Pattern Analysis and Applications (2022) 25:505–520
https://doi.org/10.1007/s10044-021-01011-w

ORIGINAL ARTICLE

SL-Animals-DVS: event-driven sign language animals dataset

Ajay Vasudevan[1] · Pablo Negri[2,3] · Camila Di Ielsi[2] · Bernabe Linares-Barranco[1] · Teresa Serrano-Gotarredona[1]

Received: 28 August 2020 / Accepted: 7 July 2021 / Published online: 16 July 2021

Abstract
Non-intrusive visual-based applications supporting the communication of people employing sign language for communication are always an open and attractive research field for the human action recognition community. Automatic sign language interpretation is a complex visual recognition task where motion across time distinguishes the sign being performed. In recent years, the development of robust and successful deep-learning techniques has been accompanied by the creation of a large number of databases. The availability of challenging datasets of Sign Language (SL) terms and phrases helps to push the research to develop new algorithms and methods to tackle their automatic recognition. This paper presents 'SL-Animals-DVS', an event-based action dataset captured by a Dynamic Vision Sensor (DVS). The DVS records non-fluent signers performing a small set of isolated words derived from SL signs of various animals as a continuous spike flow at very low latency. This is especially suited for SL signs which are usually made at very high speeds. We benchmark the recognition performance on this data using three state-of-the-art Spiking Neural Networks (SNN) recognition systems. SNNs are naturally compatible to make use of the temporal information that is provided by the DVS where the information is encoded in the spike times. The dataset has about 1100 samples of 59 subjects performing 19 sign language signs in isolation at different scenarios, providing a challenging evaluation platform for this emerging technology.

Keywords Event-based dataset · Sign language · Spiking neural network architectures

1 Introduction

Sign Language (SL) has a complexity far beyond other traditional gesture problems faced by computer sciences, such as Human-Robot interaction, traffic signals. It allows a signer to express sentiments, poetry, science, temporal changes, and have, at the same time, limitless nuances of meanings. This complex morphology is achieved based on both a complex co-occurrence of visuospatial patterns and the skills of the signer [7].

Even though SL is widely employed by the hard of hearing community, which is estimated to be 466 million according the World Health Organization [46], most individuals are not knowledgeable in SL, which translates into lesser opportunities in education and employment. Many efforts have been made worldwide to facilitate the learning of SL for non-signers. For example, China has launched a standardization of the Chinese Sign Language [30], in order to group the several dialects in their country. The Center for Sign Language (University College Capital) has documented the Danish Sign Language [38], which is available online since 2008 is another example of progress in this field. The recently published ASL-LEX project is the largest and most complete publicly available repository of information about the American Sign Language lexicon [10].

Automatic Sign Language Recognition (ASLR) applications arise in the field of human-computer interaction. Its main objective is to develop complex frameworks integrating different kinds of sensors, computer systems, and machine-learning algorithms devoted to capture, analyze, and translate sign language. One example is translation systems with the ability to detect continuous signed sentences

✉ Pablo Negri
pnegri@dc.uba.ar

Ajay Vasudevan
ajay@imse-cnm.csic.es

[1] Instituto de Microelectronica de Sevilla (IMSE-CNM), CSIC and Universidad de Sevilla, Sevilla, Spain

[2] Computer Department, University of Buenos Aires (UBA-FCEN), Buenos Aires, Argentina

[3] Institute of Research in Computer Sciences (ICC), CONICET-UBA, Buenos Aires, Argentina

and translate them into grammatical sentences in another language (e.g., written English), the same as translating spoken language into grammatical sentences in a sign language. Other example involves the growing presence of assistive humanoid robots deployed in airports, metro, hospitals, etc., which can bring information to people in general, and should be able to understand sign language. Also, ASLR can be useful in educational frameworks and assist the learning and training of Sign Language for non-signers.

Technically, SL delivers information to the receiver as a compound of hand and body gestures which also involves facial expressions, lip motions, body postures, and upper limbs within a 3D space in front of the signer. Despite differences with spoken language, transmission rate for speech and sign is the same in the case of fluent signers [14, 16]. For example, when interacting with pre-lingually deaf persons, who were either born deaf or lost their hearing before they learned their native spoken language, their fluent communication have a very high speed and can be very hard to follow. Even for people who know SL.

To capture the motion of the signer, we propose in this work the use of an neuromorphic event-driven Dynamic Vision Sensor (DVS) camera [25, 34, 35]. By operating in a continuous and asynchronous way, the pixels in the retina of the event-based sensing system respond to changes in the illumination. The main advantage in using a DVS is tackling the problem of high speed actions with very reduced latency without having to wait for full frames [33]. Furthermore, DVS outputs are naturally processed by Deep Spiking Neural Networks (SNN) [27]. Also emerging nano-scale devices that can learn and process through spikes (events) are capable of high-level cognitive computations [8, 17, 39]. Thus, DVS sensors and neuromorphic hardware systems would be better suited to follow the high speed signs than traditional frame-based video sensors and recognition systems, and can be combined with other visual sensors devoted on other tasks, i.e., lips reading. However, this new technology should bridge the performance gap in comparison with traditional video gesture recognition, which has a very long development history, successful methodologies, and more populated evaluation datasets.

In this work, we analyze in detail a new Sign Language dataset of isolated words recorded by a DVS. It is composed of more than 1100 samples of 59 subjects performing 19 signs in isolation corresponding to animals, and denominated as **SL-Animals-DVS** [3, 40]. We train and test three state-of-the-art Spiking Neural Network approaches on this dataset: DECOLLE [21], SLAYER [36] and STBP [47]. Lastly, we also give an estimation of the power consumption by the network using TrueNorth hardware.

This paper is a revised and expanded version of the paper [40] presented at the 1st. Workshop on Faces and Gestures in E-health and Welfare (FaGEW) of the IEEE International Conference on Automatic Face and Gesture Recognition 2020. In this work, we add details of the data produced by the DVS and the structure of the sign language. A third state-of-the-art Spiking Neural Network algorithm is implemented to evaluate their performance in the SL-Animals-DVS. Also, we benchmark the performance of all the SNN models on another publicly available dataset, against the SL-Animals-DVS accuracy.

The rest of the paper is organized as follows. In Sect. 2, we review datasets involving language signal and gestures in general. Section 3 analyzes the **SL-Animals-DVS** dataset and Sect. 4 describes traditional and new architectures of classifiers experimenting on the dataset. Results are discussed in Sect. 5, and, finally, conclusions are drawn in Sect. 6.

2 Related works

2.1 Sign language and datasets from traditional sensors

The recognition of Sign Language actions has caught the attention of the research community for decades. Several SL datasets captured by vision sensors were developed to train automatic recognition systems and are available online. Some of them involve the recognition of only isolated words or phrase levels [29], or continuous type corresponding to complete sentences or full coherent thoughts [41, 49].

Two approaches are mainly employed to tackle the problem: sensor-based and vision-based [13]. Sensor-based requires the use of electronic instruments to track the hands and measure their motion, position and speed. The DataGlove is one example of this kind of hardware sensor. It consists on five flex sensors for the fingers and includes one accelerometer with a gyro sensor on the back of the palm. Mori et al. [32] use the Data-Glove to collect a corpus of 20 words performed by 8 persons and employ an heuristic methodology for the classification.

The vision-based approaches record the signs as static images or video sequences,[1] were the classic recording scenario consists of a single-camera recording the signers [6]. Furthermore, because SL evolves in a 3D space, the depth of the scene can be captured by active techniques that project structured light, such as Microsoft Kinect, which provides RGB-D ('D' means depth information) videos, or Intel RealSense.

Purdue RVL-SLLL Database [29] consists of 14 fluent deaf signers performing different types of handshapes

[1] Because we are interested in the analysis of a stream of temporal events, we will not consider static images.

in isolation and single motion primitives (39 cases), ASL alphabet and numbers (69 cases), and 10 phrases with different prosody meanings. In total, the dataset is composed of 2576 RGB videos captured under 2 lighting conditions: diffuse illumination to suppress shadows, and direct illumination to enhance contrast.

The MMI-Database, also known as SIGNUM, is a German Sign Language database [41]. It has on 15,600 video recordings of 20 native signers reproducing 780 sentences based on 450 primitive signs. Each sentence is meaningful, grammatically well-formed and ranges from two to eleven signs in length. The recordings sessions occurred under controlled environment with diffuse lighting and a blue background.

In [24], 20 categories of the most commonly employed Sign Language signs at the museum of China were collected by a Microsoft Kinect sensor on RGB-D videos, and incorporate infrared, contour and skeleton information of each recording. This dataset, also known as SLVM, has a corpus of 6800 samples, corresponding to 17 subjects repeating 20 times the same sign.

The LIBRA-10 dataset [4] consists of 10 dynamic signs from Brazilian Sign Language. For each class, there are approximately 300 depth dynamic videos collected using the Intel-RealSense sensor; however, they do not give the number of persons involved in the experiment. Cerna et al. [11] introduce LIBRAS-UFOP, a publicly available challenging dataset of Brazilian Sign Language captured by Kinet sensor. It provides a complete multimodal data (RGB-D and squeleton) recording of 5 fluent signers performing 56 representative words several times. The set of 3040 data sequences is divided into 4 categories indicating the intra-class correlation: such as motion, articulation point and hand configuration.

Yuan et al. [49] recently introduced the Chinese Sign Language Dataset (CSLD). It consists on 50,000 RGB-D video recordings captured by a Microsoft Kinect sensor, from 25 female and 25 male deaf students. Each signer performs 1000 phrases randomly selected from a gallery of 10,000 Chinese utterances which correspond to lifestyle, education and common everyday life.

The state-of-the-art includes other large datasets, but they all lack user-independent scenario, as the number of signers is very limited, such as [19, 43], with 9 signers, or only 4 subjects for RWTH-BOSTON-400 [15]. Even if the training of automatic sign recognition using this sets will not be necessarily robust to new instances [41], they contribute with a wide set of sentences which can be useful as posterior evaluations of pre-trained systems.

