Hand-gesture recognition based on EMG and event-based camera sensor fusion: a benchmark in neuromorphic computing

Enea Ceolini [1], Charlotte Frenkel [1,2], Sumit Bam Shrestha [3], Gemma Taverni [1], Lyes Khacef [4], Melika Payvand [1], Elisa Donati [1]

[1] *Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland.*
[2] *ICTEAM Institute, Université catholique de Louvain, Belgium.*
[3] *Temasek Laboratories @ National University of Singapore, Singapore.*
[4] *Université Côte d'Azur, CNRS, LEAT, France.*

Correspondence*:
Corresponding Author
elisa@ini.uzh.ch

ABSTRACT

Hand gestures are a form of non-verbal communication used by individuals in conjunction with speech to communicate. Nowadays, with the increasing use of technology, hand-gesture recognition is considered to be an important aspect of Human-Machine Interaction (HMI), allowing the machine to capture and interpret the user's intent and respond accordingly. The ability to discriminate human gestures can help in several applications such as assisted living, healthcare, neuro-rehabilitation, and sports. Recently, multi-sensor data fusion mechanisms have been investigated to improve discrimination accuracy. In this paper, we present a sensor fusion framework that integrates complementary systems: the electromyography (EMG) signal from muscles and visual information. This multi-sensor approach, while improving accuracy and robustness, introduces the disadvantage of high computational cost, which grows exponentially with the number of sensors and the number of measurements. Furthermore, this huge amount of data to process can affect the classification latency which can be crucial in real-case scenarios such as prosthetic control. Neuromorphic technologies can be deployed to overcome these limitations since they allow real-time processing in parallel at low power consumption. In this paper, we present a fully neuromorphic sensor fusion approach for hand-gesture recognition comprised of event-based vision sensor and three different neuromorphic processors. In particular, we used the event-based camera, called DVS, and two neuromorphic platforms, Loihi and ODIN+MorphIC. The EMG signals were recorded using traditional electrodes and then converted into spikes to

be fed into the chips. We collected a dataset of 5 gestures from sign language where visual and electromyography signals are synchronized. We compared a fully neuromorphic approach to a baseline implemented using traditional machine learning approaches on a portable GPU system. According to the chips constraints, we designed specific spiking neural networks (SNNs) for sensor fusion that showed classification accuracy comparable to the software baseline. These neuromorphic alternatives have increased inference time, between 20% and 40%, with respect to the GPU system but have a significantly smaller energy-delay product (EDP) which makes them between 30x and 600x more efficient. The proposed work represents a new benchmark that moves the neuromorphic computing towards a real-world scenario.

Keywords: Hand-gesture classification, Spiking Neural Networks (SNNs), Electromyography (EMG) Signal Processing, Event-based camera, Sensor Fusion, Neuromorphic Engineering

1 INTRODUCTION

Hand-gestures are considered a powerful communication channel for information transfer in daily life. Hand-gesture recognition is the process of classifying meaningful gestures of the hands, and is receiving renewed interest. The gestural interaction is a well-known technique that can be utilized in a vast array of applications (Yasen and Jusoh, 2019), such as sign language translation (Cheok et al., 2019), sports (Loss et al., 2012), Human-Robot Interaction (HRI) (Liu and Wang, 2018; Cicirelli et al., 2015), and more generally in Human-Machine Interaction (HMI) (Haria et al., 2017). Hand-gesture recognition systems also target medical applications, where they are detected via bioelectrical signals instead of vision. In particular, among the biomedical signals, electromyography (EMG) is the most used for hand-gesture identification and for the design of prosthetic hand controllers (Donati et al., 2019; Chen et al., 2020; Benatti et al., 2015).

EMG measures the electrical signal resulting from muscle activation. The source of the signal is the motor neuron action potentials generated during the muscle contraction. Generally, EMG can be detected either directly with electrodes inserted in the muscle tissue, or indirectly with surface electrodes positioned above the skin (surface EMG (sEMG), for simplicity we will refer to it as EMG). The EMG is more popular for its accessibility and non-invasive nature. However, the use of EMG to discriminate hand-gestures is a non-trivial task due to several physiological processes in the skeletal muscles underlying their generation.

One way to overcome these limitations is to use a multimodal approach, combining EMG with recordings from other sensors. Multi-sensor data fusion is a direct consequence of the well-accepted paradigm that certain natural processes and phenomena are expressed under completely different physical guises (Lahat et al., 2015). In fact, multi-sensor systems provide higher accuracy by exploiting different sensors that measure the same signal in different but complementary ways. The higher accuracy is achieved thanks to a redundancy gain that reduces the amount of uncertainty in the resulting information. Recent works show a growing interest toward multi-sensory fusion in several application areas such as developmental robotics (Droniou et al., 2015; Zahra and Navarro-Alarcon, 2019), audio-visual signal processing (Shivappa et al., 2010; Rivet et al., 2014), spatial perception (Pitti et al., 2012), attention-driven selection (Braun et al., 2019) and tracking (Zhao and Zeng, 2019), memory encoding (Tan et al., 2019), emotion recognition (Zhang et al., 2019), multi-sensory classification (Cholet et al., 2019), HMI (Turk, 2014), remote sensing and earth observation (Debes et al., 2014), medical diagnosis (Hoeks et al., 2011), and understanding brain functionality (Horwitz and Poeppel, 2002).

In this study we consider the complementary system comprising of a vision sensor and EMG measurements. Using EMG or camera systems separately presents some limitations, but their fusion has several advantages, in particular EMG-based classification can help in case of camera occlusion, whereas the vision classification provides an absolute measurement of hand state. This type of sensor fusion which combines vision and proprioceptive information is intensively used in biomedical applications, such as in transradial prosthetic domain to improve control performance (Markovic et al., 2014, 2015), or to focus on recognizing objects during grasping to adjust the movements (Došen et al., 2010). This last task can also use Convolutional Neural Networks (CNNs) as feature extractors (Ghazaei et al., 2017; Gigli et al., 2018).

While improving accuracy and robustness, the multiple input modalities also increase the computational cost, due to the amount of data to process in real-time that can affect the communication between the subject and the prosthetic hand. Neuromorphic technology offers a solution to overcome these limitations giving the possibility to process the multiple inputs in parallel in real-time, and with very low power consumption. Neuromorphic systems consist of circuits designed with principles based on the biological nervous systems that, similar to their biological counterparts, process information using energy-efficient, asynchronous, event-driven methods (Liu et al., 2014). These systems are often endowed with on-line learning abilities that allow adapting to different inputs and conditions. Lots of neuromorphic computing platforms have been developed in the past for modeling cortical circuits and their number is still growing (Merolla et al., 2014; Benjamin et al., 2014; Furber et al., 2014; Meier, 2015; Qiao et al., 2015; Moradi et al., 2017; Neckar et al., 2018; Davies et al., 2018; Frenkel et al., 2019a,b; Thakur et al., 2018).

In this paper we present a fully-neuromorphic implementation of sensor fusion for hand-gesture recognition. The proposed work is based on a previous work of sensor fusion for hand-gesture recognition, using standard machine learning approaches implemented in a mobile phone application for personalized medicine (Ceolini et al., 2019a). The paper showed how a CNN performed better, in terms of accuracy, than a Support Vector Machine (SVM) on the hand-gesture recognition task. The novelty introduced here is that the sensor fusion is implemented on a fully neuromorphic system, from event-based camera sensor to the classification phase, performed by using three event-based neuromorphic circuits: Intel's Loihi research processor (Davies et al., 2018) and a combination of the ODIN and MorphIC Spiking Neural Network (SNN) processors (Frenkel et al., 2019a,b). The two neuromorphic systems present different features, in particular, depending on the number of neurons available and on the input data, we implemented different SNN architectures. For example, for visual data processing, a spiking CNN is implemented in Loihi while a spiking Multi-Layer Perceptron (MLP) is chosen for ODIN + MorphIC, (see Section 2.3). For the case of EMG, the data was collected using the Myo armband that senses electrical activity in the forearm muscles. The data was later converted into spikes to be fed into the neuromorphic systems. Here, we propose a feasible application to show the neuromorphic performance in terms of accuracy, energy consumption and latency (stimulus duration + inference time). The performance metric for the energy consumption is the Energy-Delay Product (EDP), a metric suitable for most modern processor platforms defined as the average energy consumption multiplied by the average inference time. The inference time is defined as the time elapsed between the end of the stimulus and the classification. To validate the neuromorphic results, we are comparing to a baseline consisting of the network implemented using a standard machine learning approach where the inputs are fed as continuous EMG signals and video frames. We propose this comparison for a real case scenario as a benchmark, in order for the neuromorphic research field to advance into the mainstream of computing (Davies, 2019).

