a polynomial regressor to estimate the gaze point. Additionally, [11] proposes a method for simultaneously detecting and tracking faces and eyes in driver monitoring systems, using a fully convolutional recurrent neural network architecture to determine the driver's level of fatigue.

C. DATA GENERATION USING SIMULATORS AND EMULATORS OF EVENT-BASED CAMERAS

The problem of lack of data is a common challenge in new research areas. This is mostly the case with novel sensors like event-based cameras, where the availability of databases to train machine learning-based algorithms is limited. For example, only the Face Detection [8], DAVIS Face [9], and Neuromorphic Helen [11] datasets are available for face detection, eye tracking, and blink recognition, respectively. To address this problem, several researchers have developed simulators and emulators to generate realistic event-based data from traditional frame-based data. One example is ESIM [42], which simulates an event camera from an adaptive rendering scheme that selectively samples frames. Another example is v2e [43], which aims to replicate the working principle of event cameras to generate high-fidelity DVS event streams with a realistic temporal resolution, high dynamic range, pixel-level Gaussian event, finite intensity-dependent bandwidth, and intensity-dependent noise [43]. More details about the v2e emulator fundamentals will be provided later. This tool allows the generation of two new large and diverse datasets for event-based facial expression recognition.

III. GESTURE AND FACIAL EXPRESSION RECOGNITION: A COMPARATIVE ANALYSIS

This section presents a comprehensive methodology for evaluating event-based gesture and facial expression recognition. We introduce the pipeline of Fig. 2, commonly used in most existing works that classify event data, particularly those based on deep learning techniques. This pipeline includes an event representation module that captures the event's spatial and temporal distribution, an event feature extractor, and a classifier layer. Sometimes a sequence learning module is also added (a network that learns about the temporal evolution of events).

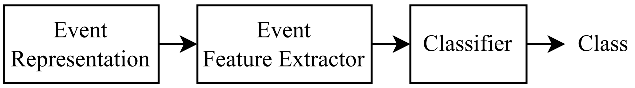


FIGURE 2. Pipeline for event-based recognition methods. This pipeline addresses the need to: (i) a representation that captures the event's spatial distribution and asynchronous temporal evolution, (ii) build a feature extractor that properly encodes the input information, (iii) accurately classify the respective event streams.

In the following, we study three learning models, each of which will be evaluated for the respective recognition task.

A. EVENT-BASED RECOGNITION METHODS EVALUATED

Event-based recognition methods are a benchmark for research, as they have diverse applications in fields such as human-computer interaction and robotics. In the following subsections, we describe the methods used in the comparative analysis later presented in this work.

1) END-TO-END LEARNING OF REPRESENTATIONS FOR ASYNCHRONOUS EVENT-BASED DATA

A *End-to-End Learning of Representations* method used for event-based classification is proposed in [5]. This method utilises a sequence of differentiable operations to convert events into grid-based representations, which are then integrated into a learning model for end-to-end event representations. The above-mentioned representation, known as the Event Spike Tensor (EST), effectively preserves spatial, temporal, and polarity information by converting the event set $\mathcal{E}(t_N)$ into a grid-based representation. The EST representation provides three sub-representations: Two-Channel Image, Event Frame and Voxel Grid. Subsequently, the novel event representation is fed into a CNN, with a ResNet-34 backbone, to learn from events. Furthermore, the architecture employs a fully connected layer to detect the processed features' global configurations and a softmax activation function to assign a probabilistic distribution to each class under study. The described pipeline is illustrated in Fig. 3, based on [5].

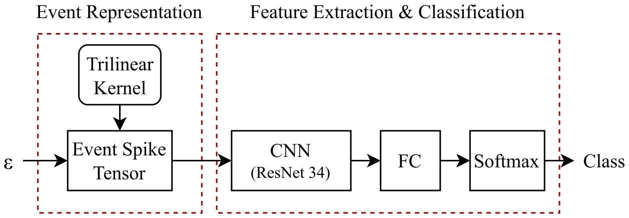


FIGURE 3. End-to-End Learning of Representations for Asynchronous Event-Based Data pipeline, inspired from [5]. The event stream ε is processed by applying a trilinear kernel to generate the EST representation, which preserves spatial, temporal and polarity information. The representation is then fed into an adapted CNN (based on a Resnet-34 as a backbone), with a fully connected layer (FC) to detect global configurations of the processed features and a softmax activation function.

This architecture has been applied to event-based object recognition (N-Caltech101 dataset [24]) and event-based gesture recognition (DVS 128 Gesture Dataset [32]). In the case of gesture recognition, the method achieved an accuracy of 93.82%, representing the best performance until 2019.

2) ASYNCHRONOUS SPARSE CONVOLUTIONAL NETWORKS

A second method for the event-based gesture and object recognition is *Asynchronous Sparse Convolutional Networks* [6], usually known as Asynet, which aims to exploit the spatiotemporal sparsity of event data in convolutional architectures through a novel approach. This method focuses on creating a sparse representation of events and integrating it with a Submanifold Sparse Convolutional network (SSC).

The sparse representation proposed by the authors is an image-like representation that can be processed by standard CNNs, preserving the spatial distribution of the events but discarding their temporal dispersion. To recover the temporal sparsity of the event stream ε, the authors focus on changing the image-like representation when a new event arrives, creating a novel Sparse Recursive Representation (SSR).

As for the SSC network, it is a novel type of CNN designed to exploit the spatiotemporal sparsity of event data. The SSC uses a sparse representation of events as input and applies convolutional operations to extract features and make predictions. Unlike traditional CNNs, which generate fuzzy activation maps and thus reduce sparsity, the SSC network only computes the convolution operation on active sites, leading to sparse activation maps in subsequent layers. This approach preserves the event data's sparse nature and reduces the computational consumption required.

Based on the type of convolutional network used, the authors [6] propose two variants: *Dense* and *Sparse* (referred to as Asynet-I and Asynet-II, respectively). Both variants perform asynchronous processing of the sparse representation, however, they differ in that Asynet-I uses a traditional CNN while Asynet-II uses the novel SSC network. The Asynchronous Sparse Convolutional Networks pipeline is illustrated in Fig. 4.

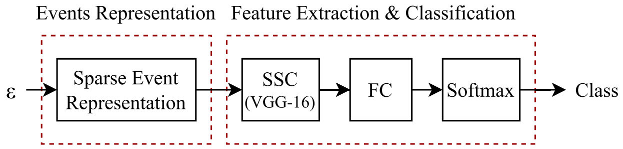

Events Representation
Feature Extraction & Classification
ε
Sparse Event Representation
SSC (VGG-16)
FC
Softmax
Class

FIGURE 4. Asynchronous Sparse Convolutional Networks pipeline, inspired from [6]. The event stream ε is processed to generate the Sparse Recursive Representation. It is then fed into a Submanifold Sparse Convolutional network (SSC), which has a VGG-16 backbone, fully connected layers, and a softmax activation function for the final prediction. The SSC network is specifically designed to exploit the spatiotemporal sparsity of the event data.

Experimental results on the DVS128 Gesture dataset show that Asynet has better performance on gesture recognition tasks (94.66% accuracy [6]) compared to End-to-End Learning of Representations (93.82% accuracy [5]). The authors attribute this improvement mainly to the novel Sparse Recursive Representations and the SSC Network for asynchronous events. However, it is important to note that this method has high inference times.