ChaLearn Datasets [18, 42] provide a corpus of body-hand gestures RGB-D videos with a similar kind of Sign Language. In [18], they proposed the Montalbano dataset, where each signer was recorded speaking fluent italian and performing natural communicational gestures. The corpus consists in 13,858 sequences, where 20 sign classes are performed by 27 subjects. Wan et al. extend in [42] precedent ChaLearn databases, to obtain a large dataset of 249 classes of signs from several lexicons or vocabularies, such as sign language, underwater signs, helicopter and traffic signals, pantomimes, symbolic gestures, italian gestures and body language. The dataset has more 50,000 RGB-D videos from 21 subjects, which were split into Isolated and Continuous sets for different recognition tasks.

2.2 Event-based gestures datasets

Because Dynamic Vision Sensor (DVS) is a nascent vision technology, the development of spiking automatic learning systems has drawn the attention of many scientific groups. The performances reported on different recognition tasks are still in evolution, but far below that obtained by frame-based sensors. There exist only a few publicly available databases of hand gestures [12, 26, 28] and body gestures [5] captured by DVS, but there is none related with sign language. This also shows the need of challenging neuromorphic datasets oriented to support further development of this technology.

ROSHAMBO17 [26] is a DVS dataset of hand gestures involving the rock, scissors and paper symbols performed by 15 subjects with a variety of positions, poses, distances, etc. The dataset also included background and non-signed frames as a fourth class (noise). The spike flows were sampled into a frame representation accumulating fixed 0.5K, 1K, 2K, and 4K events. The corpus has 5 million 64×64 pixels size images.

Maro and Benosman [28] propose a dataset for hand gestures devoted to operate a smartphone with a DVS. Their dictionary has six basic gestures recorded in real conditions: One hand holding the smartphone, while the other hand performs the movement and is captured by the DVS. A total of 35 subjects were recorded under sitting and walking conditions at indoors or outdoors. Among them, 12 people are visually impaired subjects. The dataset contains 1621 clips.

Neuro ConGD Dataset [12] recorded sixteen hand gesture classes along with an additional class blank. This database has the problem of subject-independent bias as well, since only 8 persons were recorded for the 2040 instances.

The IBM DVS-GESTURE dataset is proposed by [5] and consists of 11 hands and body gestures, not related with SL, from 29 subjects under 3 illumination conditions. The gestures include **hands clap**, **left arm clockwise**, **right hand wave**, etc. The state-of-the-art widely employs DVS-GESTURE to evaluate the performance for gesture recognition on different Spiking Neural Networks architectures and learning approaches [21, 36, 44, 47, 48]. Best result reported by the authors is 94.59%, and SLAYER reports an accuracy of 93.64% on average, which shows that the dataset

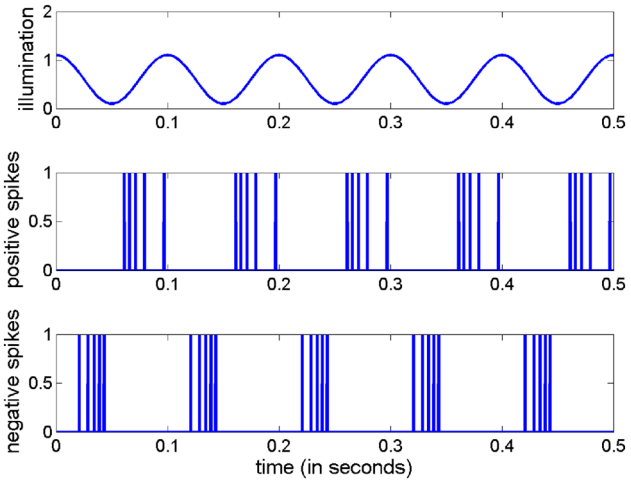


Fig. 1 Simulation of the operation of a DVS pixel when stimulated with sinusoidal illumination (upper subfigure). The pixel will generate output events or spikes when the illumination increases (middle subfigure) and negative events when the illumination decreases (lower subfigure)

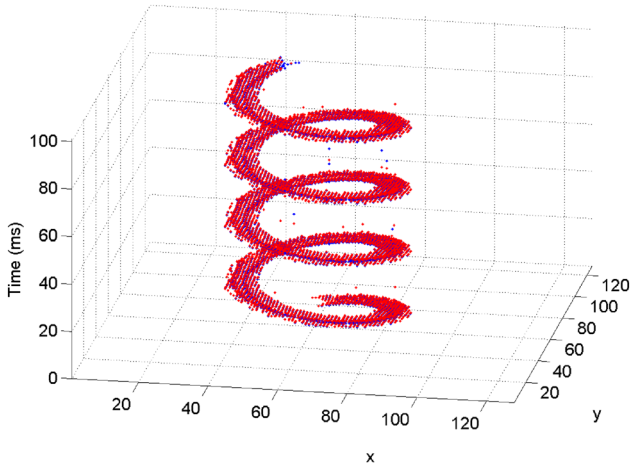


Fig. 2 3-dimensional representation of the events generated by a cross symbol rotating in front of a DVS sensor at 40 Hz

is fast saturating in performance and more challenging sets are needed. Furthermore, in following sections, we also analyze some drawbacks on the constitution of this dataset, from the point of view of intra-class variability.

Section 3 details the **SL-Animals-DVS** dataset. To collect the samples, we take into consideration some particularities that were exposed in this review. First, the number of different signers is greater than most of the publicly available sets. Also, the samples correspond to different scenarios and lightning conditions which incorporate diversity on input data. Furthermore, the nature of the signs do not have a spatial bias as we detected in DVS-GESTURE (see Sect. 3.5). Lastly, the number of signs is higher in comparison with other DVS datasets.

3 Experimental dataset

This section describes the data provided by the DVS technology, and presents and analyzes our **SL-Animals-DVS** experimental dataset.

3.1 DVS event information

For acquiring the dataset, we used a DVS sensor with a 128 × 128 resolution retina developed at IMSE [35]. Each pixel in the DVS sensor responds to local temporal changes in the illumination impinging on the pixel. Figure 1 illustrates the operation of a DVS pixel under an illumination intensity which follows a sinusoidal wave of 10Hz frequency. The pixel will generate a number of output events when the relative change in the illumination goes over a certain threshold; therefore, the density of output events is proportional to the relative illumination change. The pixels generate signed events. So that, when the illumination increases, the pixel generates positive events (middle subfigure), whereas when the illumination decreases the pixel generates events with negative polarity (lower subfigure). The pixels will generate no spike under static illumination. That way, by segregating dynamic foreground from a static background and reducing the amount of data, the DVS sensor generates a compressed information to be processed. This characteristic makes the DVS sensor especially suited for efficient high-speed recognition of dynamic gestures.

The DVS camera outputs events in the Address Event Representation (AER) protocol which is used to transmit signals between neuromorphic systems asynchronously [37]. Every event in AER is composed of the time of the event, the x, y coordinates of the pixel location generating the event and the polarity of the event: $s_i = \{t_i, x_i, y_i, \text{pol}_i\}$. The minimum and maximum value of pixel location coordinates x_i and y_i is constrained by the dimension of the retina size which is 128 × 128. pol_i takes value 1 or − 1 indicating a positive or negative polarity event occurrence which corresponds to positive or negative change in the illumination at that location. Figure 2 plots the events s_i generated by a dark cross symbol painted on a white background which is rotating in front of a DVS sensor at 40 Hz during 100 ms. The plot represents with red dots the negative events ($\text{pol}_i = -1$) and the positive events ($\text{pol}_i = 1$) are represented with blue dots. The vertical axes represents the spike timing t_i in milliseconds and the horizontal axis are the x_i, y_i coordinates of the events. Figure 3 is the result of histograming for 1 ms the events generated by the cross symbol painted on a rotating disk. The DVS sensor achieves an output rate of 30Meps (Mega events per second). It provides a final time resolution of 1μs

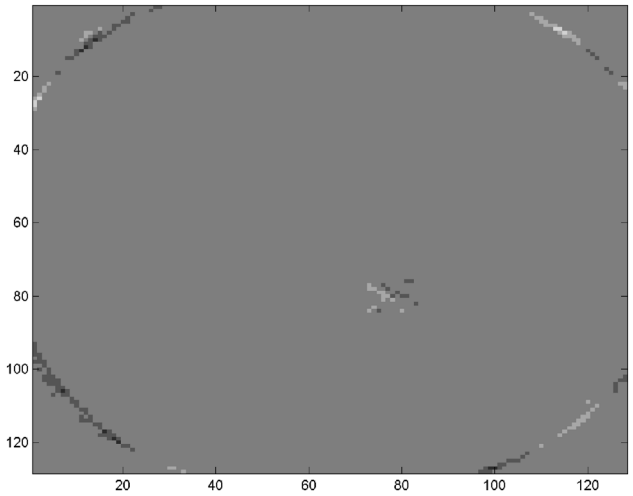


Fig. 3 Histogram of the events generated by the rotating cross symbol painted on a rotating disk for 1 ms

due to the process of capture and timestamps the events. As can be observed, the exact spatio-temporal information of the dynamic contents in the observed scene is preserved with a 1μs precision.

An AER spike flow triggered by the DVS within a window of time is described by $S(\tau) = \{s_1, s_2, \ldots, s_N\}$, where τ corresponds to the temporal window, and N is the number of spikes in the spike flow S. In practice, the starting time stamp when the recording commencement is zero, and finishes at T: $\{\forall\, i,\ 0 < i \leq N \rightarrow t_i \in [0, T]\}$.

The signs are recorded with the jAER software and the data in the AER protocol is suitable to be used in a SNN [2].