2 MATERIAL AND METHODS

In the following, we describe the overall system components. We start from the description of the sensors used to collect the hand-gesture data, namely the event-based camera, Dynamic Vision Sensor (DVS), and the EMG armband sensor, Myo. We then describe the procedure with which we collected the dataset used for the validation experiments presented here and that is publicly available. Afterwards, the two neuromorphic systems under consideration, namely Loihi and ODIN + MorphIC, will be described focusing on their system specifics, characteristics and the model architectures that will be implemented on them. Finally, we describe the system that we call baseline and that represents the point of comparison between a traditional von-Neumann approach and the two neuromorphic systems.

2.1 DVS and EMG Sensors

2.1.1 DVS Sensor

The DVS (Lichtsteiner et al., 2006) is a neuromophic camera inspired by the visual processing in the biological retina. Each pixel in the sensor array responds asynchronously to logarithmic changes in light. Whenever the incoming illumination increases or decreases above a certain threshold, it generates a polarity spike event. The polarity corresponds to the sign of the change, ON polarity for increasing in light and OFF polarity for decreasing in light. The output is a continuous and sparse train of events interchangeably called spikes throughout this paper, carrying the information of the active pixels in the scene (represented in Figure 2). The static information is directly removed on the hardware side and only the dynamic one, corresponding to the movements in the scene, is actually transmitted. In this way the DVS can reach low latency, down to 10 μs, reducing the power consumption needed for computation and the amount of transmitted data. Each spike is encoded using the Address Event Representation (AER) communication protocol (Deiss et al., 1999) and is represented by the address of the pixel (in x-y coordinates), the polarity (1 bit for the sign), and the timestamp (in microsecond resolution).

2.1.2 EMG Sensor

In the proposed work, we collected the EMG corresponding to the hand gestures by using the Myo armband by Thalmic Labs Inc. The Myo armband is a wearable device provided with eight equally spaced non-invasive EMG electrodes and a Bluetooth transmission module. The EMG electrodes detect the signals from the forearm muscles activity and afterwards the acquired data is sent to an external electronic device. The sampling rates for Myo data are fixed at $200\mathrm{Hz}$ and the data is returned as a unitless 8-bit unsigned integer for each sensor representing ‘activation’ and does not translate to millivolts (mV).

2.2 DVS-EMG Dataset

The dataset is a collection of 5 hand gestures recorded with the two sensor modalities: muscle activity from the Myo and visual input, in form of DVS events. Moreover, the dataset also provides the video recording using a traditional frame-based camera, referred to as Active Pixel Sensor (APS) in the paper. The frames from the APS are used as ground truth and as input in the baseline models. The APS-frames provided in the dataset are gray-scale, 240x180 resolution. The dataset contains recordings from 21 subjects: 12 males and 9 females of age from 25 to 35, (see Data Availability Statement for the full access to the dataset). The structure is the following: each subject repeats 3 sessions, in each session the subject performs 5 hand gestures: *pinky*, *elle*, *yo*, *index* and *thumb* (see Figure 1), repeated 5 times. Each single gesture recording lasts $2\mathrm{s}$. The gestures are separated by a relaxing time of $1\mathrm{s}$, in order to remove any residual activity from the previous gesture. Every recording is cut in 10 chunks of $200\mathrm{ms}$ each, this duration was selected to match the requirements of a real-case scenario of low latency prosthesis control where there is a need for the classification and creation of the motor command within 250 ms (Smith et al., 2011). Therefore, the final number of samples results to be 21(subjects) x 3(trials) x 5(repetitions) x 5(gestures) x 10(chunks) for a total of 15750. The Myo records the superficial muscle activity at the middle forearm from 8 electrodes with a sampling rate of $200\mathrm{Hz}$. During the recordings the DVS was mounted on a random moving system to generate relative movement between the sensor and the subject hand. The hand stands static during the recording to avoid noise in the Myo sensor and the gestures are performed in front of a static white background, see Figure 1 for the full setup.

2.2.1 Implementation on neuromorphic devices

SNNs in general and their implementation on neuromorphic devices require inputs as spike trains. In the case of the DVS, the sensor output is already in form of spikes and polarity. The only requirement that we need to take into account is the limited number of neurons in the available neuromorphic processors. For this reason, we decided to crop the 128×128 input of the DVS to 40×40 centered on the hand-gesture. On the contrary, for the EMG, a conversion in the event-based domain is required. The solution used here is the delta-modulator ADC algorithm, based on a sigma-delta modulator circuit (Corradi and Indiveri, 2015). This mechanism is particularly used in low frequency, high performance and low power application (Lee et al., 2005) such as biomedical circuits. Moreover, this modulator represents a good interface for neuromorphic devices because it has much less circuit complexity and lower power consumption than multi-bit ADCs.

The delta-modulator algorithm transforms a continuous signal into two digital pulse outputs, UP or DOWN, according to the signal derivative. The UP (DOWN) spikes are generated every time the signal exceeds a positive (negative) threshold, like the ON (OFF) events from the DVS. As described before, the signal is sampled at 200Hz, this means that a new sample is acquired every $5\ \text{ms}$. To increase the time resolution of the generated spike train, which otherwise would contain too few spikes, the EMG signals are over-sampled to a higher frequency before undergoing the transformation into spikes (Donati et al., 2019).

For our specific EMG acquisition features, we set the threshold at 0.05 and an interpolation factor of 3500, these values have been selected from previous studies which looked at quality of signal reconstruction (Donati et al., 2018, 2019).

2.3 Neuromorphic processors

2.3.1 ODIN + MorphIC

The ODIN (Online-learning DIgital spiking Neuromorphic) processor occupies an area of only 0.086mm^2 in 28nm FDSOI CMOS (Frenkel et al., 2019a)[1]. It consists of a single neurosynaptic core with 256 neurons and 256^2 synapses. Each neuron can be configured to phenomenologically reproduce the 20 Izhikevich behaviors of spiking neurons (Izhikevich, 2004). The synapses embed a 3-bit weight and a mapping table bit that allows enabling or disabling Spike-Dependent Synaptic Plasticity (SDSP) locally (Brader et al., 2007), thus allowing for the exploration of both off-chip training and on-chip online learning setups.

[1] The HDL source code and documentation of ODIN are publicly available at `https://github.com/ChFrenkel/ODIN`.

MorphIC is a quad-core digital neuromorphic processor with 2k LIF neurons and more than 2M synapses in 65nm CMOS (Frenkel et al., 2019b). MorphIC was designed for high-density large-scale integration of multi-chip setups. The four 512-neuron crossbar cores are connected with a hierarchical routing infrastructure that enables neuron fan-in and fan-out values of 1k and 2k, respectively. The synapses are binary and can be either programmed with offline-trained weights or trained online with a stochastic version of SDSP.

Both ODIN and MorphIC follow a standard synchronous digital implementation, which allows their operation to be predicted with one-to-one accuracy by custom Python-based chip simulators. As both chips rely on crossbar connectivity, CNN topologies can be explored but are limited to small networks due to an inefficient resource usage in the absence of a weight reuse mechanism (Frenkel et al., 2019b). The selected SNN architectures are thus based on fully-connected MLP topologies. Training is carried out in Keras with quantization-aware stochastic gradient descent following a standard ANN-to-SNN mapping approach (Hubara et al., 2017; Moons et al., 2017; Rueckauer et al., 2017), the resulting SNNs process the EMG and DVS spikes without further preprocessing.

In order to process the spike-based EMG gesture data, we selected ODIN so as to benefit from 3-bit weights. Indeed, due to the low input dimensionality of EMG data, satisfactory performance could not be reached with the binary weight resolution of MorphIC. A 3-bit-weight `16-230-5` SNN is thus implemented in ODIN, this setup will be referred to as the EMG-ODIN network.

For the DVS gesture data classification, we selected MorphIC in order to benefit from its higher neuron and synapse resources. ON/OFF DVS events are treated equally and their connections to the network are learned, so that any of them can be either excitatory or inhibitory. Similarly to a setup previously proposed for MNIST benchmarking (Frenkel et al., 2019b), the input 40$\times$40-pixel DVS event streams can be subsampled into four 20$\times$20-pixel event streams and processed independently in the four cores of MorphIC, thus leading to an accuracy boost when combining the outputs of all subnetworks, subsequently denoted as subMLPs. The four subMLPs have a `400-210-5` topology with binary weights, this setup will thus be referred to as the DVS-MorphIC network.

In order to ease sensor fusion, the hidden layer sizes of the EMG-ODIN and DVS-MorphIC networks and the associated firing thresholds were optimized by parameter search so as to balance their activities. These hidden layers were first flattened into a 1070-neuron layer, then a 5-neuron output layer was retrained with 3-bit weights and implemented in ODIN. This setup will be referred to as the Fusion-ODIN network, which thus encapsulates EMG processing in ODIN, DVS processing in MorphIC and sensor fusion in ODIN. From an implementation point of view, mapping the MorphIC hidden layer output spikes back to ODIN for sensor fusion requires an external mapping table. Its overhead is excluded from the results provided in Section 3.