3) LEARNING FROM THE TEMPORAL SEQUENCE OF ASYNCHRONOUS EVENTS: ESTM

The previous subsections described state-of-the-art methods, providing novel event representations and efficient asynchronous processing. However, both methods have limitations in capturing the temporal dispersion and

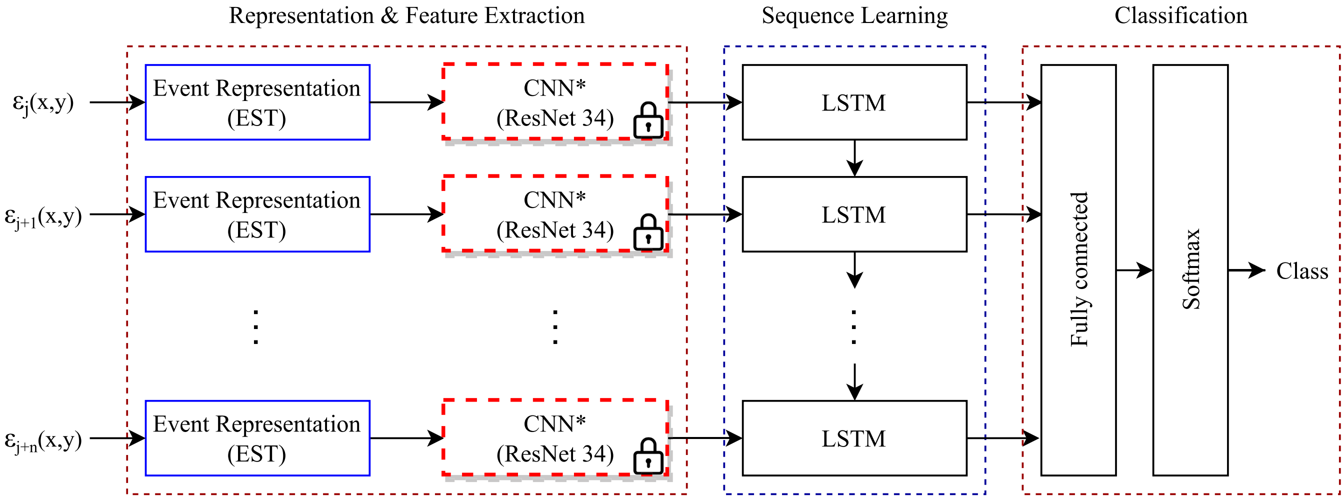


FIGURE 5. ESTM general framework architecture for the event-based gesture and facial expression recognition. This architecture consists of (i) a sequence of event streams, which can be generated with an ESBT/ESBN approach ($\mathcal{E}_j^t(x, y)$ or $\mathcal{E}_j^\epsilon(x, y)$, respectively), (ii) the use of the Event Spike Tensor as an event representation, (iii) a feature extractor module responsible for identifying and encoding the events' key patterns, (iv) the addition of temporal memory networks that allow temporal learning of the sequence of events, (v) a classification module that categorises the events into the respective class.

sequential correlation of event streams. As a result, the event-based object recognition models cannot fully learn from the temporal information in asynchronous event sequences. To address this issue, we present an ESTM approach combining the Event Spike Tensor representation [5] with a Recurrent Neural Network, specifically a Long Short Term Memory (LSTM) network. This approach aims to learn from event sequences' spatial and temporal information to improve object recognition.

In this enhanced method, the EST event representation captures events' spatial, temporal, and polarity information. Then, a CNN with fixed weights extracts spatial and temporal dispersion features. Finally, we integrate an LSTM network between the previously trained static backbone and the fully connected classifier layer, as shown in Fig. 5. Thus, the LSTM network can model the temporal dependencies in asynchronous event sequences, enhancing the described method's performance.

In this architecture, the system receives an event sub-stream $\mathcal{E}_j$ as input. Based on the considerations made by [37], two possible approaches can be adopted for the elaboration of event substreams:

- Event Stacking Based on Time (ESBT): a substream of events is sequenced in $\mathcal{E}_j^t(x, y)$ from a fixed duration of time of the temporal size of the sample (denoted as Δt). Here, the time duration of the event stream is divided into n equal-scale portions, and then $\mathcal{E}_j^t(x, y)$ are built by sequencing the events in each time interval $\left[\frac{(j-1)\Delta t}{n}, \frac{j\Delta t}{n}\right]$. Thus, each $\mathcal{E}_j^t(x, y)$ sub-stream is generated from time windows, i.e. $\mathcal{E}_j^t(x, y) = f(j, \Delta t)$.
- Event Stacking Based on Number of events (ESBN): a substream of events is sequenced in $\mathcal{E}_j^\epsilon(x, y)$ using a constant number of events for each sample (denoted as $\Delta\epsilon$). Here, the event stream is divided into m equal-scale portions, and then $\mathcal{E}_j^\epsilon(x, y)$ are built by sequencing the events in each time interval $\left[\frac{(j-1)\Delta\epsilon}{m}, \frac{j\Delta\epsilon}{m}\right]$. Thus, each $\mathcal{E}_j^\epsilon(x, y)$ sub-stream is generated from time windows, i.e. $\mathcal{E}_j^\epsilon(x, y) = f(j, \Delta\epsilon)$.

The two approaches described above – ESBT and ESBN – have a key role in the comparative analysis proposed in this paper. These two parameters will allow us to determine the relevance and priority of each variable in the performance of event-based learning methods. By comparing the performance of event-based object recognition methods using both ESBT and ESBN, we can better understand how the temporal and spatial sparsity of events influences the learning of the models, which is essential for the practical application of these systems. This, for this work, we will analyse its performance for both event-based gesture and event-based facial expression recognition.

B. SIMULATION/EMULATION METHODS

An intrinsic challenge in event-based gesture and facial expression recognition is the scarcity of databases. The availability of event-based cameras that capture only the event stream (e.g., DVS event-based vision sensor) or those that also capture the frames of traditional cameras (e.g., DAVIS event-based vision sensor) is limited, which makes it difficult to employ a rigorous methodology to develop new databases. To address this issue, [43] proposes the v2e algorithm, which simulates the working principle of event cameras to generate realistic DVS event streams.

This algorithm, illustrated in Fig. 6, uses the SuperSloMo interpolation model to achieve high temporal resolution and calculates the logarithm of the luminance intensity to achieve a high dynamic range. By comparing the logarithmic curve with the activation thresholds, the algorithm generates emulated events and associated frames.

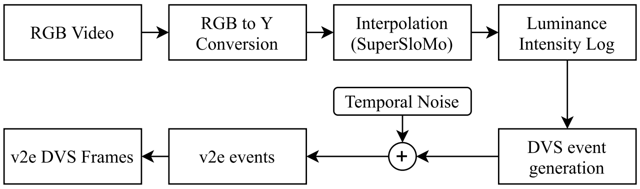


FIGURE 6. ***V2E From video frames to DVS events*** **pipeline, based on [43]. This method generates events from conventional frame-based video and simulates the behaviour of a neuromorphic sensor, providing a realistic emulation.**

This approach can be useful in situations where: (i) it is not possible to acquire real event data, (ii) it is necessary to control the variability of the data to perform a fair comparison between different methods, or/and (iii) the scarcity of event-based databases is a major challenge. Then, this method enables the evaluation of event-based algorithms on a large collection of conventional video datasets and helps bridge the gap between event-based and frame-based recognition research.