3.2 Dataset recording

The corpus of the **SL-Animals-DVS** dataset consists of 59 non-deaf experimental subjects performing 19 Spanish Sign Language signs corresponding to animals. Because all the signers were not familiar with the SL, the set of animals signs proved to be a very intuitive option which justified their choice. The signs were first presented by a model on a YouTube video [9]. Each subject replicates the signs while observing the video in front of the DVS placed at a distance between 2 and 2.5 m, in such a way that all of their upper-body was visible and focused. The signer freely used left or right hand/arm to perform the signs. The DVS records on a single file the one subject following the video. To tag the record, we define the starting and endpoint of a sign when the Youtube model changes the sign demonstration for a new one. The record file is then divided into 19 samples, one sample for each sign. It is worth noting that one sample corresponds to the subject performing several times the same sign, we will analyze this later. The dataset is then composed

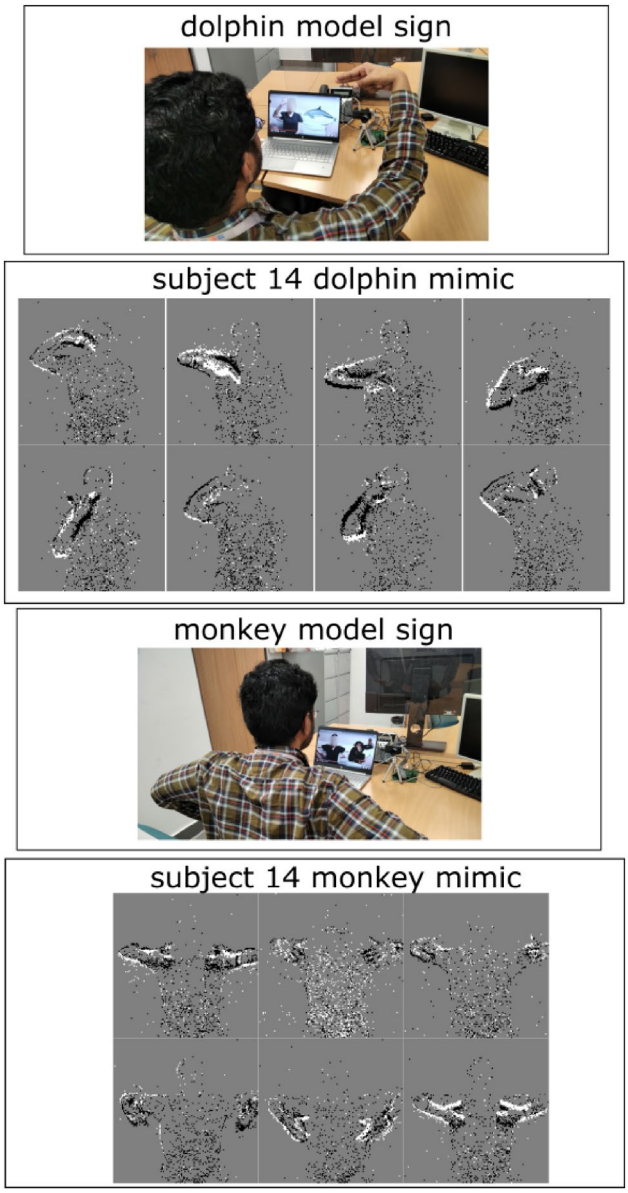


Fig. 4 The figure shows the experimental protocol to capture the signers using the DVS, while they replicate signs performed by the YouTube model [9]

of 59 recordings, and 59 tag files with the starting and end point of the sign samples. Thus, the total number of signs samples reaches 1121.

Figure 4 shows two scenes of [9] of the experimental set-up where the model makes the sign of a dolphin and a monkey, while displaying a picture of the corresponding animal. The sign of the **dolphin** is made by one hand with the thumb up, the index and medium extended, and a regular movement. The subfigure immediately below plots eight histograms obtained by accumulating $N = 2000$ successive events from the DVS recording of subject 14 performing the **dolphin** sign. To compute each histogram, at each spatial

location we accumulate the total number of events occurring at the corresponding location but taking into account the event polarity. In these histograms, the locations with value 0 (receiving no event or the same number of positive and negative events) correspond to a gray level, while the white level corresponds to locations receiving a net contribution of events with positive polarity and the black level corresponds a net contribution of negative polarity.

The lower panel of Fig. 4 illustrates an example of the **monkey** sign as well as a sequence of histograms obtained from the DVS recording of the same individual reproducing the sign.

The recording was conducted in 4 sessions at different locations under different lighting conditions. Set `S1` and `S2` contains 10 and 23 recordings, respectively, and were recorded with natural light coming from a side window. Set `S3` corresponds to 18 users recorded indoors with artificial lightning from a neon light source. Finally, the 7 records of set `S4` were captured indoors under a strong frontal sunlight. The difference between recording session `S1` and recording session `S2` is that in session `S1` we employed a lens with higher angle than the one used for the others sessions. Figure 5 shows two records corresponding to the same sign but performed by two different persons at sessions `S1` and `S2`. In total, **SL-Animals-DVS** is composed of 1121 records corresponding to 59 unique sequences performing the 19 signs and where no person was recorded twice.

Figure 6 is a comparison of sign **bird** being performed under different lighting conditions. It shows the occurrence of events for action performed by a single user across time t_0 to t_8, where all frames histogram the stream of events during the same temporal window. The frames were equally separated in time from the start to the end of the action. As can be observed, in set `S3` the concentration of events happens on the body of the user rather than on the required motion in the hands. We will see on Sect. 5, that this noise, which is due to the indoor lighting conditions reflecting on the patterned clothing of the use, adversely affects the accuracy when the samples from the `S3` are included in the experiments.

3.3 Structure of sign languages

At the lexical level individual signs differ from one another by exploiting the possibilities that provide a wide choice of handshapes, movements and locations. For example, the signs of summer, ugly and dry follow the same handshape and movement. They only differ in the specific body location where the signer develops the movement path (forehead, nose or chin) [7].

In that way, the location of the sign is a first clue which follows a systematic pattern for the recognition. Secondly, the handshape information and lastly the dynamic of the motion are used as clues. In sign recognition studies, it has been observed that the identification of a sign needs at least 35% of development, 240 ms on average, instead of a spoken word which needs 85%, or 330 ms on average [14].

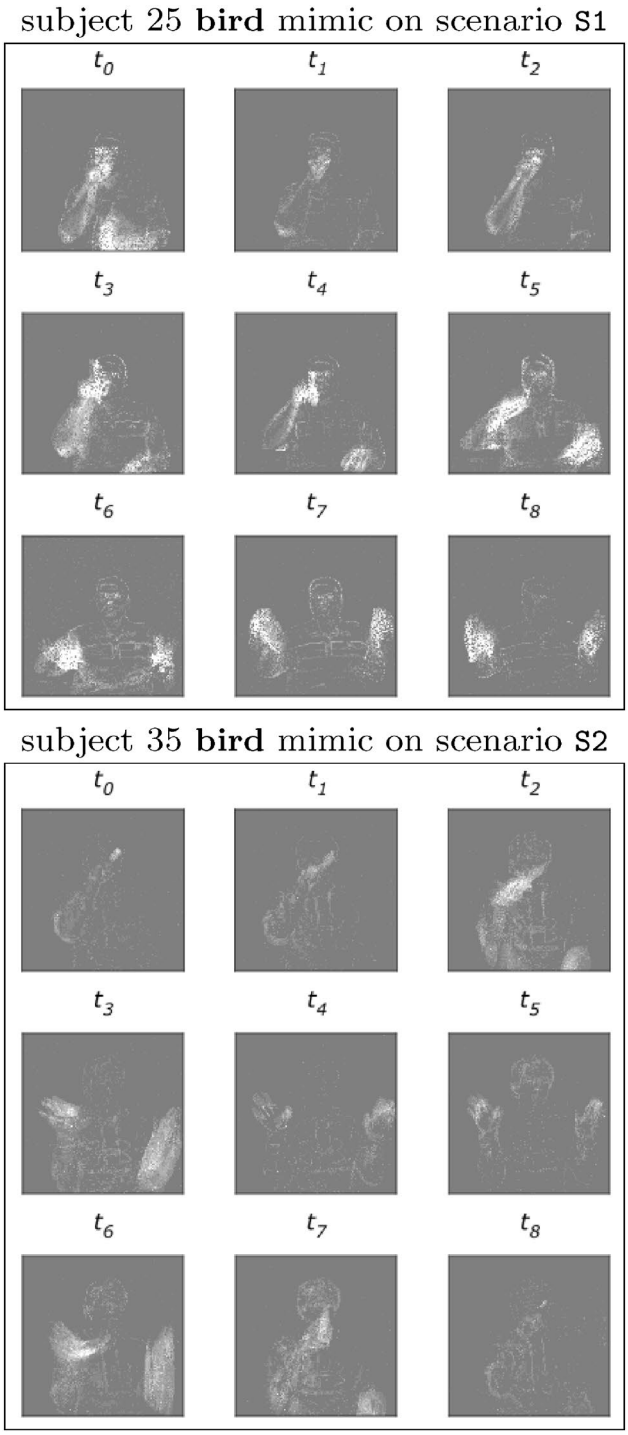


Fig. 5 The figure shows histogram frames of two recordings belonging to two users from sessions `S1` and `S2`. The recordings correspond to the same sign **bird**. Negative polarity was not included in the picture