2.3.2 Loihi and its training framework SLAYER

Intel's Loihi (Davies et al., 2018) is an asynchronous neuromorphic research processor. Each Loihi chip consists of 128 neurocores, with each neurocore capable of implementing up to 1024 current based (CUBA) Leaky Integrate and Fire (LIF) neurons. The network state and configuration is stored entirely in on-chip SRAMs local to each core, this allows each core to access its local memories independently of other cores without needing to share a global memory bus (and in fact removing the need for off-chip memory). Loihi supports a number of different encodings for representing network connectivity, thus allowing the user to choose the most efficient encoding for their task. Each Loihi chip also contains three small synchronous x86 processors which help monitor and configure the network, as well as assisting with the injection of spikes and recording of output spikes.

SLAYER (Shrestha and Orchard, 2018) is a backpropagation framework for evaluating the gradient of any kind of SNN (i.e. spiking MLP and spiking CNN) directly in the spiking domain. It is a dt-based SNN backpropagation algorithm that keeps track of the internal membrane potential of the spiking neuron and uses it during gradient propagation. There are two main guiding principles of SLAYER: temporal credit assignment policy and probabilistic spiking neuron behavior during error backpropagation. Temporal credit assignment policy acknowledges the temporal nature of a spiking neuron where a spike event at a particular time has its effect on future events. Therefore, the error credit of an error at a particular time needs to be distributed back in time. SLAYER is one of the few methods that consider temporal effects during backpropagation. The use of probabilistic neurons during backpropagation helps estimate the spike function derivative, which is a major challenge for SNN backpropagation, with the spike escape rate function of a probabilistic neuron. The end effect is that the spike escape rate function is used to estimate the spike function derivative, similar to the surrogate gradient concept (Zenke and Ganguli, 2018; Neftci et al., 2019). With SLAYER, we can train synaptic weights as well as axonal delays and achieve state of the art performances (Shrestha and Orchard, 2018) on neuromorphic datasets.

SLAYER uses the versatile Spike Response Model (SRM) (Gerstner, 1995) which can be customized to represent a wide variety of spiking neurons with a simple change of spike response kernels. It is implemented[2] atop the PyTorch framework with automatic differentiation support (Paszke et al., 2017) with the flexibility of feedforward dense, convolutional, pooling and skip connections in the network.

[2] SLAYER-PyTorch is publicly available at `https://github.com/bamsumit/slayerPytorch`.

SLAYER-PyTorch also supports training with the exact CUBA Leaky Integrate and Fire neuron model in Loihi (Davies et al., 2018). To train for the fixed precision constraints on weights and delays of Loihi hardware, it trains the network with the quantization constraints and then trains using the strategy of shadow variables (Courbariaux et al., 2015; Hubara et al., 2016) where the constrained network is used in forward propagation phase and the full precision shadow variables are used during backpropagation.

We used SLAYER-PyTorch to train a Loihi compatible network for the hand-gesture recognition task. The networks were trained offline using GPU and trained weights and delays were used to configure the network on Loihi hardware for inference purposes. All the figures reported here are for inference using Loihi with one algorithmic time tick in Loihi of 1 ms.

A spiking MLP of architecture `16-128d-128d-5` was trained for EMG gestures converted into spikes (Section 2.2.1). Here, `128d` means the fully connected layer has 128 neurons with trained axonal delays. The Loihi neuron with current and voltage decay constants of 1024 (32 ms) was used for this network.

For the gesture classification using DVS data we used both a spiking MLP, with the same architecture as the one deployed on MorphIC and described in Section 2.3.1, and a spiking CNN with architecture `40x40x2-8c3-2p-16c3-2p-32c3-512-5`. Here, `XcY` denotes a convolution layer with `X` kernels of shape `Y-by-Y`, while `2p` denotes a `2-by-2` max pooling layer. Zero padding was applied for all convolution layers. No preprocessing on the spike events was performed, the ON/OFF events are treated as different input channels, hence the input shape `40x40x2`. For this network, current and voltage decay constants for the Loihi neurons were set to 1024 (32 ms) and 128 (4 ms).

Finally, a third network where the penultimate layer neurons of DVS and EMG networks were fused together was trained. Only the last fully connected weights (`640-5`) were trained. The parameters of the network before fusion were preserved. The current and voltage decay constants of 1024 (32 ms) and 128 (4 ms) respectively were used for the final fusion layer neurons. From now on, we will refer these three networks as EMG-Loihi, DVS-Loihi, and Fusion-Loihi whenever there is ambiguity.

2.4 Traditional machine learning baselines

Machine Learning (ML) methods, and in general data-driven approaches, are currently the dominant tools to solve complex classification tasks since they give the best performance compared to other approaches. We compare the performance of the two fully neuromorphic systems described in the above sections, against a traditional machine learning pipeline that uses frame-based inputs, i.e. traditionally sampled EMG signals and traditionally sampled video frames. In order for the comparisons to be fair, for the traditional approach we maintain the same constraints imposed by the neuromorphic hardware. In particular, we used the same neural network architectures as those used in the neuromorphic systems. Note that two different networks were implemented, spiking MLP and spiking CNN (see Figure 3 for more details on the architectures). For this reason, we have two different baseline models that are paired to the two considered neuromorphic systems.

2.4.1 EMG Feature extraction

Traditional EMG signal processing consists of various steps. First, signal pre-processing is used to extract useful information by applying filters and transformations. Then, feature extraction is used to highlight meaningful structures and patterns. Finally, a classifier maps the selected features to output classes. In this section we describe the EMG feature extraction phase, in particular we consider time domain features used for the classification of gestures with the baseline models. We extracted two time domain features generally used in literature (Phinyomark et al., 2018), namely Mean Absolute Value (MAV) and Root Mean Square (RMS) shown in Equations 1. The MAV is the average of the muscles activation value and it is calculated by a stride-moving window. The RMS is represented as amplitude relating to a gestural force and muscular contraction. The two features are calculated across a window of 40 samples, corresponding to 200 ms:

$$MAV(x_c) = \frac{1}{T}\sum_{t=0}^{T}|x_c(t)| \qquad RMS(x_c) = \sqrt{\frac{1}{T}\sum_{t=0}^{T}x_c^2(t)} \tag{1}$$

where $x_c(t)$ is the signal in the time domain for the EMG channel with index c and T is the number of samples in the considered window, which was set to be $T = 40$ ($N = 200$ ms) across this work. The features were calculated for each channel separately and the resulting values were concatenated in a vector $\mathbf{F}(n)$ described in Equation 2:

$$\mathbf{F}(n) = [F(x_1), ..., F(x_C)]^T \tag{2}$$

where $\mathbf{F}$ is MAV or RMS, n is the index of the window and C is the number of EMG channels. The final feature vector $\mathbf{E}(n)$ for window n is shown in Equation 3, it is used for the classification and is obtained by concatenating the two single feature vectors

$$\mathbf{E}(n) = \left[\mathbf{MAV}(n)^T, \mathbf{RMS}(n)^T\right]^T \tag{3}$$

2.4.2 Baseline ODIN + MorphIC

As described in Section 2.3.1, a CNN cannot be efficiently implemented on crossbar cores, which is the architecture ODIN and MorphIC rely on. We will therefore rely solely on fully-connected MLPs networks for both visual and EMG data processing. For the visual input, we used the same subMLP-based network structure as the one described in Section 2.3.1, but with gray-scale APS frames.The 40x40 cropped APS frames are sub-sampled and fed into four 2-layer subMLPs of architecture `400-210-5`, as shown in Figure 3 panel (b). The outputs of the 4 subMLPs are then summed when classifying with a single sensor and are concatenated for the fusion network. The EMG neural network is a 2-layer MLP of architecture `16-230-5`. The fusion network is obtained as described above for the Loihi baseline.

2.4.3 Baseline Loihi

As described in Section 2.3.2, we used a spiking MLP and a spiking CNN to process and classify DVS events. For the Loihi baseline, we kept the exact same architectures, except for the axonal delays. Moreover, both architectures of the baseline receive the corresponding gray-scale APS frames instead of the DVS events. The baseline MLP architecture and the CNN architectures are shown in Figure 3 panel (a) and (b) respectively. Note that the number of parameters between the baseline networks and the spiking networks implemented on Loihi is slightly different since the input has 1 channel (gray-scale) in the case of the baseline that uses APS frames while it has 2 channels (polarity) in the input for Loihi.

The MLP architecture used for the EMG classification is instead composed of 2 layers of 128 followed by one layer of 5 units. While the input stays of the same size (16) with respect to the network implemented on Loihi, the input features are different since the baseline MLP receives MAV and RMS features while the Loihi receives spikes obtained from the raw signal.