IV. DATASETS

This study performs a comparative analysis of event-based gesture and facial expression recognition methods using several datasets. Specifically, we use two event-based gesture benchmark databases and generate two completely new databases for event-based facial expression recognition using a state-of-the-art emulator. The following section provides an overview of the databases employed for this analysis. The databases can be accessed from the website https://sites.google.com/uoh.cl/uoh-ris-lab/datasets.

A. EVENT-BASED GESTURE DATA

1) IBM DVS128 GESTURE DATASET

A pioneering database for event-based gesture recognition research is DVS128 Gesture [32] –also called IBM DVS128 Gesture Dataset– developed by IBM Research, in collaboration with UC San Diego and UZH-ETH Zurich. This dataset was recorded using an Event-Based Vision Sensor (DVS128) with a 128 × 128 spatial resolution in a controlled environment with varying lighting conditions. This combination of challenging yet realistic recording conditions makes the IBM DVS128 Gesture dataset a valuable benchmark for evaluating the performance of gesture recognition algorithms. The gestures in the IBM DVS128 Gesture dataset include hand and finger movements, which are crucial for a wide range of human-computer interaction applications. It is structured into 11 hand and arm gestures classes, which are grouped into multiple trials collected from 29 subjects under three different lighting conditions, as partially shown in Fig. 7.

These features make the dataset representative of the gesture recognition problem, and its relevant contribution is evident in the 250+ literature citations and the seven state-of-the-art methods that have achieved an accuracy of over 95% for all 11 classes, as reported in Table 1. That is why the IBM DVS128 Gesture dataset is widely used as a benchmark for evaluating new gesture recognition and classification algorithms.

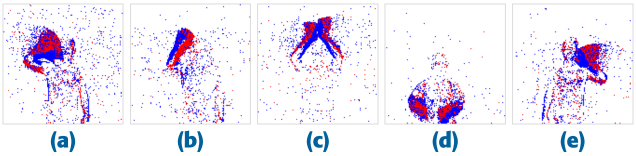


FIGURE 7. IBM DVS128 Gesture Dataset - Events representation with 128 × 128 spatial resolution and 33*ms* time window. Gestures: (a) Right-hand wave; (b) Right hand clockwise; (c) Hand clapping; (d) Drums; (e) Left-hand wave.

2) NavGesture DATASET

A second event-based gesture recognition dataset is NavGesture [34]. It consists of recording gestures made by subjects using a wearable device equipped with an event-based vision sensor and an accelerometer. The gestures in this dataset include hand and finger movements that are typically used for navigation tasks of smart mobile devices, i.e., down swipe, up swipe, left swipe, right swipe, select, and home (as partially shown in Fig. 8).

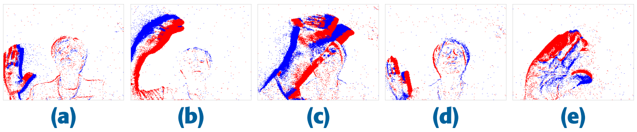


FIGURE 8. NavGesture Sit Dataset - Events representation with 304 × 240 spatial resolution and 33*ms* time window. Gestures: (a) Swip Left; (b) Swip Up; (c) Home; (d) Swip Right; (e) Select.

The dataset was recorded in a realistic environment using the ATIS sensor (304 × 240 spatial resolution) with varied background scenes (sitting for static conditions and walking for dynamic indoor and outdoor conditions), illumination conditions, and subject populations, making it a challenging and representative test bed for evaluating gesture recognition algorithms. Additionally, the NavGesture dataset provides an opportunity to study the performance of algorithms under different motion and gesture types.

As this is a recent contribution, the state of the art shows that it has not been compared with other methods: only an accuracy of 95.9% for the sit scenario and 92.6% for the walk scenario was reported in [34]. To achieve these metrics, the authors used a time-surface representation as a descriptor of the spatiotemporal neighbourhood of an event and built their own learning network.

B. EVENT-BASED FACIAL EXPRESSION DATA

In many problems, when a new sensor is made available, scarcity of databases has always been a major issue [44]. Traditional data collection methods can be difficult, time-consuming, and expensive. Consequently, v2e is a valuable resource for research development in this area.

Using these emulated event streams and literature-inspired facial expression recognition methods makes it feasible to address the aforementioned limitations and offer new contributions. This will be described in the next subsection to perform a comparative analysis of event-based facial expressions.

1) USING v2e FOR EVENT EMULATION

In the present work, we aim to evaluate the proposed facial expression recognition methods using two event-based datasets. Due to the data scarcity for event-based facial expression recognition, we synthetically generate event-based equivalents of two frame-based facial expression benchmarks: the CK+ and the MMI datasets. The CK+ dataset is a widely used benchmark for facial expression recognition and contains image sequences of individuals exhibiting various facial expressions [45]. The MMI Facial Expression Database is another well-known benchmark, recorded in static and dynamic conditions [46]. To convert these frame-based image sequences into event-based data, we employ the v2e emulation method [43], generating event streams with high temporal resolution and dynamic range. The parameters used in the v2e method are detailed in Table 2.

TABLE 2. Configuration parameters adopted for the generation/emulation of new event-based databases, using v2e for event-based comparative analysis of gesture and facial expression recognition.

DVS timestamp resolution	1 ms
DVS exposure duration	5 ms
Positive event threshold	0.15
Negative event threshold	0.15
Sigma threshold variation	0.03
Event camera device emulated	DAVIS 346
Dimensions	346x260
Cut-off frequency	30 Hz

These parameters are used to create the Event CK+ dataset (e-CK+) and the Event MMI Facial Expression Database (e-MMI). The event-based equivalent datasets provide a more representative evaluation scenario for event-based facial expression recognition algorithms than the frame-based image sequences from CK+ and MMI. Therefore, in the following subsections, we will describe in more detail the databases used and their properties, as well as report visually on the newly generated databases.

2) EVENT CK+ DATASET (e-CK+)

Facial expression recognition is an important area in computer vision and has numerous applications in fields such as psychology, sociology, and human-computer interaction. Therefore, the taxonomic definition of human emotions in a broad valence and arousal two-dimensional space [47] inspired the authors of [45] for the development and release of the Cohn-Kanade Extended image database (CK+).

The CK+ database is widely used in facial expression recognition and was created to address the limitations of the original Cohn-Kanade database (CK) [48], which only contained posed expressions. The CK database includes a diverse range of spontaneous facial expressions, with 123 subjects recorded under several scenarios. The CK+ database, on the other hand, is an extension of the CK database and includes 93 subjects, with a greater emphasis on capturing the transition of expressions, simplifying the two-dimensional space of valence and arousal into seven fundamental expressions: Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral (as partially shown in Fig. 9). Also, this dataset contains 327 video sequences labelled, recording at 30 frames per second with a spatial resolution of 640 × 490 or 640 × 480 pixels [45].

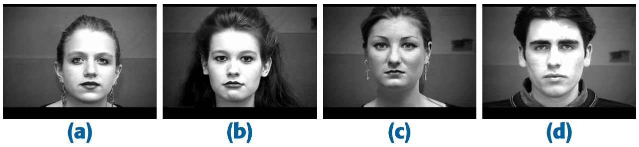


FIGURE 9. The Extended Cohn-Kanade Dataset (CK+) [45], [48]: A complete dataset for action unit and emotion-specified expression, released to promote research on the automatic detection of individual facial expressions. A few example expressions from the CK+ dataset: (a) Anger; (b) Fear; (c) Happiness; (d) Sadness.