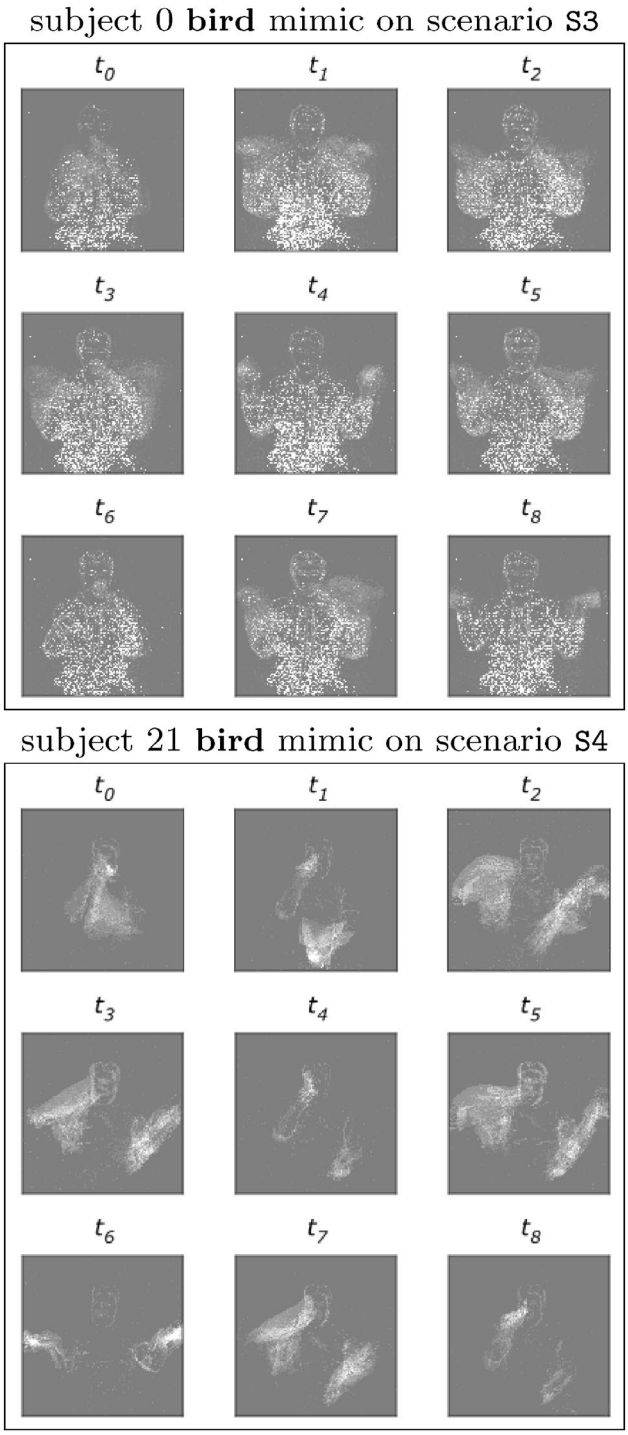


Fig. 6 The figure shows histogram frames of two recordings belonging to two users from sessions S4 and S3. The recordings correspond to the same sign **bird**. Negative polarity was not included in the picture to enhance the noise differences on both scenarios. Note the presence of noise in the sequence recorded in S3 is mostly placed at the clothes, while the events corresponding to the sign motion are present but in fewer number degrading the gesture discrimination capability of the event-based classifier

Several signs in **SL-Animals-DVS** start with a hand translation to a specific body location. Afterward, the handshape and the gesture movement define the corresponding animal sign. For example, in the **lion** sign, both hands perform a wave motion from the forehead side to the bottom.

The pertinent question involves the definition of the starting point of the sign: before the hands reach the starting location, or after that. In this work, we will consider the sign once both hands start moving from the knees (this is the *resting position*) to the starting location, and launch the motion.

In Table 1, detailed description of the signs can be seen, giving a notion of the complexity an automatic classifier is facing in the **SL-Animals-DVS** dataset. It depicts individual information about each sign performed by the Youtube model such as the duration in seconds, and the number of times she repeats the same sign in front of the camera (*Repet* column). The table also performs a detailed analysis of the sign dynamic, such as one hand or two hands motion. A two-hand sign can also be classified as symmetric when both hands perform the same movement at both sides of the body, i.e., the **lion** sign described before. Otherwise, in the **spider** sign one hand moves up and down while the other hand stays in a fixed position. This is a non-symmetric two hand sign. Figure 4 shows the dolphin sing, which is of non-symmetric one hand type, and the monkey sign, which is a symmetric type sign.

It is important to notice that there is no restriction about what hand should be fixed in the **spider** sign, thus, some signers fix the left hand while others the right hand. The *both handed* column in Table 1 points out the signs which were played by a different hand by different signers. This clearly increases the within class variability challenging the automatic classifier.

Body Related columns specifies which signs have a starting point at some specific part of the body (face or chest). These signs are considered as a *compound* category indicating a body spatial relationship.

3.4 Records analysis

Figures 7 and 8 give individual statistics of each sign. The solid line represents the average number of events per second for each sign for the 59 subjects. The dashed line plots the average number of events generated by a sign sequence divided by the number of repetitions performed by each subject.

The analysis of Figs. 8 and 7 exhibits that sequences in session S3, which were captured with no natural light, have double number of events for almost all the signs. This is due to the effect of the frequent switching of the fluorescent lamps, which is sensed by the DVS. Thus, the DVS constantly triggers events from the textured clothes of some signers, instead of detecting the moving parts of

Table 1 The details of the SL-DVS dataset model from [9]

Animal	#Class	Duration (s)	Repet.	1 Hand	2 Hands		Body related		Both-handed
					Symmetric	Non-symmetric	Head	Chest	
Cat	1	6	–	–	–	X	–	–	X
Dog	2	5	10	X	–	–	X	–	X
Camel	3	5	8	–	–	X	–	–	X
Cow	4	4	3	X	–	–	X	–	X
Sheep	5	4	8	–	X	–	–	X	–
Goat	6	5	–	–	X	–	X	–	–
Wolf	7	4	3	–	X	–	X	–	–
Squirrel	8	3	–	–	–	X	–	–	X
Mouse	9	4	4	–	–	X	–	–	X
Dolphin	10	4	3	X	–	–	–	–	X
Shark	11	6	5	–	–	X	–	–	X
Lion	12	4	3	–	X	–	X	–	–
Monkey	13	4	7	–	X	–	–	–	–
Snake	14	4	4	X	–	–	X	–	X
Spider	15	4	3	–	–	X	–	–	X
Butterfly	16	4	–	–	X	–	–	–	–
Bird	17	5	3	–	X	–	X	–	X
Duck	18	3	3	X	–	–	X	–	X
Zebra	19	3	4	X	X	–	–	X	–

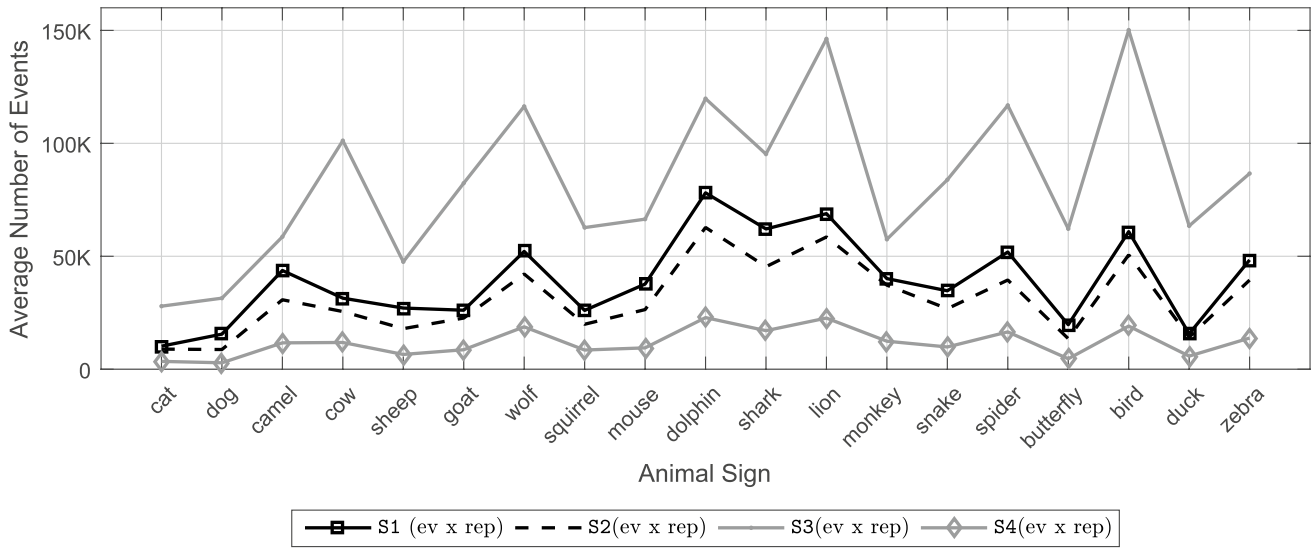


Fig. 7 The figure depicts the average number of events in each repetition for each animal sign at the four recording sessions

the action being performed. On the other side, sequences in `S4` have lower number of events. In that case, the DVS has weakly sensed the moving parts of the signer who is under direct sun. Sessions `S1` and `S2` have almost the same profile, but the former has higher number of events, which is maintained for almost all the signs. We found that the open angular lens captures more details of the motions, and as a consequence, it triggers a higher number of events.