To obtain the fusion network, we eliminate the last layer (classification layer) from both the single sensor networks, concatenate the two penultimate layers of the single sensor networks and add a common classification layer with 5 units, one per each class.

2.4.4 Training and Deployment

The models are trained with Keras using Adam optimizer with standard parameters. First, the single modality networks are trained separately, each for 30 epochs. For sensor fusion, output layer retraining is also carried out for 30 epochs. In order to compare the baselines against the neuromorphic systems in terms of energy consumption and inference time, we deployed the baseline models onto the NVIDIA Jetson Nano, an embedded system with a 128-Core Maxwell GPU with 4GB 64-bit LPDDR4 memory 25.6 GB/s[3].

3 RESULTS

Table 1 summarizes the results for Loihi and ODIN+MorphIC with the respective baselines. More details are described in the following sections.

[3] `https://developer.nvidia.com/embedded/jetson-nano-developer-kit`

3.1 Loihi results

The classification performances of these three networks, EMG-Loihi, DVS-Loihi, and Fusion-Loihi, with three-fold cross-validation and inferenced using 200 ms data are tabulated in Table 2. The core utilization, dynamic power consumption and inference time in the Loihi hardware are also listed in Table 2. The dynamic power is measured as the difference of total power consumed by the network and the static power when the chip is idle. Since one algorithmic time tick is 1ms long, inference time represents the speedup factor compared to real time.

With the spiking MLP implemented on Loihi, we obtained an accuracy of $50.3 \pm 1.5\%$, $83.1 \pm 3.4\%$ and $83.4 \pm 2.1\%$ for the hand-gesture classification task using EMG, DVS and fusion respectively. Being that these results were significantly worse than the ones obtained with the spiking CNN, we do not report them in Table 1 and Table 2 and prefer to focus our analysis on the CNN which is better suited for visual tasks. This poor performance is due to temporal resolution of Loihi that causes a drop in the number of spikes in the MLP architecture while this does not happen in the CNN architecture.

The EMG network does not perform as well as in the baseline as shown in Table 1. The reason for this discrepancy can be found in the fact that the baseline method uses EMG from the raw signal of the sensor. However, to process this signal using neuromorphic chips (Loihi and ODIN+MorphIC), the EMG signal is encoded into spikes. With this encoding, part of the information is lost (as is the case for any encoding). Therefore, the baseline method has the advantage of using a signal that has more information and thus it outperforms the neuromorphic approach. Note that these Loihi networks are restricted to 8-bit fixed precision weights and 6-bit fixed precision delays.

To evaluate the performance over time of the Loihi networks, stimulus duration versus testing accuracy is plotted in Figure 4. We can see that the EMG-Loihi network continues to improve with longer stimulus duration. Table 1 and Figure 4 show the results of the Loihi baseline. From an accuracy point of view the baseline reaches a higher classification accuracy only in the EMG classification, while both the visual classification and fusion are on par with the Loihi networks and show only a non-significant difference. In terms of inference time, the baseline running on the GPU system is systematically faster than Loihi, but never more than 40% faster. As expected, the energy consumption of the GPU system is significantly higher than the Loihi system. Loihi is around 30x more efficient than the baseline for what concerns the fusion network and more than 150x and 40x more efficient for what concerns the EMG and DVS processing respectively. Figure 4 shows in more details the effect of stimulus duration on the classification accuracy. As expected, EMG is the modality that suffers more from classification based on short segments (Smith et al., 2011), reaching for both the neuromorphic system and the baseline the best accuracy only after 200 ms, while the accuracy for vision and fusion modalities saturate much more quickly, in around 100 ms for the neuromorphic system and 50 ms for the baseline. The traditional system reaches its best performance after 50 ms while the neuromorphic system reaches its best performance after 200ms. One should, however, also note that the DVS sensor contains only the edge information of the scene whereas the baseline network uses the image frame. Therefore, the spiking CNN requires some time to integrate the input information from DVS. Despite the inherent delays in a spiking CNN, the Loihi CNN can respond to the input within a few ms of inputs. However, for the vision modality, notice that, because the frame rate of the camera is 20 fps, there is no classification before 25ms. Therefore, for short stimulus duration, the neuromorphic system has higher accuracy than the traditional system.

3.2 ODIN + MorphIC results

Inference statistics for a 200 ms sample duration are reported in Table 3 for the EMG-ODIN, DVS-MorphIC and Fusion-ODIN networks. Chip utilization is computed as the percentage of neuron resources taken by the hidden and output layers in ODIN and MorphIC, while the power consumption P of the crossbar cores of both chips can be decomposed as

$$P = P_{\text{leak}} + P_{\text{idle}} f_{\text{clk}} + E_{\text{SOP}} r_{\text{SOP}}, \tag{4}$$

where P_{leak} is the chip leakage power and $P_{\text{leak}} + P_{\text{idle}} f_{\text{clk}}$ represents the static power consumption when a clock of frequency f_{clk} is connected, without network activity. The term $E_{\text{SOP}} r_{\text{SOP}}$ thus represents the dynamic power consumption, where E_{SOP} is the energy per synaptic operation (SOP) and r_{SOP} is the SOP processing rate, each SOP taking two clock cycles. Detailed power models extracted from chip measurements of ODIN and MorphIC are provided in (Frenkel et al., 2019a) and (Frenkel et al., 2019b), respectively. The results reported in Tables 1 and 3 are obtained with ODIN and MorphIC optimizing for power, under the conditions summarized in Table 4. The dynamic power consumption reported in Table 4 reflects the regime in which ODIN and the four cores of MorphIC run at the maximum SOP processing rate $r_{\text{SOP}} = f_{\text{clk}}/2$.

A limitation of the crossbar-based architecture of ODIN and MorphIC is that each neuron spike leads to a systematic processing of all neurons in the core, thus potentially leading to a significant amount of dummy operations (Frenkel et al., 2019b). Taking the example of the DVS-MorphIC network with a crossbar core of 512 neurons (Figure 3, panel (b)), each input spike leads to 512 SOPs, of which only 210 are useful for hidden layer processing. Similarly, each spike from a hidden layer neuron leads to 512 SOPs, of which only 5 are actually used for output layer processing. The induced overhead is thus particularly critical for output layer processing, which degrades both the energy per inference and the inference time.[4] However,

[4] As discussed in (Frenkel et al., 2019b), a simple extension providing post-synaptic start and end addresses would avoid these dummy SOPs and allow for an efficient processing of fully-connected layers, which will be included in future generations of the chips.

this problem is partly mitigated in the Fusion-ODIN network for output layer processing. Indeed, when resorting to an external mapping table (Section 2.3.1), hidden layer spikes can be remapped back to the sensor fusion output layer of ODIN with specific single-SOP AER events (Frenkel et al., 2019a), thus avoiding the dummy SOP overhead and leading to a lower energy and inference time compared to the standalone EMG-ODIN and DVS-MorphIC networks (Tables 1 and 3). As described in Section 2.3.1, the fusion results exclude the mapping table overhead.

The comparison of the results obtained with ODIN + MorphIC to those obtained with its GPU baseline counterpart (Table 1 and Figure 5) leads to conclusions similar to those already drawn with Loihi in Section 3.1, with the difference that while the GPU system is significantly faster, between 2x to 10x faster, the ODIN + MorphIC neuromorphic system is between 500× and 3200× more energy-efficient. Moreover, it appears from Figure 5 that the EMG-ODIN, DVS-MorphIC and Fusion-ODIN networks basically perform at chance level for a 10-ms stimulus duration. This comes from the fact that the firing thresholds of the networks were selected based on a 200-ms stimulus duration, which leads the output neurons to remain silent and never cross their firing threshold when insufficient input spike data is provided. This problem could be alleviated by reducing the neuron firing thresholds for shorter stimulus durations.

3.3 EDP and computational complexity

Figure 6 shows a comparison between the Loihi system and the ODIN + MorphIC system in terms of EDP, number of operations per classification and a ratio between these two quantities. While panel (a) reports the same numbers as in Table 1, panels (b) and (c), allow for a more fair comparison of energy consumption between the two neuromorphic systems. From panel (b), we can see how the number of operations is similar for the EMG networks, being both MLPs for the two neuromorphic systems. Differently, the numbers of operations for the visual input and the fusion differ substantially between the two systems due to the use of a CNN in the Loihi system. Taking this into account, we can see in panel (c) that the normalized energy consumption tends to be similar for both systems more than the EDP in panel (a) is.