Both databases contain images and videos annotated with facial landmarks and expression labels, making them valuable resources. The CK and CK+ databases have been widely used in benchmarking and evaluating the performance of several facial expression recognition algorithms.

In this work, the v2e realistic event generation algorithm is applied to the CK+ dataset to generate the Event CK+ dataset (e-CK+), which contains both the original image sequence and the corresponding asynchronous stream of events, for the seven expressions classes (as partially shown in Fig. 10).

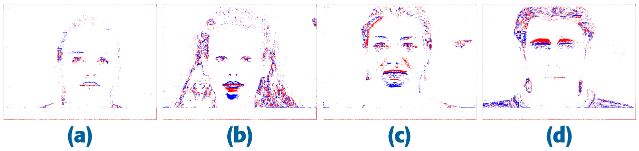


FIGURE 10. The Event-based Extended Cohn-Kanade Dataset (e-CK+): A novel event-based dataset for emotion-specified expression. A few example expressions from the e-CK+ dataset: (a) Anger; (b) Fear; (c) Happiness; (d) Sadness. The images show the events projected into the image plane.

This new dataset is a pioneering state-of-the-art database for event-based facial expression recognition –based on a frame-based vision benchmark– with events characterised by their asynchrony, polarity and spatiotemporal dispersion.

3) EVENT MMI FACIAL EXPRESSION DATABASE (e-MMI)

Another relevant frame-based facial dataset is the Multi-Modal Affect Facial Expression database (MMI) [46], [49], an ongoing project which aims to provide large volumes of visual facial expression data to the facial expression analysis community.

To address the problem of affect recognition, the MMI dataset was conceived as a resource for building and evaluating facial expression recognition algorithms. The database contains recordings of the full temporal pattern of a facial expression (in controlled laboratory conditions), and consists of over 2900 high-resolution videos and still images from 75 subjects, for the nine expression classes: Anger, Disgust, Fear, Happiness, Sadness, Surprise, Scream, Bored, Sleepy (as partially shown in Fig. 11). It is fully annotated for the presence of facial expressions, making it a rich and diverse resource for academic research.

FIGURE 11. The MMI Facial Expression Dataset [46], [49]. A few example expressions from the MMI dataset: (a) Surprise; (b) Anger; (c) Happiness; (d) Disgust.

The MMI database is particularly valuable as it contains annotations of both static and dynamic facial expressions, allowing for the study of subtle and nuanced expressions. Furthermore, the MMI database provides multiple annotations of each expression, including facial landmarks, action units, and expression labels, making it a comprehensive resource for researchers. The MMI database has been widely used as a benchmark for evaluating the performance of facial expression recognition algorithms and has been instrumental in advancing the field.

Analogous to CK+, the v2e realistic event generation algorithm is applied to build the event MMI Extended Dataset (e-MMI), which contains both the original image sequence and the corresponding asynchronous event stream, for the nine expression classes (as partially shown in Fig. 12).

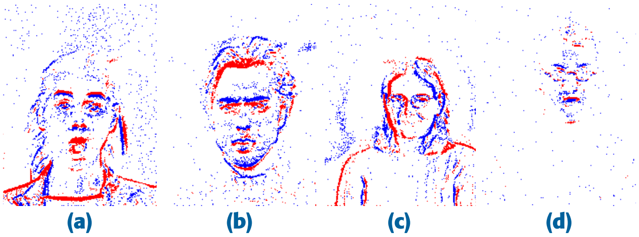


FIGURE 12. The Event-based MMI Facial Expression Dataset (e-MMI): A novel event-based dataset for emotion-specified expression. A few example expressions from the e-MMI dataset: (a) Surprise; (b) Anger; (c) Happiness; (d) Disgust. The images show the events projected into the image plane.

This second new benchmark database for event-based facial expression recognition will be fundamental for the comparative analysis of this study, as it will give objectivity and perspective to the analysis of the results to be reported in the next section.

V. RESULTS AND DISCUSSION

This section presents a comparative analysis of event-based gesture and facial expression recognition using the three state-of-the-art classification models mentioned above: End-to-End Learning of Representations (EST), Asynet CNN Dense (AsyI), Asynet SSC Sparse (AsyII), and ESTM. The performance of these models was evaluated using standard metrics such as accuracy, and by varying two relevant parameters: the size of the time window and the size of the number of events in a sample.

For this purpose, we used two databases for each task: DVS128 Gesture and NavGesture for gesture recognition, the novel Event-based Extended Cohn Kanade (e-CK+) and Event-based MMI (e-MMI) for facial expression recognition. Then, we evaluate each model in different configurations (with corresponding databases), and later, a comparison is reported and discussed.

A. EVENT-BASED GESTURE RECOGNITION

In the following subsections, we will analyse event-based gesture recognition methods using two datasets: IBM's DVS128 and NavGesture. In these data sets, we carefully portioned the data into time windows of different sizes (10 ms, 33 ms and 100 ms) and event windows of different sizes (500 events, 1500 events and 4500 events) to provide a comprehensive examination.

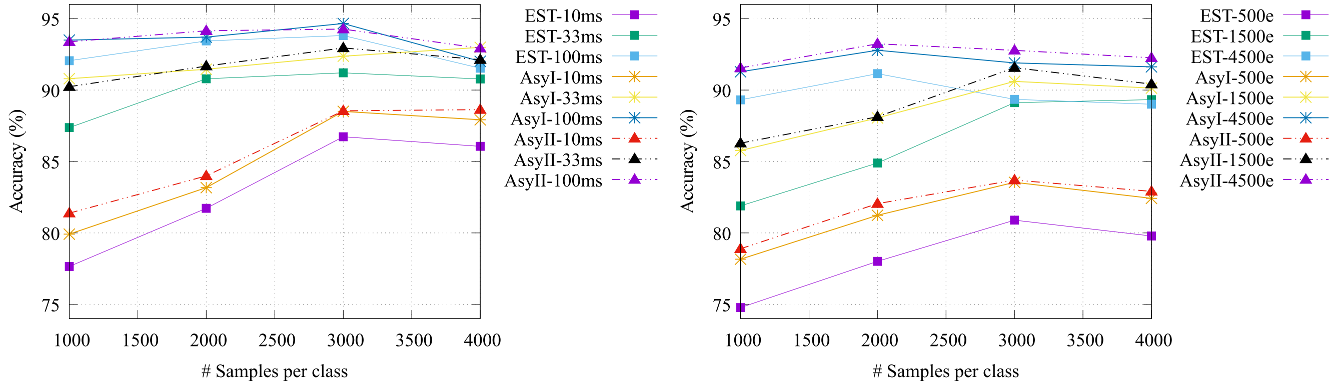
1) IBM DVS128 GESTURE DATASET

This subsection provides the results of the comparative analysis of three event-based recognition methods –EST, AsyI, and AsyII– on the IBM DVS128 Gesture dataset. The results are displayed in Fig. 13. To further evaluate the performance of these methods, two key parameters have been varied in the comparative analysis: the temporal sample size (Fig. 13a), and the number of events per sample (Fig. 13b). In addition, the number of samples per class was also varied, with values ranging from 1000 to 4000.

a: REGARDING THE TEMPORAL LENGTH OF THE SAMPLE

The results of the comparative temporal analysis of the event-based gesture recognition methods on the IBM DVS Gesture 128 database are reported in Fig. 13a.