The signs having lower event rates correspond to those where one or both hands perform small movements. For example, the **squirrel** sign represented with two fingers movement, the index and middle. This leads to a very small motion, which is captured by a lesser number of triggered

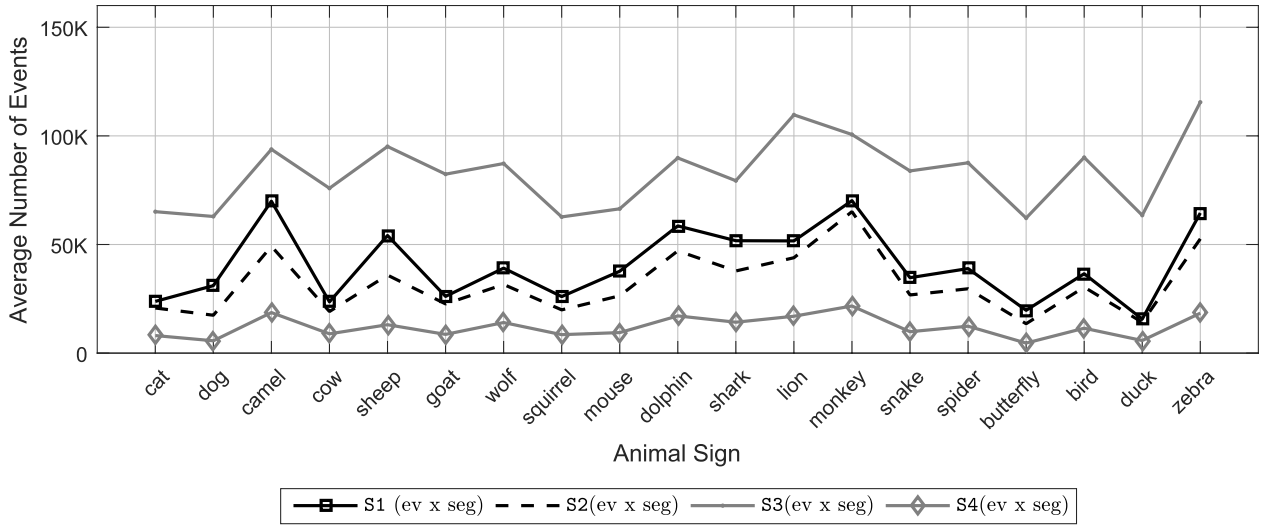


Fig. 8 The figure depicts the average number of events in a temporal window of one second for each animal sign at the four recording sessions

events. Another example is the **cow** sign, where one hand with the index and small fingers outwards are placed on the head, and a quick turn of the wrist is made to emphasise the action. The effect of artificial lights is evident since for S3, the **cow** which has minimal movement is one of the signs with a higher average number of events.

Signs involving wide and fast movements including both arms trigger a very high number of events. This is the case of **camel**, **wolf**, **shark**, **spider**, **monkey** and **bird**. In case of the **dolphin** sign only one arm performs the wide movements.

Finally, Figs. 7 and 8 show information which can be employed to choose one of the two considered approaches to design the sign recognition system. One possible approach is using a fixed delta time τ window to build training samples from event sequences of the dataset. We call this approach temporal windowing. The other approach is using as training samples the complete sequences taken from the start of the sign until the end of the motion. We call this the sign-based approach.

3.5 Spatial analysis of signs

To analyze the complexity on the **SL-Animals-DVS** dataset, we made a comparison with the 10 actions of the DVS-GESTURE dataset [5]. In Figs. 9 and 10, the histogram of events for the duration of each action is plotted. It shows the areas where events for every action are concentrated. The figures are plotted for actions performed by one user and the subfigure correspond to the different actions in the two databases. We can observe few actions in the DVS-GESTURE dataset which are spatially biased i.e., they are performed at a chosen location of an image. Specifically, for example **right hand wave** and **left hand wave**, the hands are localized on the left or right of the frame and there is no other action occurring in the same spatial area. For a classification algorithm, the recognition of this actions is easily solved. Some actions of the DVS-GESTURE dataset have a complexity related to the direction of motion of the same action in the same location. This is the case for the gesture corresponding to clockwise/counter-clockwise movements.

In our dataset, the actions cannot be classified based on only the location of events as it has a wider spread of sign types happening at the same location and most of them are not concentrated in specific areas.

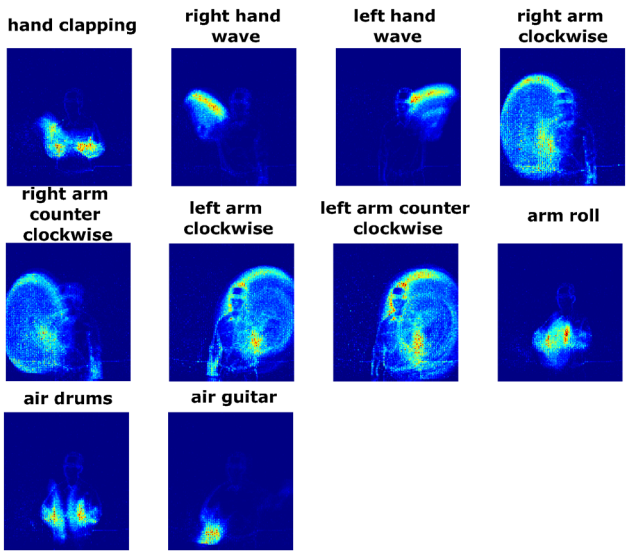


Fig. 9 Heat map of events for actions in the DVS-GESTURE dataset

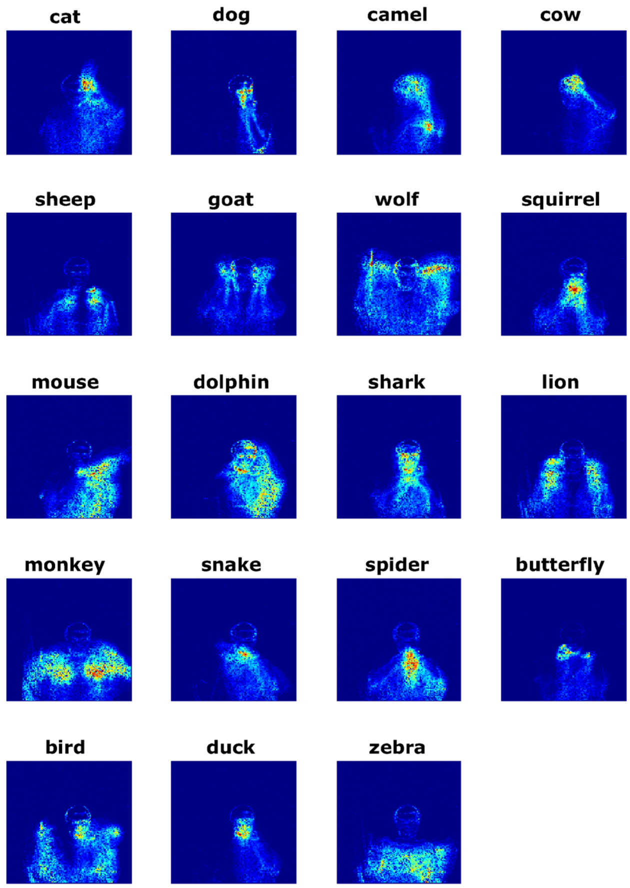


Fig. 10 Heat map of events in the SL-Animals-DVS dataset

4 Automatic classification methods

We employ three classification methods implementing neuromorphic SNN architectures to benchmark the results when applied on the new **SL-Animals-DVS**, namely the SLAYER [36], Spatio Temporal Back Propagation (STBP) [47] and Deep Continuous Local Learning (DECOLLE) [21].

4.1 Spiking neuron model

Neurons in SNN model communicate with each other through spikes, voltage pulse signals through their synapses that accumulate into their internal states. Formally, the instantaneous membrane potential $u_j(t)$ of cell j in the Spike Response Model (SRM) [20] is expressed as double sum as:

$$u_j(t) = \sum_{i=1}^{N_I} \sum_{f=1}^{N_i} w_i \varepsilon(t - t_i^f - d_i) + \eta(t - t_o^l) \tag{1}$$

where we assume that neuron j has N_I input synapses and the ith synapse transmits N_i spikes. Each spike is fired at time t_i^f and their contribution to the cell potential is governed by the spike response function ε, w_i is the synaptic weight and d_i the synaptic delay for the ith synapse. In the short-term memory SRM, only the last fired spike t_o^l contributes to the refractoriness function $\eta(t - t_o^l)$ (please refer to [45] for more details about this term).

Let be $\Delta_i^f = t - t_i^f - d_i$. The spike response function $\varepsilon(\Delta_i^f)$ which describes the effect of the pre-synaptic spike on the internal state of the postsynaptic neuron is expressed as:

$$\varepsilon(\Delta_i^f) = \begin{cases} \frac{\Delta_i^f}{\tau} exp(1 - \frac{\Delta_i^f}{\tau}) & \text{when } \Delta_i^f > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where τ indicates the decay time of postsynaptic potentials, which determines the shape of the spike response function. Equation 2 clearly shows that the real firing time of the pre-synaptic neuron i will be $t_i^f - d_i$. When the internal state variable $u_j(t)$ crosses the firing threshold θ, the neuron j fires an output spike. Then, the membrane potential of the neighboring connected neurons are updated with response kernels scaled by the connecting synaptic weights, following eq. 1. In this work we will consider a zero delay for each neuron: $d_i = 0$

4.2 SLAYER

SLAYER trains a SNN with Backpropagation Through Time where the error is distributed through the network layers and can simultaneously learn synaptic weights and axonal delays [36]. In the learning phase, SLAYER defines for each training sample an output target spike rate with greater number of spikes for the known correct class and lesser number of spikes for the wrong classes. Then, the error to be propagated is calculated by means of a spike flow distance measure. To calculate the distance, the spike flows are smoothed by a kernel and the distance or error is the integrated difference along time between the target smoothed spike flow and the output smoothed spike flow.

In this work, we use as training samples the complete sequences taken from the start of the sign until the end of the motion. Hence, it is a gesture based approach.

In SLAYER we do not need to define a desired spike flow i.e., the spike expectation at each time step. We only need to define the total number of desired spikes in an time interval T_{int} since the error is the difference of the integrated spike flows in T_{int}. For an output spike flow S and a desired spike flow $\hat{S}$, the error e is calculated as

$$\mathrm{e}^{n_l}(t) = \left(\int_{T_{\text{int}}} S^{(n_l)}(\tau) \mathrm{d}\tau - \int_{T_{\text{int}}} \hat{S}(\tau) \mathrm{d}\tau \right), t \in T_{\text{int}} \tag{3}$$

Therefore, it is sufficient to set the number of spikes for the second term and not the specific pattern of spike flow itself because the second term is an integration of the spike flow. The loss is zero outside the interval T_{int}.