4 DISCUSSIONS

As it has been discussed in (Davies, 2019), there is a real need for a benchmark in the neuromorphic engineering field to compare the metrics of accuracy, energy, and latency. ML benchmarks such as ImageNet for image classification (Deng et al., 2009), Chime challenges for speech recognition (Barker et al., 2015) and Ninapro dataset containing kinematic and surface EMG for prosthetic applications (Atzori et al., 2014) are not ideal for neuromorphic chips as they require high performance computing for processing. For example, floating point bit resolution, large amounts of data and large power consumption. There have been some efforts in creating relevant event-based datasets such as N-MNIST (Orchard et al., 2015), the spiking version of the widespread MNIST digits recognition dataset, N-TIDIGITS18 (Anumula et al., 2018), the spiking version of the spoken digits recognition dataset from LDC TIDIGITS, and DVS gesture recognition dataset from IBM (Amir et al., 2017). These datasets are either toy examples, or are not meant for real-world applications. Here, we are introducing a hand gesture benchmark in English sign language (e.g. ILY) using the DVS and Myo sensors. This kind of benchmark can be directly used as a preliminary test for Brain-Machine Interface (BMI)/personalized medicine applications. We have collected this dataset from 21 people and in this paper have benchmarked it on three digital neuromorphic chips, measuring the accuracy, energy and inference time. We believe this work takes an important first step in the direction of a real use-case (e.g. rehabilitation, sports applications, and sign interpretation) which we would like to encourage the community to use.

Although the dataset we provided is on static gestures, the DVS and the spiking EMG signals provide the capability for low-power processing using event-based neuromorphic chips and enable embedded systems with online on-site processing without having to send the data to remote sensors. Therefore, this work is an important first step towards edge-computing applications. The static dataset also helps with reducing the noise from the EMG signals as we mentioned in Section 2.2. However, this does not move away from the real application as we have shown in a live demo in (Ceolini et al., 2019).

The selected multi-sensor data fusion, that combines vision and EMG sensors, derives from the need of multiple sources to help the classification in real-scenario cases. Although the results show a small improvement due to the EMG sensors, they still provide some classification in case of not ideal light conditions or camera occlusions. In addition, for specific applications such as neuroprosthetic control, the EMG is integrated in the prosthetic device and, eventually, the camera can act as support input helping during calibration or more advanced tasks, such as sensory-motor closed loop (Jiang et al., 2012).

Since the event-based neuromorphic chips require inputs in the form of events, the continuous sensory signals have to be encoded into spikes for an event-driven processing. This quantization loses information (and hence accuracy) in comparison to the analog information processing in trade-off with the low power consumption of event-based systems which is required for edge computing. To compensate for the loss of information and accuracy, it is important to merge information from multiple sensors in a sensory fusion setup. In this setting, the information loss by quantization from one sensor can be made up for by another one. This is similar to how humans and animals perceive their environment through diverse sensory channels: vision, audition, touch, smell, proprioception, etc. From a biological perspective, the fundamental reason lies in the concept of degeneracy in neural structures (Edelman, 1987), which means that any single function can be carried out by more than one configuration of neural signals, so that the biological system still functions with the loss of one component. It also means that sensory systems can educate each other, without an external teacher (Smith and Gasser, 2005). The same principles can be applied for artificial systems, as information about the same phenomenon in the environment can be acquired from various types of sensors: cameras, microphones, accelerometers, etc. Each sensory-information can be considered as a modality. Due to the rich characteristics of natural phenomena, it is rare that a single modality provides a complete representation of the phenomenon of interest (Lahat et al., 2015).

There are mainly two strategies for multi-modal fusion in the literature (Cholet et al., 2019): (1) data-level fusion (early fusion) where modalities are concatenated then learned by a unique model, and (2) score-level fusion (late fusion) where modalities are learned by distinct models and only after their predictions are fused with another model that provides a final decision. Early fusion, including feature-level fusion, suffers from a compatibility problem (Peng et al., 2016) and does not generalize well. Additionally, neural-based early fusion increases the memory footprint and the computational cost of the process, by inducing a full connectivity at the first classification stages. It is an important factor to take in consideration when choosing a fusion strategy (Castanedo, 2013), especially for embedded systems. Therefore, we follow a late fusion approach with a classifier-level fusion, that has been shown to perform better than feature-level fusion for classification tasks (Guo et al., 2014; Peng et al., 2016; Biagetti et al., 2018). It is close to score-level fusion by combining the penultimate layers of the base (unimodal) classifiers in a meta-level (multimodal) classifier that uses the natural complementarity of different modalities to improve the overall classification accuracy.

In this context, to have a fair comparison, the central question is the difference between the completely traditional approaches, such as the CNN and MLP baselines, versus the event-based neuromorphic one. In the baseline, the EMG features are manually extracted and the classification is done on the extracted features. Note that this pipeline is completely different from the event-based neuromorphic approach which extracts the features directly from the events. Another important thing to mention here is that although we have encoded the signals separately, this sensory information can be directly encoded to events at the front-end. This has already been established for audio and visual sensors (Lichtsteiner et al., 2006; Chan et al., 2007) and there have also recently been design efforts for other signals such the biomedical ones (Corradi and Indiveri, 2015).

To have a reference point for comparison, we trained the same network architecture used for the two neuromorphic setups. As it can be seen in Table 1, the baseline accuracy on the fusion is on par with both Loihi and ODIN+MorphIC, despite the lower bit resolution on the neuromorphic chips in comparison with the 32-bit floating point resolutions on GPU in the baseline approach. We speculate this is because the SLAYER training model already takes into account the low bit precision and thus calculates the gradients respectively. Similar to that, ODIN and MorphIC take a quantization-aware training approach which calculates the weights based on the available on-chip precision. As can be seen from all the experiments in Table 1, the classification accuracy using only the EMG sensor is relatively low. However, it is to be noticed that this is a result of having a model which is trained across subjects and there are multiple sources of variability across subjects: i) The placement of the EMG sensor is not necessarily in the same position (with respect to the forearm muscles) for every subject. ii) Every subject performs the gestures in a unique manner iii) The muscle strength is different for every subject. In addition, since the EMG is directly measured from surface electrodes, it acquires noise while traveling through the skin, background noise from electronics, ambient noise, and so forth. In a real-world application, the network model can be trained on a single subject's data yielding much higher accuracy. Moreover, having the online learning abilities on the neuromorphic chip can aid in adapting these models to every subject uniquely. Such online learning modules are already existent in Loihi as well as in ODIN and MorphIC, which can be exploited in the future for boosting the classification accuracy of EMG signals. Also, it becomes apparent that the fusion accuracy is close, even if higher of about 4%, to the accuracy achieved with the DVS single sensor. However, the importance of the EMG signal is in the wearable application since it is a natural way to control prosthesis and it is a direct measure of the activity and movement in the muscles. Given the noisy nature of the EMG signal, it is critical to combine it with the visual input to boost the accuracy. But even given the noisy nature of the signal, it still allows to retrieve relevant information which helps boosting the accuracy of the fusion.

It is worth noting that while the accuracy between the spiking MLP on Loihi and ODIN+MorphIC are directly comparable, the results regarding the spiking CNN on Loihi and the spiking MLP on ODIN+MorphIC are not. This is because the two architectures use different features and resources on their respective neuromorphic systems (as already described in Section 2.3). Based on this, there are different constraints present in the two chips. Traditionally, a CNN architecture is used for image classification which is the network we used on the Loihi chip given the large number of neurons that are available (128k) on this general purpose platform. However, since ODIN and MorphIC are small-scale devices compared to Loihi, the number of neurons are a lot more constrained (i.e. 256 neurons for ODIN, 2k for MorphIC). Therefore, we resorted to using a fully-connected MLP topology instead of a CNN for image classification in MorphIC.

Regarding the latency, it is important to mention that for real-world prosthetic applications, the latency budget is below 250 ms (Smith et al., 2011). This means that if the processing happens within this budget, the patient will not feel the lag of the system. Hence, optimizing the system for having lower latency than 200 ms will not be beneficial as the patient will not feel the latency below 200 ms. Therefore, within this budget, other parameters can be optimized. The neuromorphic approach is very advantageous in this case since it trades-off power with latency but it stays within the latency budget that is required. Contrarily, the GPU system has an overall faster inference time but uses much more energy. It is worth mentioning that our results are reported in accelerated time, however, the EMG and DVS are slowly changing signals and thus even though the classification is done very fast, the system has to wait for the inputs to arrive. Therefore, it is as if the system is being run in real-time. Here, there is a trade-off between the memory that is storing the streaming data for processing and the dynamic energy consumption. The accelerated time allows the lower energy consumption as the system is on for a shorter time, however, this comes with the caveat that the input has to be buffered for at least 200 ms in off-chip memory, therefore inducing a power and resource overhead.

The final comparison provided by Figure 6 shows how the two systems have a similar energy consumption when this is normalized by the number of operations done to run the network and obtain one classification output. While ODIN + MorphIC consumes less per classification in absolute terms, when considering the number of operations, it performs comparably to Loihi. When deploying a neuromorphic system one has to take into account all these aspects. Meaning not only there is a trade-off between speed and energy consumption but there is also one between accuracy and energy consumption given the fact that a more complex network architecture may have more predictive power while coming with a higher energy demand. Overall, one has to look for the best trade-off in the context of a particular application, the malleability of neuromorphic hardware enables this adaptation to the task-dependent constraints within a framework of state of the art results with respect to system performance.