These findings indicate that, regardless of the number of samples per class, increasing the temporal length of the event samples allows for better learning of the gestures' dynamics, as evidenced by the significant improvement in accuracy observed in the EST method (9.1% increase from 10ms to 33ms and 2.6% increase from 33ms to 100ms). However, it is also noted that as the length of the time window increases, the rate of accuracy improvement decreases. In some cases, an excessive increase in the time window may result in



(a) (b)

FIGURE 13. Performance evaluation of three state-of-the-art event-based recognition methods on the IBM DVS128 Gesture dataset, varying: (a) the temporal length of the sample (10ms, 33ms, 100ms); (b) the number of events in the sample (500 events, 1500 events, 4500 events). In addition, the number of samples per class is varied (from 1000 to 4000) to analyse its impact on the performance of the methods and to measure the trade-off with the other parameters studied. EST refers to [5], AsyI to Asynet Dense [6], and AsyII to Asynet Sparse [6].

overfitting, as demonstrated by the decrease in accuracy (−0.9%) reported in the AsyI method, from 33ms to 100ms.

As for the number of samples per class, Fig. 13a reveals a cutoff point, where event-based classification methods reach an optimal performance level, at 3000 samples per class. Beyond this point, performance decreases due to overfitting, where the classifier learns to fit known patterns and fails to generalise to previously unseen data.

As for the evaluated method, the AsyII method –which employs SSC convolution designed to leverage the sparse and asynchronous distribution of events– demonstrates the best overall performance, as shown in every configuration of the comparative analysis (Fig. 13a). In particular, the AsyII method achieves the highest accuracy (94.27%) in the 100ms time window and 3000 samples per class configuration.

These results suggest that proper time window selection is essential in capturing the events that effectively describe the dynamics of the analysed gestures. The AsyII method yields promising results, demonstrating its effectiveness in event-based gesture recognition.

b: REGARDING THE NUMBER OF EVENTS PER SAMPLE

The results of the comparative analysis of the event-based gesture recognition methods on the IBM DVS Gesture 128 database, as the number of events in each sample varies, are reported in Fig. 13b.

From the above results, and analogous to the previous comparative analysis, it is evident that increasing the number of events results in improved learning of gestural dynamics, regardless of the number of samples per class (e.g., in the case of EST with 2000 samples per class, accuracy increased by 6.9% from 500 events to 1500 events, and by 6.3% from 1500 events to 4500 events). However, it is also observed that the percentage of new learning decreases as the number of events increases, and an excessive increase can lead to over-fitting, as indicated by the decrease in accuracy of −0.3 percent for EST with 4000 samples per class from 1500 events to 4500 events.

As for the number of samples per class, Fig. 13b reveals two cutoff points: 2000 samples per class for the event-based recognition methods with 4500 events per sample and 3000 samples per class for the other methods. This suggests the existence of a trade-off between the number of events in a sample and the number of samples per class, as a sample with a significant number of events may already hold enough information for gesture recognition.

Once again, AsyII performed the best among the evaluated methods but with a different configuration. The highest accuracy of 93.24% was recorded with 2000 samples per class and 4500 events per sample.

The above considerations indicate that an adequate event window is key to learning the dynamics of gestures in event-based recognition tasks.

2) NavGesture DATASET

This subsection provides the results of the comparative analysis of three event-based recognition methods –EST, AsyI, and AsyII– on the NavGesture dataset. The results are presented in Table 3a and 3b for NavGesture Sit and NavGesture Walk, respectively. To further evaluate the performance of these methods, two key parameters have been varied in the comparative analysis: the temporal sample size and the number of events per sample. The number of samples per class has been set at 3000, as it was the best value for most methods in the IBM DVS128 Gesture dataset.

Based on the previous results, the comparative analysis performed on the NavGesture Dataset indicates that the three event-based recognition methods under evaluation demonstrate better accuracy in a static scenario (as shown in Table 3a), as opposed to a dynamic scenario (as shown in Table 3b). This behaviour can be attributed to the reduced complexity in identifying gesture structure events

TABLE 3. Performance evaluation of three state-of-the-art event-based recognition methods on the: (a) NavGesture Sit dataset; (b) NavGesture Walk dataset, varying the temporal length of the sample (10ms, 33ms, 100ms) and the number of events in the sample (500 events, 1500 events, 4500 events). The number of samples per class has been set at 3000. The EST refers to [5], AsyI to Asynet Dense [6], AsyII to Asynet Sparse [6].

	NavGesture - Sit					
	Time window			Event window		
	10ms	33ms	100ms	500e	1500e	4500e
EST	87.6	**89.1**	87.3	79.9	87.6	**89.5**
AsyI	92.1	**93.5**	92.0	80.6	88.9	**90.1**
AsyII	94.4	**94.6**	92.4	83.1	90.5	**92.3**

(a)

	NavGesture - Walk					
	Time window			Event window		
	10ms	33ms	100ms	500e	1500e	4500e
EST	88.3	**88.5**	87.2	75.2	87.3	**88.7**
AsyI	89.5	**90.6**	88.1	78.6	88.4	**90.9**
AsyII	89.7	**91.3**	90.8	84.5	88.9	**91.2**

(b)

in a scenario free of external perturbations caused by the environmental movement. The results suggest that a stable, controlled environment improves gesture recognition by clearly representing the underlying event structure.

For the temporal analysis of events, the best performance for the three methods employed is recorded with an event time window of 33ms. In NavGesture Sit Dataset, Table 3a reports an accuracy of: 89.1% for EST, 93.5% for Asy I and 94.6% for Asy II. In NavGesture Walk Dataset, Table 3b shows an accuracy of: 88.5% for EST, 90.6% for AsyI, and 91.3% for AsyII. However, it should be noted that as the window time becomes shorter (10 ms) or longer (100 ms), the performance of the methods decreases, with the presence of overfitting in the learning phase of the algorithms becoming more apparent. This highlights the importance of selecting an appropriate event sample size in event-based gesture recognition.

For the number of events in each sample, the best performance for the three methods employed is recorded at 4500 events. In NavGesture Sit Dataset, Table 3a reports an accuracy of: 89.5% for EST, 90.1% for Asy I and 92.3% for Asy II. In NavGesture Walk Dataset, Table 3b shows an accuracy of: 88.7% for EST, 90.9% for AsyI, and 91.2% for AsyII. From these results, three preliminary analyses can be inferred.

The first analysis shows that neither time nor event windows consistently perform better. In the NavGesture Sit scenario (Table 3a), using time windows with a size of 33ms resulted in higher accuracy than using event windows with 4500 events. Conversely, in the NavGesture Walk scenario (Table 3b), event windows with 4500 events showed slightly higher accuracy than time windows with 33ms. This disparity suggests that the dynamic conditions of the environment are better captured with a fixed event window than a time window. The second analysis focuses on the number of events in the sample. The analysis shows that a larger number of events –specifically 4500 events– results in improved learning compared to smaller event sample sizes of 500 or 1500 events. The nature of the NavGesture database, consisting of gestures related to navigation on a smartphone device, highlights the importance of capturing enough events to identify the gesture in action accurately. A final analysis shows that AsyII performs best in all scenarios. In the case of the event window for NavGesture Walk (Table 3b), this method excels as it employs an SCC that exploits the sparse distribution of events (91.2% accuracy).

These previous analyses illustrate that, depending on the nature of the action and the environment under study, the choice of the sample (either time window or event window) will impact the performance of the recognition task.