The derivative of the spike function is the key part of any SNN with supervised learning with back propagation since the spiking activity is itself not differentiable. SLAYER deals with this problem by considering the probability of spiking state change in presence of a perturbation as an estimate for the derivative of the spike function. For backpropagation with SLAYER any optimization technique like simple gradient descent or Adam [22] can be used.

4.3 STBP

Spatio Temporal Backpropagation (STBP) is an algorithm to train high performance SNNs where neuronal dynamics are defined by both the spatial and temporal dimensions [47]. STBP works using a temporal windowing based approach. The event-based samples in the dataset are converted to a series of binary images by activating positions in the retina with a high number of coincident events inside a time interval. The time interval between each frame is kept constant irrespective of the number of triggered events.

STBP works with a regular SNN where each neuron connects to itself in time. Therefore, the neuron state i.e., its internal potential is also affected by its temporal memory i.e., its value at the previous time step. The internal states of the cells follow the SRM model.

The error for Backpropagation is calculated as the mean square error of the difference between the label vector and the averaged output spikes of the final layer of the SNN within a time window. To address the non-differentiability of spiking activity in STBP, the derivate of spiking activity is approximated by model functions,

$$h(u_j) = \frac{1}{a}\text{sign}\left(|u_j - \theta| < \frac{a}{2}\right), \tag{4}$$

where u_j is the membrane potential of cell j, θ is the threshold and a controls the curve steepness.

4.4 DECOLLE

Deep Continuous Local Learning (DECOLLE) is a SNN model for online learning using plasticity dynamics [21]. It deals with specific issues of SNN models, such as the non-differentiable spiking neurons, the continuous-time dynamics which raises a temporal credit assignment problem, and the constraints of local information which cannot backpropagate the errors at the top layer.

The neuron and synapse models used in DECOLLE follow SRM dynamics with a relative refractory mechanism. In the training phase, each layer of the SNN feeds local readout units through fixed random connections, producing the auxiliary targets $\hat{Y}$. The random readout is obtained by multiplying the neural activations A_j^l with a random and fixed matrix G_{kj}^l on each layer l: $Y_k^l = \sum_j G_{kj}^l A_j^l$. The global loss function is then defined as the sum of the layerwise loss functions defined on the random readouts, i.e., $\mathcal{L} = \sum_{l=1}^{N} L^l(Y^l)$. To ensure reasonable firing rates and prevent sustained firing, this equation is modified with two regularizers.

To perform weight updates, DECOLLE uses surrogate gradients allowing to propagate the gradient forward and making the plasticity rule temporally local. Thus, each layer indirectly learns useful hierarchical features that will later minimize the cost at the top layer i.e., DECOLLE performs backpropagation on every layer and within the same time step. Errors are propagated through the random connections to train weights coming into the spiking layer. This is how a layer is capable of learning deep spatio-temporal representations from spikes relying solely on local information.

When applied the Mean Square Error (MSE) loss for layer l the DECOLLE rule for updating synaptic weights becomes:

$$\Delta w_{ij}^l = -\eta\, \text{error}_i^l \sigma'(U_i^l) P_j^l, \tag{5}$$

$$\text{error}_i^l = \sum_k G_{ki}^l (Y_k^l - \hat{Y}_k^l) \tag{6}$$

In this case, we have one modulatory factor (error_i), one post-synaptic factor ($\sigma'(U_i^l)$) that consists of a surrogate gradient of the non-differentiable step function applied to the membrane potential, and one pre-synaptic factor (P_j^l) that describes the traces of the membrane and drives the weight update rule.

5 Methods and results

5.1 Training methodology

We perform a K-fold cross-validation procedure to estimate training and test results of the different SNN networks on the SL-Animals-DVS dataset. It consists of dividing the complete set into K parts following a leave-signers-out approach, where the actions performed by the same user are never at the training set and the testing set at the same time.

Using $K = 4$, the training set has the signs of around 42 signers, and the testing set the signs of the remaining 14 signers. Results on each testing set are accumulated on the same confusion matrix.

To evaluate the robustness of the classifiers facing event noises, we also perform training and testing of each SNN classifier on the SL-Animals-DVS dataset but removing the noisy set `S3`.

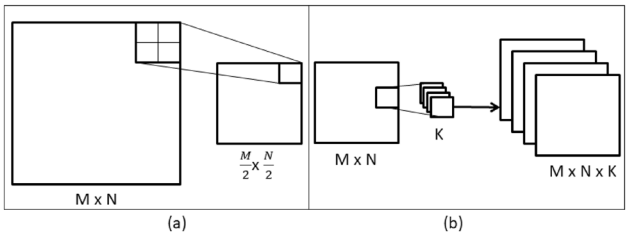


Fig. 11 **a** Shows a pooling layer and **b** shows a convolutional layer with K kernels producing K output maps

5.2 Classifiers results

The results are summarized in Table 2, showing the accuracy for each method with a K-fold cross validation approach for the DVS-GESTURE and SL-Animals-DVS datasets.

In SLAYER, the network type chosen is comparable to the network used in evaluating DVS-GESTURE dataset and it is empirically balanced in the tradeoff of training time versus accuracy. The error is calculated as the difference between the accumulated number of output spikes and the desired number of spikes. The desired number of spikes was set to 180 for the correct class and 30 for the wrong classes. Only the first 1.5 s of the action recording were used with a time step of 5 mS. For a network [128 × 128×2 - 2p - 8c5 - 2p - 16c5 - 2p - 100fc - 19o] where the third dimension of the input shape has two channels that correspond to the two polarity values, *p* denotes a pooling layer, *c* denotes a convolutional layer and *fc* denotes a fully connected layer with each network type prefixed and suffixed by the network size and kernel size, respectively, the network achieved a testing accuracy of 78.03% ± 3.08%.

The network shape is chosen to reduce the memory requirement for processing therefore reducing the learning time. The use of only 1.5 seconds of every recording is justified since every sign is performed within this time and we can avoid overlap of signs within this time frame. In Fig. 11 we show the different type of layers used and their resultant sizes dependent on the number of maps in the layer. In Fig. 12 we show a typical architecture used for the SLAYER method.

Using recordings from all of the `S1`, `S2`, `S3` and `S4` sets and performing a K-fold cross validation we get worse accuracy than using only the `S1`, `S2` and `S4` sets. We obtain accuracies [54.54, 62.18, 65.45, 61.45] for the four trials with a mean accuracy of 60.9% and a standard deviation of 4.58%.

For STBP, the recordings were converted to a series of frames. The first 1.5 seconds of each recording was reduced to 50 frames by accumulating the spikes in an equal time period and thresholding. We used a batch size of 19 in the training phase. For a network [128 × 128×2 - 20c3 - 2p - 50c3 - 2p - 200fc - 19o], the testing accuracy was 71.45% ± 1.74% with a K-fold cross validation split on the `S1`, `S2` and `S4` sets with K=4. The accuracy when using `S1`, `S2`, `S3` and `S4` sets and performing a k-fold cross validation drops to 56.20% ± 1.52%.

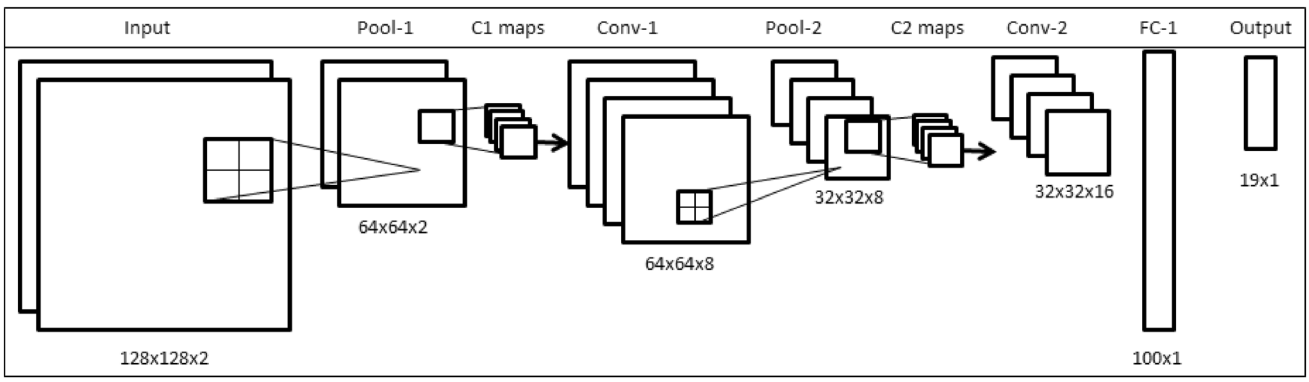
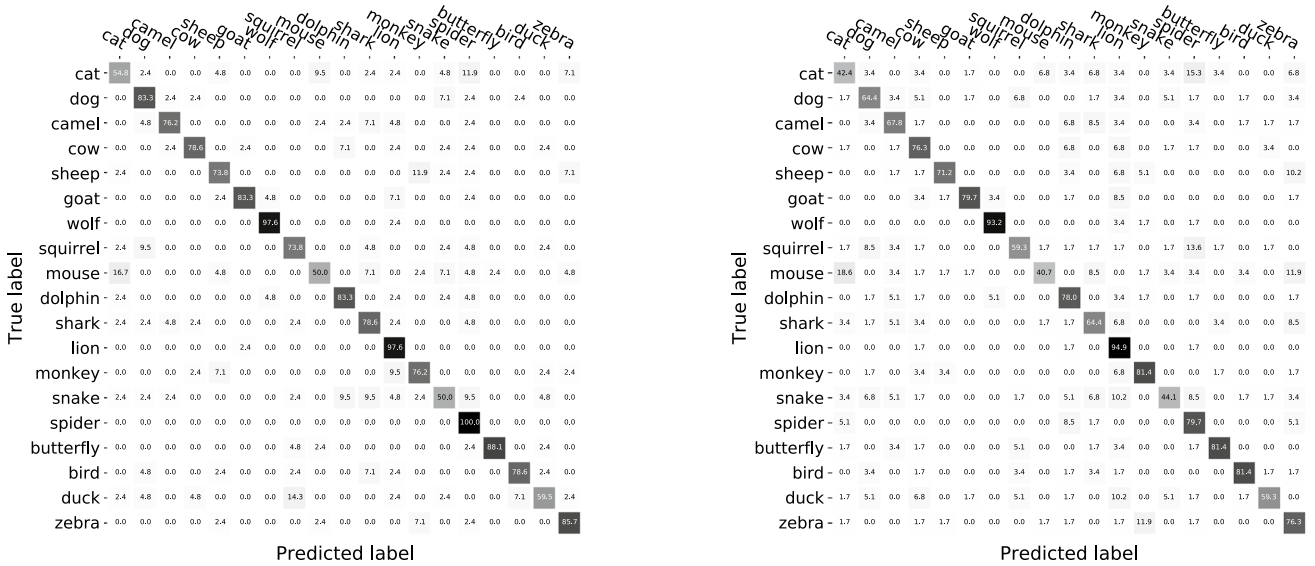