CONFLICT OF INTEREST STATEMENT

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

AUTHOR CONTRIBUTIONS

EC, CF, SBS contributed equally to the work. EC, GT, MP, ED participate equally to the development of the work idea and collected the dataset. EC, LK were responsible for the baselines experiments. CF and SBS implemented the ODIN+MorphIC and Loihi pipelines respectively. SBS implemented the SLAYER framework and adapted it for the specific application. Everyone contributed to the writing of the paper.

FUNDING

This work is supported by the EU's H2020 MSC-IF grant NEPSpiNN (Grant No. 753470), the Swiss Forschungskredit grant FK-18-103, the Toshiba Corporation, the Swiss Forschungskredit grant FK-19-106, the SNSF grant No. 200021_172553 the fonds européen de développement régional FEDER, the Wallonia within the "Wallonie-2020.EU" program, the Plan Marshall, the FRS-FNRS of Belgium, and the H2020 MC SWITCHBOARD ETN (Grant No. 674901).

The authors declare that this study received funding from Toshiba Corporation. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

The authors would like to acknowledge the 2019 Capocaccia and Telluride Neuromorphic Workshops and all theirs participants for the fruitful discussions, Intel Corporation for access to Loihi neuromorphic platform. We thank Prof. B. Miramond, Prof. D. Bol, Prof. S. Liu, Prof. T. Delbruck and Prof. G. Indiveri. Finally, we thank Garrick Orchard for supporting us with the use of the Loihi platform and the useful comments to the paper.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the Zenodo, open access repository (Ceolini et al., 2019b), `http://doi.org/10.5281/zenodo.3663616`. Ceolini, Enea, Taverni, Gemma, Payvand, Melika, & Donati, Elisa. (2020). EMG and Video Dataset for sensor fusion based hand gestures recognition (Version 3.0). All the code used for the reported experiments can be found at `https://github.com/Enny1991/dvs_emg_fusion`

REFERENCES

Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Nolfo, C. D., et al. (2017). A low power, fully event-based gesture recognition system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7388–7397. doi:10.1109/CVPR.2017.781

Anumula, J., Neil, D., Delbruck, T., and Liu, S.-C. (2018). Feature representations for neuromorphic audio spike streams. *Frontiers in neuroscience* 12, 23

Atzori, M., Gijsberts, A., Castellini, C., Caputo, B., Hager, A.-G. M., Elsig, S., et al. (2014). Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific data* 1, 140053

Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2015). The third 'chime' speech separation and recognition challenge: Dataset, task and baselines. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* , 504–511

Benatti, S., Casamassima, F., Milosevic, B., Farella, E., Schönle, P., Fateh, S., et al. (2015). A versatile embedded platform for emg acquisition and gesture recognition. *IEEE transactions on biomedical circuits and systems* 9, 620–630

Benjamin, B. V., Gao, P., McQuinn, E., Choudhary, S., Chandrasekaran, A. R., Bussat, J.-M., et al. (2014). Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations. *Proceedings of the IEEE* 102, 699–716

Biagetti, G., Crippa, P., and Falaschetti, L. (2018). Classifier level fusion of accelerometer and semg signals for automatic fitness activity diarization. *Sensors* 18, 2850. doi:10.3390/s18092850

Brader, J. M., Senn, W., and Fusi, S. (2007). Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural computation* 19, 2881–2912

Braun, S., Neil, D., Anumula, J., Ceolini, E., and Liu, S. (2019). Attention-driven multi-sensor selection. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–8. doi:10.1109/IJCNN.2019.8852396

Castanedo, F. (2013). A review of data fusion techniques. *TheScientificWorldJournal* 2013, 704504. doi:10.1155/2013/704504

Ceolini, E., Taverni, G., Khacef, L., Payvand, M., and Donati, E. (2019). Live demostration: Sensor fusion using emg and vision for hand gesture classification in mobile applications. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. 1–1. doi:10.1109/BIOCAS.2019.8919163

Ceolini, E., Taverni, G., Khacef, L., Payvand, M., and Donati, E. (2019a). Sensor fusion using emg and vision for hand gesture classification in mobile applications. *arXiv preprint arXiv:1910.11126*

[Dataset] Ceolini, E., Taverni, G., Payvand, M., and Donati, E. (2019b). EMG and Video Dataset for sensor fusion based hand gestures recognition. doi:10.5281/zenodo.3228846

Chan, V., Liu, S.-C., and van Schaik, A. (2007). Aer ear: A matched silicon cochlea pair with address event representation interface. *IEEE Transactions on Circuits and Systems I: Regular Papers* 54, 48–59

Chen, C., Yu, Y., Ma, S., Sheng, X., Lin, C., Farina, D., et al. (2020). Hand gesture recognition based on motor unit spike trains decoded from high-density electromyography. *Biomedical Signal Processing and Control* 55, 101637

Cheok, M. J., Omar, Z., and Jaward, M. H. (2019). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics* 10, 131–153

Cholet, S., Paugam-Moisy, H., and Regis, S. (2019). Bidirectional associative memory for multimodal fusion : a depression evaluation case study. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–6. doi:10.1109/IJCNN.2019.8852089

Cicirelli, G., Attolico, C., Guaragnella, C., and D'Orazio, T. (2015). A kinect-based gesture recognition approach for a natural human robot interface. *International Journal of Advanced Robotic Systems* 12, 22

Corradi, F. and Indiveri, G. (2015). A neuromorphic event-based neural recording system for smart brain-machine-interfaces. *IEEE transactions on biomedical circuits and systems* 9, 699–709

Courbariaux, M., Bengio, Y., and David, J.-P. (2015). Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*. 3123–3131

Davies, M. (2019). Benchmarks for progress in neuromorphic computing. *Nature Machine Intelligence* 1, 386–388

Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., et al. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro* 38, 82–99

Debes, C., Merentitis, A., Heremans, R., Hahn, J., Frangiadakis, N., van Kasteren, T., et al. (2014). Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7. doi:10.1109/JSTARS.2014.2305441

Deiss, S. R., Douglas, R. J., Whatley, A. M., et al. (1999). A pulse-coded communications infrastructure for neuromorphic systems. *Pulsed neural networks* , 157–178

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (Ieee), 248–255

Donati, E., Payvand, M., Risi, N., Krause, R., Burelo, K., Indiveri, G., et al. (2018). Processing emg signals using reservoir computing on an event-based neuromorphic system. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (IEEE), 1–4

Donati, E., Payvand, M., Risi, N., Krause, R. B., and Indiveri, G. (2019). Discrimination of emg signals using a neuromorphic implementation of a spiking neural network. *IEEE transactions on biomedical circuits and systems*

Došen, S., Cipriani, C., Kostić, M., Controzzi, M., Carrozza, M. C., and Popović, D. B. (2010). Cognitive vision system for control of dexterous prosthetic hands: experimental evaluation. *Journal of neuroengineering and rehabilitation* 7, 42

Droniou, A., Ivaldi, S., and Sigaud, O. (2015). Deep unsupervised network for multimodal perception, representation and classification. *Robotics and Autonomous Systems* 71, 83 – 98. doi:https://doi.org/10.1016/j.robot.2014.11.005. Emerging Spatial Competences: From Machine Perception to Sensorimotor Intelligence

Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection* (New York, US: Basic Books)

Frenkel, C., Lefebvre, M., Legat, J.-D., and Bol, D. (2019a). A 0.086-mm^2 12.7-pj/sop 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28-nm cmos. *IEEE Transactions on Biomedical Circuits and Systems* 13, 145–158

Frenkel, C., Legat, J.-D., and Bol, D. (2019b). Morphic: A 65-nm 738k-synapse/mm^2 quad-core binary-weight digital neuromorphic processor with stochastic spike-driven online learning. *IEEE Transactions on Biomedical Circuits and Systems* 13, 999–1010

Furber, S. B., Galluppi, F., Temple, S., and Plana, L. A. (2014). The spinnaker project. *Proceedings of the IEEE* 102, 652–665

Gerstner, W. (1995). Time structure of the activity in neural network models. *Phys. Rev. E* 51, 738–758. doi:10.1103/PhysRevE.51.738

Ghazaei, G., Alameer, A., Degenaar, P., Morgan, G., and Nazarpour, K. (2017). Deep learning-based artificial vision for grasp classification in myoelectric hands. *Journal of neural engineering* 14, 036025

Gigli, A., Gregori, V., Cognolato, M., Atzori, M., and Gijsberts, A. (2018). Visual cues to improve myoelectric control of upper limb prostheses. In *2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob)* (IEEE), 783–788