3) LEARNING GESTURES FROM THE TEMPORAL SEQUENCE

The achieved performance with the three state-of-the-art event-based object recognition methods indicates a high learning rate for the IBM DVS128 Gesture and NavGesture (Sit and Walk) database. But can this learning rate be improved? An interesting approach is to train from the temporal sequence of the event samples, given the correlation between them. We employ the ESTM method described earlier: the novel EST representation, an adapted CNN (with a Resnet-34 as a backbone), and an LSTM. Since the LSTM requires sequential input, three 10 ms time windows and three 33 ms time windows are provided to the ESTM, as shown in Table 4.

TABLE 4. Performance evaluation of the ESTM event-based gesture recognition method on the: IBM DVS128 Gesture Dataset (with 11 classes), NavGesture Sit Dataset, NavGesture Walk Dataset. The number of samples per class has been set at 3000.

Method	30ms (3 x 10ms)	100ms (3 x 33ms)
DVS128 Gesture (11 cl.)	96.9	**99.5**
NavGesture Sit	91.4	**93.2**
NavGesture Walk	89.7	**90.9**

Compared to the previous results (Fig. 13 and Table 3), Table 4 shows significantly better performance in accuracy rates for event-based gesture recognition tasks, for both the IBM DVS128 Gesture –99.5%– and NavGesture –93.2% and 90.9% for sit and walk, respectively– datasets. This is due to the temporal encoding of events, which allows for a better understanding of gesture dynamics. The EST representation can adequately characterise the studied events, the modified CNN can extract the spatial characteristics of the events, and the LSTM models the temporal sequence of the events.

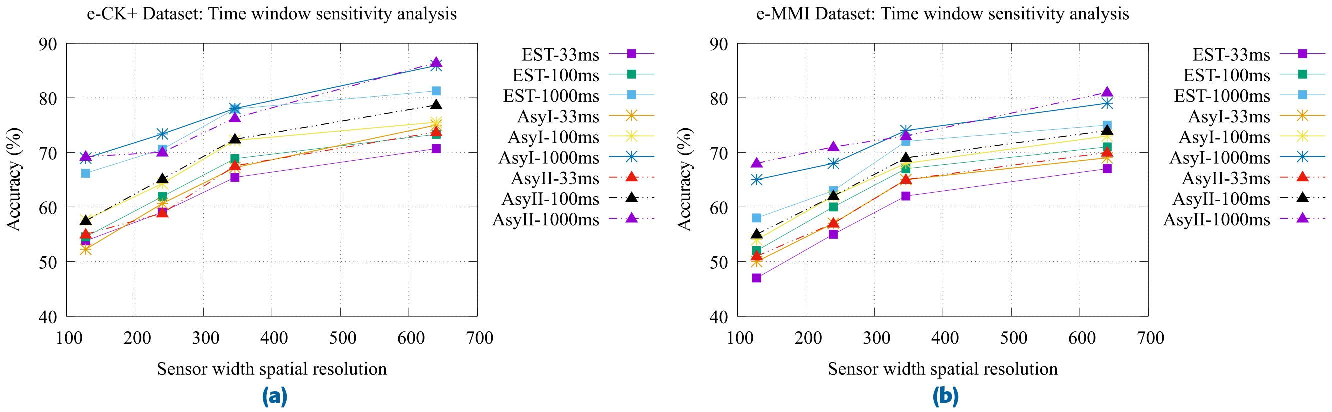


FIGURE 14. Performance evaluation of three state-of-the-art event-based recognition methods, varying the temporal length of the sample (10ms, 33ms, 100ms) and the width spatial resolution (128, 240, 346, 640), on the two new event databases: (a) the e-CK+ Dataset; (b) the e-MMI Dataset. The number of samples per class has been set at 3000. EST refers to [5], AsyI to Asynet Dense [6], AsyII to Asynet Sparse [6].

B. EVENT-BASED FACIAL EXPRESSION RECOGNITION

In the following subsections, we will analyse event-based facial expression recognition methods using two new datasets: e-CK+ and e-MMI. In these data sets, we carefully portioned the data into time windows of different sizes (33 ms, 100 ms, and 1000 ms) to provide a comprehensive examination. Since both databases are emulated, we also analysed the spatial resolution of the generated events to identify if this parameter impacts the classification method learning.

1) e-CK+ DATASET AND e-MMI DATASET

This subsection provides the results of the comparative analysis of three event-based recognition methods –EST, AsyI, and AsyII– for the facial expression recognition task. The results are presented in Fig. 14a and Fig. 14b for the new e-CK+ and e-MMI datasets, respectively. The number of samples per class has been set at 3000.

Based on the previous results, the following analyses can be inferred. Firstly, and independently of the spatial resolution of the events, it is evident that the larger the temporal size of the event sample, the better the performance of the classifiers under study: EST, AsyI and AsyII. In particular, the latter two methods show the best accuracy with time windows of 1000ms, for the e-CK+ and e-MMI databases. However, when compared to gesture recognition, the optimal temporal sample size for facial expression recognition is considerably larger, and this is because both databases were emulated at a timestamp resolution of 1ms, and not at a resolution of microseconds (μs) as it is in real event cameras. Furthermore, the spatiotemporal evolution of facial expressions requires a longer analysis time than gestures, as the full sequence of deformable object dynamics needs to be adequately described.

As regards the spatial resolution of events, it is clear that algorithms achieve better learning rates as the spatial resolution increases. This is consistent with facial expression recognition since identifying the corresponding facial expressions requires a detailed analysis of each fiducial point. Given that fiducial points are usually of very low resolution, having a larger spatial area will allow for better analysis and identification of each facial expression.

Finally, referring to the performance metrics, it is evident that neither of the two problems achieves optimal accuracy. Specifically, for e-CK+, the best method (AsyII-1000ms with a spatial resolution of 640 × 480) fails to surpass 90% accuracy, and for e-MMI, the best and same previous method fails to reach 85% accuracy. For a recognition task, these metrics are poor. However, they can be explained by the fact that they are events emulated from frames, which do not capture the full nature of the dynamics of facial expressions (either by temporal and/or spatial resolution of the events). However, as a first approach in this line of research, these values become a baseline that we will try to overcome in the following subsection.

2) LEARNING FACIAL EXPRESSIONS FROM THE TEMPORAL SEQUENCE

The achieved performance with the three state-of-the-art event-based object recognition methods indicates a good learning rate for the e-CK+ and e-MMI datasets. But again, can this learning rate be improved? By training from the time sequence of the event samples, it is possible.

For this, we employ the ESTM method described earlier. Regarding the sequential input, three 10 ms time windows and three 33 ms time windows are provided to the ESTM, as shown in Table 5.

Compared to the previous results (Fig. 14a and Fig. 14b), Table 5 shows significantly better performance in accuracy rates for event-based facial expressions recognition tasks, for both the e-CK+ –89.1%– and e-MMI –83.7%– datasets.

The use of ESTM captures the temporal correlation of the sequence of movements of facial expressions. However, neither of the two methods manages to surpass the previously mentioned thresholds (90% for e-CK+ and 85% for e-MMI), which could be a limitation

TABLE 5. Performance evaluation of the ESTM event-based facial expression recognition method on the: e-CK+ Dataset, e-MMI Dataset. The number of samples per class has been set at 3000. The spatial resolution has been set at 640 × 480.