Fig. 12 The full architecture of a Spiking Neural Network, i.e., SLAYER

Table 2 Accuracy of the three datasets evaluated using SLAYER, STBP and DECOLLE classifiers

Method	Set		
	DVS-GESTURES	SL-DVS S1,S2 & S4	SL-DVS S1,S2,S3 & S4
SLAYER	93.64%	78.03% ± 3.08%	60.09% ± 4.58%
STBP	95.9% ± 5.4%	71.45% ± 1.74%	56.20% ± 1.52%
DECOLLE	95.1% ± 3.2%	77.6% ± 6.5%	70.6% ± 7.8%



(a) SL-Animals-DVS S1, S2 & S4

(b) SL-Animals-DVS S1, S2, S3 & S4

Fig. 13 Confusion Matrix showing percentages for the two SL-Animals-DVS sets with DECOLLE classifier

The DECOLLE network implements *Lenet* architecture from the authors github [1]: [32×32×2 - 64c7 - 128c7 - 128c7 - 2p - 25fc - 19o]. During the training, slices of 500 ms randomly selected from each sign sample with a resolution of 1 ms populate the batch that feed the network. Testing sequences have 1,800 ms long with a burn-in period of 100 ms. The performance on IBM GESTURES-DVS shows an accuracy below what the authors reported on [21] as we can see on Table 2, because we implement the the K-Fold train and validation, instead of a fixed split between train and tests samples. For the SL-Animals-DVS dataset, results are 70.6% ± 7.8% and 77.6% ± 6.5% on the sets with and without the indoor set `S3`, respectively. The former is the highest score on recognition using the noisy set.

The accuracy with the STBP method being lower than the accuracy obtained with SLAYER and DECOLLE indicates that the framing of the DVS events data causes a loss in information and that a method preserving the spatio-temporal dynamics of the data is more appropriate for recognition of dynamic scenes as it is the case of the sign recognition task.

5.3 Confusion matrix analysis

Figure 13 presents the confusion matrix using DECOLLE classifier on both SL-Animals-DVS datasets. Figure 13a shows the results using `S1`, `S2` & `S4` sets, and Fig. 13b corresponds to the results incorporating the indoor set `S3`. From both confusion matrices, we can observe a good recognition rate on all symmetric two hands signs: **sheep**, **goat**, **wolf**, **lion**, **monkey**, **butterfly** and **zebra**. Conversely, both-handed signs using one or two hands report lower performance, showing the high complexity that the classifiers face. Lowest recognition rates are reported on signs with low amplitude motion and performed in front of the signer. This is even more evident when the training dataset includes the noisy set, because the texture of the clothes seems to hide the events belonging to the sign. In this case, the performance could be improved using pre-processing attention algorithms tuned to detect fastest moving parts of the scene.

The case of the **cat** sign shows how the compound signs can be misclassified on some cases. This sign has the same hands position through all the movement, but it is very similar to the starting point of **mouse** and **spider** signs. During the testing phase, a **cat** sign will trigger spikes for the output neurons of the three signs. On the other hand, **mouse** and **spider** signs motion differs from **cat** after the initial position. Thus, their dynamic will not trigger the **cat** output neuron.

Adding indoor set with noisy body texture increases the confusion of some of these signs related with the body, such as sheep and zebra (see Fig. 13b). Signs related to the head position have also better recognition rate, even if they are performed by one hand, such as **cow** sign, showing that their events are not hidden by the texture of the clothes.

5.4 Power consumption on TrueNorth hardware

The total number of spikes processed by a network to perform each recognition depends on the particular input sequence and the network parameters. For the trained SLAYER 7-layer network of shape [128×128×2 - 2p - 8c5 - 2p - 16c5 - 2p - 100fc - 19o] the estimated average number

of spikes generated by each layer during each sequence recognition is: Layer 1-6.5k spikes, Layer 2-66.6K spikes, layer 3-319.7K spikes, layer 4-108K spikes, layer 5-700K spikes, layer 6-200K spikes and layer 7-50K spikes. Thus, the network processes an average of 1.45M spikes during a forward pass. The average number of output synapses per neuron is 257.6 synapses/neuron. So, to recognize a complete sequence, the network has to compute an average of 373M synaptic connections.

The TrueNorth neuromorphic hardware consumes an average of 26pJ per synaptic event [31]. Consequently, we can estimate a value of 9.7 mJ Energy as the energy that would consume a forward pass of a complete recognition sequence running on TrueNorth. This compares favourably against GPUs which need an average power of 100–200 W to function and average energy of 0.5 J/image for a non spiking Convolutional Neural Network [23]. For a comparable video processing at 30 frames per second the average consumption of recognizing a sequence on a GPU would be 63J, four orders of magnitude higher than on TrueNorth.

5.5 Limitations of the dataset

The number of signers in the dataset is higher than in previous datasets of similar complexity but there are shortcomings that we feel should be addressed before it is suitable for a real world application.

The signers recorded are not fluent or native signers. Also, the dataset only contains isolated signs, not phrases or sentences with grammar. We record with a single 2D camera, so we lack the depth information which could be useful in recognition of signs.

The DVS employed has a small retina and it can lose details about hand/fingers configurations. The focus of the camera is such that movement of the lips is also not focused and recorded. Despite these shortcomings this dataset is very helpful to capture the motion, orientation, and position, etc. of most of the signs, and is a challenging dataset for benchmarking gesture recognition on spiking neural networks.

6 Conclusion

An event-based action dataset referred as **SL-Animals-DVS**, based on sign language gestures, is introduced and analyzed in terms of complexity, dynamics and inter-class variability. This dataset will be a good benchmark to train and evaluate event-based Automatic Sign Language Recognition systems, as it includes a greater number of unique signers as well as greater complexity of action and the noise level across recordings. Furthermore, the different dynamic and motion of several gestures, which can be played by both hands by the signers increases the intra-class variability. A direct comparison with the DVS-GESTURES dataset shows that **SL-Animals-DVS** is better suited to benchmark real implementation tasks.

In this work, three state-of-the-art SNNs training algorithms are applied on **SL-Animals-DVS** for the gesture recognition task. Two of the methods, SLAYER and DECOLLE, are gesture based training approaches, while the other, STBP, is a time-windowing based approach. The first methods obtain better results which suggests that SNN methods avoiding framing and thus preserving the spatio-temporal dynamics of the data are more appropriate for gesture recognition tasks.

Further research on methodologies encoding combined dynamics of the gestures should improve the recognition task of the complex co-occurrence of visuospatial patterns in sign language. Moreover, future efforts should be considered to increase the dataset in two main axes. Firstly, a consistent dataset should incorporate records of fluent signers. In that case, the vocabulary set could be increased, adding signs related to the geographic location where it is recorded. Secondly, adding simple phrases to the set of classes will give the possibility of applying a continuous interpretation in the recognition of the signs.

Acknowledgements This work was funded by EU H2020 grants 824164 “HERMES”, 871371 “Memscales” and 871501 “NeuroNN”, Spanish grant from the Ministry of Economy and Competitivity NANOMIND-PID2019-105556GB-C31, with support from the European Regional Development Fund. P. Negri was partially supported by the Scholarship Program Mobility of the Postgraduate Iberoamerican University Association (AUIP). C. Di Ielsi was supported by a scholarship of the Computer Department of the University of Buenos Aires. A. Vasudevan was supported by the MINECO FPI scholarship BES-2016-077757.