Guo, H., Chen, L., Shen, Y., and Chen, G. (2014). Activity recognition exploiting classifier level fusion of acceleration and physiological signals. *UbiComp 2014 - Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* , 63–66doi:10.1145/2638728.2638777

Haria, A., Subramanian, A., Asokkumar, N., Poddar, S., and Nayak, J. S. (2017). Hand gesture recognition for human computer interaction. *Procedia computer science* 115, 367–374

Hoeks, C., Barentsz, J., Hambrock, T., Yakar, D., Somford, D., Heijmink, S., et al. (2011). Prostate cancer: Multiparametric mr imaging for detection, localization, and staging. *Radiology* 261, 46–66. doi:10.1148/radiol.11091822

Horwitz, B. and Poeppel, D. (2002). How can eeg/meg and fmri/pet data be combined? *Human brain mapping* 17, 1–3. doi:10.1002/hbm.10057

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., and Bengio, Y. (2016). Binarized neural networks. In *Advances in neural information processing systems*. 4107–4115

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., and Bengio, Y. (2017). Quantized neural networks: Training neural networks with low precision weights and activations. *The Journal of Machine Learning Research* 18, 6869–6898

Izhikevich, E. M. (2004). Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks* 15, 1063–1070

Jiang, N., Dosen, S., Muller, K.-R., and Farina, D. (2012). Myoelectric control of artificial limbs—is there a need to change focus?[in the spotlight]. *IEEE Signal Processing Magazine* 29, 152–150

Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE* 103, 1449–1477. doi:10.1109/JPROC.2015.2460697

Lee, H.-Y., Hsu, C.-M., Huang, S.-C., Shih, Y.-W., and Luo, C.-H. (2005). Designing low power of sigma delta modulator for biomedical application. *Biomedical Engineering: Applications, Basis and Communications* 17, 181–185

Lichtsteiner, P., Posch, C., and Delbruck, T. (2006). A 128 x 128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers* (IEEE), 2060–2069

Liu, H. and Wang, L. (2018). Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics* 68, 355–367

Liu, S.-C., Delbruck, T., Indiveri, G., Whatley, A., and Douglas, R. (2014). *Event-based neuromorphic systems* (John Wiley & Sons)

Loss, J. F., Cantergi, D., Krumholz, F. M., La Torre, M., and Candotti, C. T. (2012). Evaluating the electromyographical signal during symmetrical load lifting. *Applications of EMG in Clinical and Sports Medicine* , 1

Markovic, M., Dosen, S., Cipriani, C., Popovic, D., and Farina, D. (2014). Stereovision and augmented reality for closed-loop control of grasping in hand prostheses. *Journal of neural engineering* 11, 046001

Markovic, M., Dosen, S., Popovic, D., Graimann, B., and Farina, D. (2015). Sensor fusion and computer vision for context-aware control of a multi degree-of-freedom prosthesis. *Journal of neural engineering* 12, 066022

Meier, K. (2015). A mixed-signal universal neuromorphic computing system. In *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE), 4–6

Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., et al. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345, 668–673

Moons, B., Goetschalckx, K., Van Berckelaer, N., and Verhelst, M. (2017). Minimum energy quantized neural networks. In *2017 51st Asilomar Conference on Signals, Systems, and Computers* (IEEE), 1921–1925

Moradi, S., Qiao, N., Stefanini, F., and Indiveri, G. (2017). A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (dynaps). *IEEE transactions on biomedical circuits and systems* 12, 106–122

Neckar, A., Fok, S., Benjamin, B. V., Stewart, T. C., Oza, N. N., Voelker, A. R., et al. (2018). Braindrop: A mixed-signal neuromorphic architecture with a dynamical systems-based programming model. *Proceedings of the IEEE* 107, 144–164

Neftci, E., Mostafa, H., and Zenke, F. (2019). Surrogate gradient learning in spiking neural networks. *ArXiv* abs/1901.09948

Orchard, G., Jayawant, A., Cohen, G. K., and Thakor, N. (2015). Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience* 9, 437

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in PyTorch. *NeurIPS Autodiff Workshop*

Peng, L., Chen, L., Wu, X., Guo, H., and Chen, G. (2016). Hierarchical complex activity representation and recognition using topic model and classifier level fusion. *IEEE Transactions on Biomedical Engineering* 64, 1369–1379

Phinyomark, A., N Khushaba, R., and Scheme, E. (2018). Feature extraction and selection for myoelectric control based on wearable emg sensors. *Sensors* 18, 1615

Pitti, A., Blanchard, A., Cardinaux, M., and Gaussier, P. (2012). Gain-field modulation mechanism in multimodal networks for spatial perception. In *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*. 297–302. doi:10.1109/HUMANOIDS.2012.6651535

Qiao, N., Mostafa, H., Corradi, F., Osswald, M., Stefanini, F., Sumislawska, D., et al. (2015). A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses. *Frontiers in neuroscience* 9, 141

Rivet, B., Wang, W., Naqvi, S. M., and Chambers, J. A. (2014). Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine* 31, 125–134. doi:10.1109/MSP.2013.2296173

Rueckauer, B., Lungu, I.-A., Hu, Y., Pfeiffer, M., and Liu, S.-C. (2017). Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience* 11, 682

Shivappa, S. T., Trivedi, M. M., and Rao, B. D. (2010). Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE* 98, 1692–1715. doi:10.1109/JPROC.2010.2057231

Shrestha, S. B. and Orchard, G. (2018). SLAYER: Spike layer error reassignment in time. In *Advances in Neural Information Processing Systems 31*, eds. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc.), 1419–1428

Smith, L. and Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life* 11, 13–29. doi:10.1162/1064546053278973

Smith, L. H., Hargrove, L. J., Lock, B. A., and Kuiken, T. A. (2011). Determining the optimal window length for pattern recognition-based myoelectric control: Balancing the competing effects of classification error and controller delay. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 19, 186–192. doi:10.1109/TNSRE.2010.2100828

Tan, A.-H., Subagdja, B., Wang, D., and Meng, L. (2019). Self-organizing neural networks for universal learning and multimodal memory encoding. *Neural Networks* doi:https://doi.org/10.1016/j.neunet.2019.08.020

Thakur, C. S., Molin, J. L., Cauwenberghs, G., Indiveri, G., Kumar, K., Qiao, N., et al. (2018). Large-scale neuromorphic spiking array processors: A quest to mimic the brain. *Frontiers in neuroscience* 12, 891

Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters* 36, 189 – 195. doi:https://doi.org/10.1016/j.patrec.2013.07.003

Yasen, M. and Jusoh, S. (2019). A systematic review on hand gesture recognition techniques, challenges and applications. *PeerJ Computer Science* 5, e218

Zahra, O. and Navarro-Alarcon, D. (2019). A self-organizing network with varying density structure for characterizing sensorimotor transformations in robotic systems. In *Towards Autonomous Robotic Systems*, eds. K. Althoefer, J. Konstantinova, and K. Zhang (Cham: Springer International Publishing), 167–178

Zenke, F. and Ganguli, S. (2018). SuperSpike: Supervised Learning in Multilayer Spiking Neural Networks. *Neural Computation* 30, 1514–1541. doi:10.1162/neco_a_01086

Zhang, Y., Wang, Z., and Du, J. (2019). Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–8. doi:10.1109/IJCNN.2019.8851942

Zhao, D. and Zeng, Y. (2019). Dynamic fusion of convolutional features based on spatial and temporal attention for visual tracking. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–8. doi:10.1109/IJCNN.2019.8852301

FIGURE CAPTIONS

Table 1. Comparison of traditional and neuromorphic systems on the task of gesture recognition for both single sensor and sensor fusion. The results of the accuracy are reported with mean and standard deviation obtained over a 3-fold cross validation.

System	Modality	Accuracy (%)	Energy (uJ)	Inference time (ms)	EDP (uJ * s)
Spiking CNN (Loihi)	EMG	55.7 ± 2.7	173.2 ± 21.2	5.89 ± 0.18	1.0 ± 0.1
	DVS	92.1 ± 1.2	815.3 ± 115.9	6.64 ± 0.14	5.4 ± 0.8
	EMG+DVS	96.0 ± 0.4	1104.5 ± 58.8	7.75 ± 0.07	8.6 ± 0.5
CNN (GPU)	EMG	68.1 ± 2.8	$(25.5 \pm 8.4) \cdot 10^3$	3.8 ± 0.1	97.3 ± 4.4
	APS	92.4 ± 1.6	$(31.7 \pm 7.4) \cdot 10^3$	5.9 ± 0.1	186.9 ± 3.9
	EMG+APS	95.4 ± 1.7	$(32.1 \pm 7.9) \cdot 10^3$	6.9 ± 0.05	221.1 ± 4.1
Spiking MLP (ODIN+MorphIC)	EMG	53.6 ± 1.4	7.42 ± 0.11	23.5 ± 0.35	0.17 ± 0.01
	DVS	85.1 ± 4.1	57.2 ± 6.8	17.3 ± 2.0	1.00 ± 0.24
	EMG+DVS	89.4 ± 3.0	37.4 ± 4.2	19.5 ± 0.3	0.42 ± 0.08
MLP (GPU)	EMG	67.2 ± 3.6	$(23.9 \pm 5.6) \cdot 10^3$	2.8 ± 0.08	67.2 ± 2.9
	APS	84.2 ± 4.3	$(30.2 \pm 7.5) \cdot 10^3$	6.9 ± 0.1	211.3 ± 6.1
	EMG+APS	88.1 ± 4.1	$(32.0 \pm 8.9) \cdot 10^3$	7.9 ± 0.05	253.0 ± 3.9

Table 2. Inference statistics of Loihi models on 200 ms-long samples.