Method	30ms (3 x 10ms)	100ms (3 x 33ms)
e-CK+	88.5	**89.1**
e-MMI	80.2	**83.7**

associated with generating emulated databases. Eventually, increasing the temporal size of the event samples may improve the performance of the classifiers. Nevertheless, the emulated events probably suffer the loss of valuable information associated with the evolution of the face fiducial points, affecting the learning curve of the different recognition methods under study.

In any case, it is relevant to highlight that this work opens new possibilities for developing event-based facial expression recognition systems in low light and high contrast environments, setting a baseline performance for future developments and proposals.

C. TRAINING TIMES

The massive event stream processing implies a high requirement of computational resources and long training times. We used an NVIDIA®DGX-1™, a purpose-built system optimised for deep learning. The specifications are presented in Table 6, where GPU cores were used for event-based database emulation (1 week for e-CK+, 1 week for e-MMI) and the application of several classifier training techniques.

TABLE 6. Hardware specifications used for experiments and tests in event-based facial expression recognition.

Hardware	Accelerator	Architecture	Boost clk
NVIDIA Dgx-1	V100	Volta	1530 MHz
Memory Clock	Bandwidth	VRAM	GPU
1.75Gbit/s	900GB/sec	32GB	GV100

Regarding inference computation times, we used an NVIDIA®GeForce GTX 1050 4GB, with the results for the different methods listed in Fig. 7. We can observe that the ESTM method takes 31.9 ms (29.8 ms more than EST and 8.5 ms more than Asynet).

TABLE 7. Methods performance. Inference times for event-based gesture and facial expression recognition methods.

Method	Reference	DGX-1 t_{exec}
EST	[5]	2.1 [ms]
Asynet	[6]	23.4 [ms]
ESTM	Our implementation	31.9 [ms]

VI. CONCLUSION

This article presents a comparative analysis of deformable object recognition, specifically focusing on gestures and facial expressions, through spatial and temporal analysis of events. As a contribution, it includes an evaluation of state-of-the-art event-based methods in deformable object recognition and introduces two new facial expression databases: e-CK+ and e-MMI.

Two highly performing methods in object recognition are selected from the state-of the-art: EST and ResNet-34 [5], and Asynet [6] (with dense and sparse convolutions). These methods are evaluated on benchmark databases such as IBM DVS 128 Gesture Dataset and NavGesture (for gestures) as well as e-CK+ and e-MMI (for facial expressions), yielding different results depending on the analysed variable. To enhance the architecture's performance proposed by [5], LSTM memory units are incorporated to capture the temporal dependency of event sequence samples in a novel architecture called ESTM. We believe that the increased accuracy of ESTM when evaluated on the various databases, compared to the original method, is due to EST's ability to characterise the spatio-temporal distribution of events properly, the CNN's capability to extract relevant features, and the LSTM's modelling of the temporal sequence.

The comparative analysis highlights how the performance of deformable object recognition methods is affected by the number of events in a sample, the temporal window size of the sample, and the spatial resolution of the events. For all these variables, it is observed that increasing these variables leads to enhanced accuracy in both gesture and facial expression classification. In the case of the number of events, this is intuitive: having a sample with more events provides more relevant information for the task. Similarly, regarding the spatial resolution of events, having sensors with higher spatial dimensionality allows for better identification of relevant structures. However, in the case of the temporal window size of the sample, it should be noted that there exists an optimal value that yields the best performance for classifiers (33ms). Values lower (10ms) or higher (100ms) than this optimal value result in poorer performance of the method, which may be due to the representation's inability to properly characterise the spatio-temporal distribution of events or the presence of a deficit or surplus of events that hinder the recognition of the respective class. Consequently, generating event samples appropriately, considering the nature of the problem, is emphasised. The choice between a time window and an event window plays a significant role depending on the specific task.

Compared with the state-of-the-art, the evaluated method (ESTM) has two analyses. In gesture recognition, ESTM achieves a 99.5% accuracy for the IBM DVS128 Gestures database, being surpassed only by Inception3D+Voting [39], which reports a 99.62% accuracy. Regarding facial expression recognition, ESTM reports an 89.1% accuracy for e-CK+ and an 83.7% accuracy for e-MMI. While these performances may be low for recognition methods, they are notable results in a field where new results are not reported. Additionally, increasing the spatial resolution of events led to a more than 30% improvement in ESTM's performance on both databases.

For future work, the goal is to replicate this comparative methodology in similar recognition tasks by developing databases using different real neuromorphic sensors (not emulated). This will enable a more comprehensive study of the influence of various factors, such as spatial resolution, temporal size, and the number of events in samples, on the performance of state-of-the-art and novel methods. Furthermore, new research lines in event-based deformable object analysis, specifically facial analysis, and examining influencing factors are intended to be explored.

ACKNOWLEDGMENT

The research in this article uses the MMI-Facial Expression Database collected by Valstar [49] and Pantic [46]. Also, this work uses the CK and CK+ Database collected by Kanade [48] and Lucey [45].

REFERENCES

[1] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 × 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Mar. 2008.

[2] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.

[3] S. Jia, "Event camera survey and extension application to semantic segmentation," in *Proc. 4th Int. Conf. Image Process. Mach. Vis. (IPMV)*. New York, NY, USA: Association for Computing Machinery, Mar. 2022, pp. 115–121, doi: 10.1145/3529446.3529465.

[4] X. Zheng, Y. Liu, Y. Lu, T. Hua, T. Pan, W. Zhang, D. Tao, and L. Wang, "Deep learning for event-based vision: A comprehensive survey and benchmarks," 2023, *arXiv:2302.08890*.

[5] D. Gehrig, A. Loquercio, K. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5632–5642.

[6] N. Messikommer, D. Gehrig, A. Loquercio, and D. Scaramuzza, "Event-based asynchronous sparse convolutional networks," in *Computer Vision—ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 415–431.

[7] S. Barua, Y. Miyatani, and A. Veeraraghavan, "Direct face detection and video reconstruction from event cameras," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[8] G. Lenz, S.-H. Ieng, and R. Benosman, "Event-based face detection and tracking using the dynamics of eye blinks," *Frontiers Neurosci.*, vol. 14, p. 587, Jul. 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2020.00587

[9] B. Ramesh and H. Yang, "Boosted kernelized correlation filters for event-based face detection," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Mar. 2020, pp. 155–159.

[10] A. N. Angelopoulos, J. N. P. Martel, A. P. Kohli, J. Conradt, and G. Wetzstein, "Event-based near-eye gaze tracking beyond 10,000 Hz," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 5, pp. 2577–2586, May 2021.

[11] C. Ryan, B. O'Sullivan, A. Elrasad, A. Cahill, J. Lemley, P. Kielty, C. Posch, and E. Perot, "Real-time face & eye tracking and blink detection using event cameras," *Neural Netw.*, vol. 141, pp. 87–97, Sep. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608021001076

[12] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[13] R. Jafri and H. R. Arabnia, "A survey of face recognition techniques," *J. Inf. Process. Syst.*, vol. 5, no. 2, pp. 41–68, 2009.

[14] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220316945

[15] R. Ghosh, A. Mishra, G. Orchard, and N. V. Thakor, "Real-time object recognition and orientation estimation using an event-based camera and CNN," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2014, pp. 544–547.