References

1. DECOLLE implementetion code. https://github.com/nmi-lab/decolle-public. Accessed 09 July 2021
2. jaer open source project: Real time sensory-motor processing for event-based sensors and systems. http://www.jaerproject.org. Accessed: 09 July 2021
3. SL-Animals-DVS dataset. http://www2.imse-cnm.csic.es/neuromorphs/index.php/SL-ANIMALS-DVS-Database. Accessed: 09 July 2021
4. Amaral L, Júnior GL, Vieira T, Vieira T (2018) Evaluating deep models for dynamic Brazilian sign language recognition. In: Iberoamerican congress on pattern recognition, pp. 930–937. Springer. https://doi.org/10.1007/978-3-030-13469-3_107
5. Amir A, Taba B, Berg D, Melano T, McKinstry J, Di Nolfo C, Nayak T, Andreopoulos A, Garreau G, Mendoza M et al (2017) A low power, fully event-based gesture recognition system. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7243–7252 (2017). https://doi.org/10.1109/CVPR.2017.781
6. Baranwal N, Nandi GC (2017) An efficient gesture based humanoid learning using wavelet descriptor and mfcc techniques. Int

J Mach Learn Cybern 8(4):1369–1388. https://doi.org/10.1007/s13042-016-0512-4
7. Bellugi U, Klima E (2001) Sign language. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social and behavioral sciences, pp. 14066–14071. Pergamon, Oxford. https://doi.org/10.1016/B0-08-043076-7/02940-5
8. Camuñas-Mesa LA, Linares-Barranco B, Serrano-Gotarredona T (2019) Neuromorphic spiking neural networks and their memristor-cmos hardware implementations. Materials 12(17):2745. https://doi.org/10.3390/ma12172745
9. Canales E (2021) iAPRENDE A SIGNAR! LSE (20 Animales). https://www.youtube.com/watch?v=IRue9cRhsDk. Accessed: 9 July
10. Caselli NK, Sehyr ZS, Cohen-Goldberg AM, Emmorey K (2017) ASL-LEX: a lexical database of American sign language. Behav Res Methods 49(2):784–801. https://doi.org/10.3758/s13428-016-0742-0
11. Cerna LR, Cardenas EE, Miranda DG, Menotti D, Camara-Chavez G (2021) A multimodal libras-ufop Brazilian sign language dataset of minimal pairs using a microsoft kinect sensor. Exp Syst Appl 167:114179. https://doi.org/10.1016/j.eswa.2020.114179
12. Chen G, Chen J, Lienen M, Conradt J, Röhrbein F, Knoll AC (2019) FLGR: fixed length GISTS representation learning for RNN-HMM hybrid-based neuromorphic continuous gesture recognition. Front Neurosci 13:73
13. Cheok MJ, Omar Z, Jaward MH (2019) A review of hand gesture and sign language recognition techniques. Int J Mach Learn Cybern 10(1):131–153. https://doi.org/10.1007/s13042-017-0705-5
14. Corina D (2001) Sign language: psychological and neural aspects. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social and behavioral sciences, pp. 14071–14075. Pergamon, Oxford. https://doi.org/10.1016/B0-08-043076-7/03492-6
15. Dreuw P, Neidle C, Athitsos V, Sclaroff S, Ney H (2008) Benchmark databases for video-based automatic sign language recognition. In: LREC
16. Emmorey K, Corina D (1990) Lexical recognition in sign language: effects of phonetic structure and morphology. Percept Motor Skills **71**(3_suppl), 1227–1252. https://doi.org/10.2466/pms.1990.71.3f.1227
17. Eryilmaz SB, Joshi S, Neftci E, Wan W, Cauwenberghs G, Wong HSP (2016) Neuromorphic architectures with electronic synapses. In: 17th international symposium on quality electronic design (ISQED), pp. 118–123. https://doi.org/10.1109/ISQED.2016.7479186
18. Escalera S, Baró X, Gonzalez J, Bautista MA, Madadi M, Reyes M, Ponce-López V, Escalante HJ, Shotton J, Guyon I (2014) Chalearn looking at people challenge 2014: dataset and results. In: European conference on computer vision, pp. 459–473. Springer. https://doi.org/10.1007/978-3-319-16178-5_32
19. Forster J, Schmidt C, Koller O, Bellgardt M, Ney H (2014) Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather. In: International conference on language resources and evaluation, pp. 1911–1916
20. Gerstner W, Kistler WM (2002) Spiking neuron models: single neurons, populations, plasticity. Cambridge University Press, Cambridge
21. Kaiser J, Mostafa H, Neftci E (2020) Synaptic plasticity dynamics for deep continuous local learning (decolle). Front Neurosci 14:424. https://doi.org/10.3389/fnins.2020.00424
22. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
23. Li D, Chen X, Becchi M, Zong Z (2016) Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs. In: IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom), pp. 477–484. https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.76
24. Liang ZJ, Liao SB, Hu BZ (2018) 3D convolutional neural networks for dynamic sign language recognition. Comput J 61(11):1724–1736. https://doi.org/10.1093/comjnl/bxy049
25. Lichtsteiner P, Posch C, Delbruck T (2006) A 128*128 120db 15us latency asynchronous temporal contrast vision sensor. pp. 566–576. https://doi.org/10.1109/JSSC.2007.914337
26. Lungu IA, Corradi F, Delbrück T (2017) Live demonstration: convolutional neural network driven by dynamic vision sensor playing RoShamBo. In: IEEE international symposium on circuits and systems (ISCAS), p 1 . https://doi.org/10.1109/ISCAS.2017.8050403
27. Maass W (1997) Networks of spiking neurons: the third generation of neural network models. Neural Netw 10(9):1659–1671. https://doi.org/10.1016/S0893-6080(97)00011-7
28. Maro JM, Ieng SH, Benosman R (2020) Event-based gesture recognition with dynamic background suppression using smartphone computational capabilities. Front Neurosci 14:275
29. Martínez AM, Wilbur RB, Shay R, Kak AC (2002) Purdue RVL-SLLL ASL database for automatic recognition of American sign language. In: IEEE international conference on multimodal interfaces, pp 167–172. https://doi.org/10.1109/ICMI.2002.1166987
30. McLeister M (2019) Worship, technology and identity: a deaf protestant congregation in urban China. Stud World Christ 25(2):220–237
31. Merolla PA, Arthur JV, Alvarez-Icaza R, Cassidy AS, Sawada J, Akopyan F, Jackson BL, Imam N, Guo C, Nakamura Y et al (2014) A million spiking-neuron integrated circuit with a scalable communication network and interface. Science 345(6197):668–673. https://doi.org/10.1126/science.1254642
32. Mori Y, Toyonaga M (2018) Data-glove for Japanese sign language training system with gyro-sensor. In: Joint 10th international conference on soft computing and intelligent systems (SCIS) and 19th international symposium on advanced intelligent systems (ISIS), pp. 1354–1357. https://doi.org/10.1007/s13042-017-0705-5
33. Pérez-Carrasco JA, Zhao B, Serrano C, Acha B, Serrano-Gotarredona T, Chen S, Linares-Barranco B (2013) Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing-application to feedforward ConvNets. IEEE Trans Pattern Anal Mach Intell 35(11):2706–2719. https://doi.org/10.1109/TPAMI.2013.71
34. Posch C, Matolin D, Wohlgenannt R (2010) A QVGA 143 db dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS. IEEE J Solid-State Circuits 46(1):259–275. https://doi.org/10.1109/JSSC.2010.2085952
35. Serrano-Gotarredona T, Linares-Barranco B (2013) A 128x128 1.5% contrast sensitivity 0.9% fpn 3us latency 4 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers. IEEE J Solid State Circuits 48(3):827–838
36. Shrestha SB, Orchard G (2018) SLAYER: spike layer error reassignment in time. In: Advances in neural information processing systems, pp 1412–1421
37. Sivilotti MA (1991) Wiring considerations in analog vlsi systems, with application to field-programmable networks. Ph.D. thesis, Computation and Neural Systems, California Inst. Technol., Pasadena, CA, USA
38. Troelsgård T, Kristoffersen JH (2008) An electronic dictionary of Danish sign language. In: Theoretical issues in sign language research conference, Florianopolis, Brazil
39. Upadhyay NK, Jiang H, Wang Z, Asapu S, Xia Q, Joshua Yang J (2019) Emerging memory devices for neuromorphic computing.

Adv Mater Technol 4(4):1800589. https://doi.org/10.1002/admt.201800589
40. Vasudevan A, Negri P, Linares-Barranco B, Serrano-Gotarredona T (2020) Introduction and analysis of an event-based sign language dataset. In: Faces and gestures in E-health and welfare (FaGEW) workshop, 15th IEEE international conference on automatic face and gesture recognition (FG), pp 441–448
41. Von Agris U, Kraiss KF (2007) Towards a video corpus for signer-independent continuous sign language recognition
42. Wan J, Zhao Y, Zhou S, Guyon I, Escalera S, Li SZ (2016) Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 56–64 https://doi.org/10.1109/CVPRW.2016.100
43. Wang H, Chai X, Hong X, Zhao G, Chen X (2016) Isolated sign language recognition with Grassmann covariance matrices. ACM Trans Access Comput (TACCESS) 8(4):1–21. https://doi.org/10.1145/2897735
44. Wang Q, Zhang Y, Yuan J, Lu Y (2019) Space-time event clouds for gesture recognition: from rgb cameras to event cameras. In: IEEE winter conference on applications of computer vision (WACV), pp 1826–1835. https://doi.org/10.1109/WACV.2019.00199
45. Wang X, Lin X, Dang X (2019) A delay learning algorithm based on spike train kernels for spiking neurons. Front Neurosci 13:252. https://doi.org/10.3389/fnins.2019.00252
46. World Health Organization: Deafness and hearing loss (2019). https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed 09 July 2021
47. Wu Y, Deng L, Li G, Zhu J, Shi L (2018) Spatio-temporal backpropagation for training high-performance spiking neural networks. Front Neurosci 12:331. https://doi.org/10.3389/fnins.2018.00331
48. Yousefzadeh A, Khoei MA, Hosseini S, Holanda P, Leroux S, Moreira O, Tapson J, Dhoedt B, Simoens P, Serrano-Gotarredona T et al (2019) Asynchronous spiking neurons, the natural key to exploit temporal sparsity. IEEE J Emerg Sel Top Circuits Syst 9(4):668–678. https://doi.org/10.1109/JETCAS.2019.2951121
49. Yuan T, Sah S, Ananthanarayana T, Zhang C, Bhat A, Gandhi S, Ptucha R (2019) Large scale sign language interpretation. In: 14th IEEE international conference on automatic face and gesture recognition (FG), pp 1–5. https://doi.org/10.1109/FG.2019.8756506