[16] H. Li, G. Li, and L. Shi, "Classification of spatiotemporal events based on random forest," in *Advances in Brain Inspired Cognitive Systems*, C.-L. Liu, A. Hussain, B. Luo, K. C. Tan, Y. Zeng, and Z. Zhang, Eds. Cham, Switzerland: Springer, 2016, pp. 138–148.

[17] G. K. Cohen, G. Orchard, S.-H. Leng, J. Tapson, R. B. Benosman, and A. van Schaik, "Skimming digits: Neuromorphic classification of spike-encoded images," *Frontiers Neurosci.*, vol. 10, p. 184, Apr. 2016. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2016.00184

[18] Q. Liu, H. Ruan, D. Xing, H. Tang, and G. Pan, "Effective AER object classification using segmented probability-maximization learning in spiking neural networks," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 2, pp. 1308–1315. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5486

[19] Y. Li, H. Zhou, B. Yang, Y. Zhang, Z. Cui, H. Bao, and G. Zhang, "Graph-based asynchronous event processing for rapid object recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 914–923.

[20] Z. Wang, Y. Hu, and S.-C. Liu, "Exploiting spatial sparsity for event cameras with visual transformers," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 411–415.

[21] L. A. Camuñas-Mesa, B. Linares-Barranco, and T. Serrano-Gotarredona, "Low-power hardware implementation of SNN with decision block for recognition tasks," in *Proc. 26th IEEE Int. Conf. Electron., Circuits Syst. (ICECS)*, Nov. 2019, pp. 73–76.

[22] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[24] G. Orchard, A. Jayawant, G. K. Cohen, and N. Thakor, "Converting static image datasets to spiking neuromorphic datasets using saccades," *Frontiers Neurosci.*, vol. 9, p. 437, Nov. 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2015.00437

[25] T. Serrano-Gotarredona and B. Linares-Barranco, "Poker-DVS and MNIST-DVS. Their history, how they were made, and other details," *Frontiers Neurosci.*, vol. 9, p. 481, Dec. 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2015.00481

[26] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "HATS: Histograms of averaged time surfaces for robust event-based object classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1731–1740.

[27] J. Kim, J. Bae, G. Park, D. Zhang, and Y. M. Kim, "N-ImageNet: Towards robust, fine-grained object recognition with event cameras," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2146–2156.

[28] S. B. Shrestha and G. Orchard, "SLAYER: Spike layer error reassignment in time," in *Advances in Neural Information Processing Systems*, vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2018, pp. 1419–1428. [Online]. Available: http://papers.nips.cc/paper/7415-slayer-spike-layer-error-reassignment-in-time.pdf

[29] J. Kaiser, H. Mostafa, and E. Neftci, "Synaptic plasticity dynamics for deep continuous local learning (DECOLLE)," *Frontiers Neurosci.*, vol. 14, p. 424, May 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2020.00424

[30] R. W. Baldwin, R. Liu, M. Almatrafi, V. Asari, and K. Hirakawa, "Time-ordered recent event (TORE) volumes for event cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2519–2532, Feb. 2023.

[31] Y. Bi, A. Chadha, A. Abbas, E. Bourtsoulatze, and Y. Andreopoulos, "Graph-based spatio-temporal feature learning for neuromorphic vision sensing," *IEEE Trans. Image Process.*, vol. 29, pp. 9084–9098, 2020.

[32] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbrück, M. Flickner, and D. Modha, "A low power, fully event-based gesture recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7388–7397.

[33] S. A. Baby, B. Vinod, C. Chinni, and K. Mitra, "Dynamic vision sensors for human activity recognition," in *Proc. 4th IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2017, pp. 316–321.

[34] J.-M. Maro, S.-H. Ieng, and R. Benosman, "Event-based gesture recognition with dynamic background suppression using smartphone computational capabilities," *Frontiers Neurosci.*, vol. 14, p. 275, Apr. 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2020.00275

[35] X. Chen, J. Wang, L. Zhang, S. Guo, L. Qu, and L. Wang, "Real-time gesture classification system based on dynamic vision sensor," in *Neural Information Processing*, H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, Eds. Cham, Switzerland: Springer, 2020, pp. 486–497.

[36] J. Kaiser, A. Friedrich, J. C. V. Tieck, D. Reichard, A. Roennau, E. Neftci, and R. Dillmann, "Embodied neuromorphic vision with continuous random backpropagation," in *Proc. 8th IEEE RAS/EMBS Int. Conf. Biomed. Robot. Biomechatronics (BioRob)*, Nov. 2020, pp. 1202–1209.

[37] Q. Wang, Y. Zhang, J. Yuan, and Y. Lu, "Space-time event clouds for gesture recognition: From RGB cameras to event cameras," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1826–1835.

[38] A. Sabater, L. Montesano, and A. C. Murillo, "Event transformer. A sparse-aware solution for efficient event data processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2676–2685.

[39] S. U. Innocenti, F. Becattini, F. Pernici, and A. D. Bimbo, "Temporal binary representation for event-based action recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10426–10432.

[40] G. Moreira, A. Graça, B. Silva, P. Martins, and J. Batista, "Neuromorphic event-based face identity recognition," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 922–929.

[41] F. Becattini, F. Palai, and A. D. Bimbo, "Understanding human reactions looking at facial microexpressions with an event camera," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9112–9121, Dec. 2022.

[42] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. 2nd Conf. Robot Learn.*, vol. 87, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., Oct. 2018, pp. 969–982. [Online]. Available: https://proceedings.mlr.press/v87/rebecq18a.html

[43] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic DVS events," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1312–1321.

[44] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-Dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, and Y. Gu, "A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications," *J. Big Data*, vol. 10, no. 1, p. 46, Apr. 2023.

[45] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[46] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Amsterdam, The Netherlands, Jul. 2005, pp. 317–321.

[47] R. Plutchik, "A psychoevolutionary theory of emotions," *Social Sci. Inf.*, vol. 21, nos. 4–5, pp. 529–553, Jul. 1982, doi: 10.1177/053901882021004003.

[48] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 46–53.

[49] M. F. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," in *Proc. Int. Conf. Lang. Resour. Eval., Workshop EMOTION*, Valletta, Malta, May 2010, pp. 65–70.

RODRIGO VERSCHAE (Member, IEEE) received the Diploma degree in electrical engineering from Universidad de Chile, Santiago, Chile, in 2003, the M.S. degree in applied mathematics from École Normale Supérieure Paris-Saclay, France, in 2006, and the Ph.D. degree in electrical engineering from Universidad de Chile, in 2010. He is currently an Associate Professor with the Institute of Engineering Sciences, Universidad de O'Higgins, Chile. Prior to the current position, he was the Director of the Institute of Engineering Sciences, Universidad de O'Higgins, and an Assistant Professor with Kyoto University. He has published more than 70 peer-reviewed publications. His research interests include computer vision, robotics, pattern recognition, and machine learning, with various applications, including agriculture, energy, health, and education. He serves as an Associate Editor for *Information* (MDPI) and *Frontiers in Robotics and AI* and an Advisory Board Member for *Sci* (MDPI).

IGNACIO BUGUENO-CORDOVA (Member, IEEE) received the degree in electrical engineering, in 2019. He is currently pursuing the master's degree with the Artificial Intelligence and Robotics Laboratory, Department of Electrical Engineering, Universidad de Chile, Chile. He is a Research Assistant with the Robotics and Intelligent Systems Laboratory, Universidad de O'Higgins. His research interests include artificial intelligence, computer vision, robotics, machine learning, and telecommunications.