Network	Accuracy %	Core Utilization	Dynamic Power (mW)	Inference Speedup
EMG-Loihi	55.74 ± 2.74	6	29.4±3.6	$(34.01 \pm 1.01)\times$
DVS-Loihi	92.14 ± 1.23	95	109.0±15.5	$(30.14 \pm 0.65)\times$
Fusion-Loihi	96.04 ± 0.48	100	137.2±7.3	$(25.82 \pm 0.24)\times$

Table 3. Inference statistics of ODIN and MorphIC models on 200 ms-long samples.

Network	Accuracy (%)	Chip utilization (%)		Dyn. power (mW)		Processing time (ms)		Inference speedup
		ODIN	MorphIC	ODIN	MorphIC	ODIN	MorphIC	
EMG-ODIN	53.65 ± 1.37	91.8	–	0.315	–	23.5	–	$8.5\times$
DVS-MorphIC	85.17 ± 4.11	–	42.0	–	3.3	–	17.3	$11.6\times$
Fusion-ODIN	89.44 ± 3.02	91.8	41.0	0.315	3.3	19.5	9.5	$10.3\times$

Table 4. Low-power operating conditions of ODIN and MorphIC at minimum supply voltage.

Chip	Supply voltage	E_{SOP}	Max. f_{clk}
ODIN	0.55V	8.4pJ	75MHz
MorphIC	0.8V	30pJ	55MHz

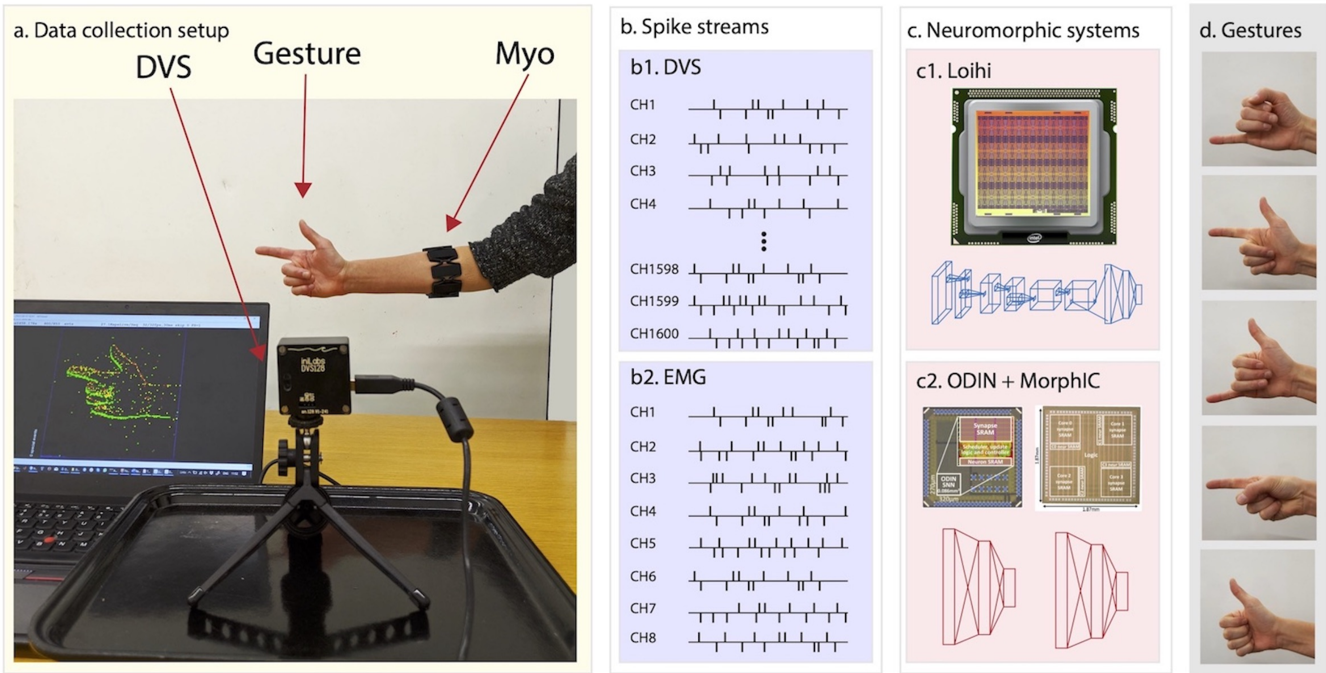

Figure 1. System overview. From left to right: (a) data collection setup featuring the DVS, the traditional camera and the subject wearing the EMG armband sensor, data streams of (b1) DVS and (b2) EMG transformed into spikes via the Sigma Delta modulation approach, the two neuromorphic systems namely (c1) Loihi and (c2) ODIN + MorphIC, (d) the hand gestures that the system is able to recognize in real time.

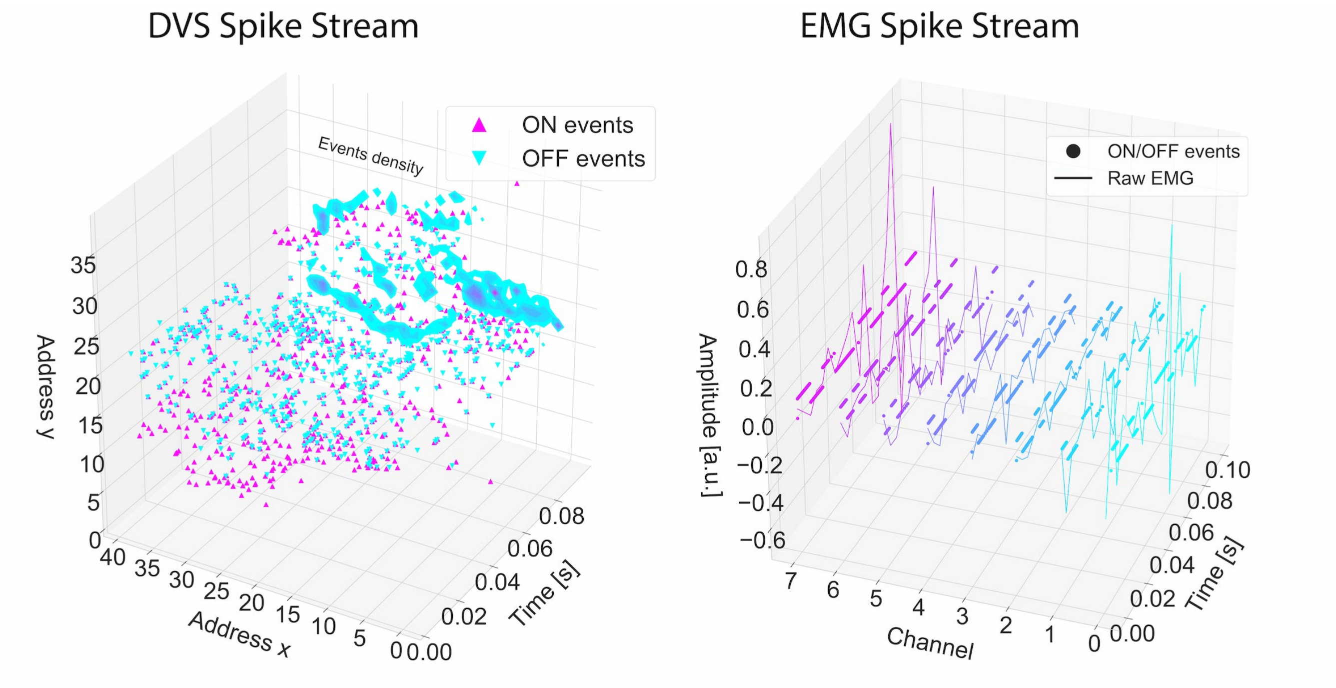

Figure 2. Example, for a gesture 'elle', of spike streams for DVS (left) and EMG (right). In the EMG figure the spikes are represented by dots while the continuous line is the raw EMG. Different channels have different colors.

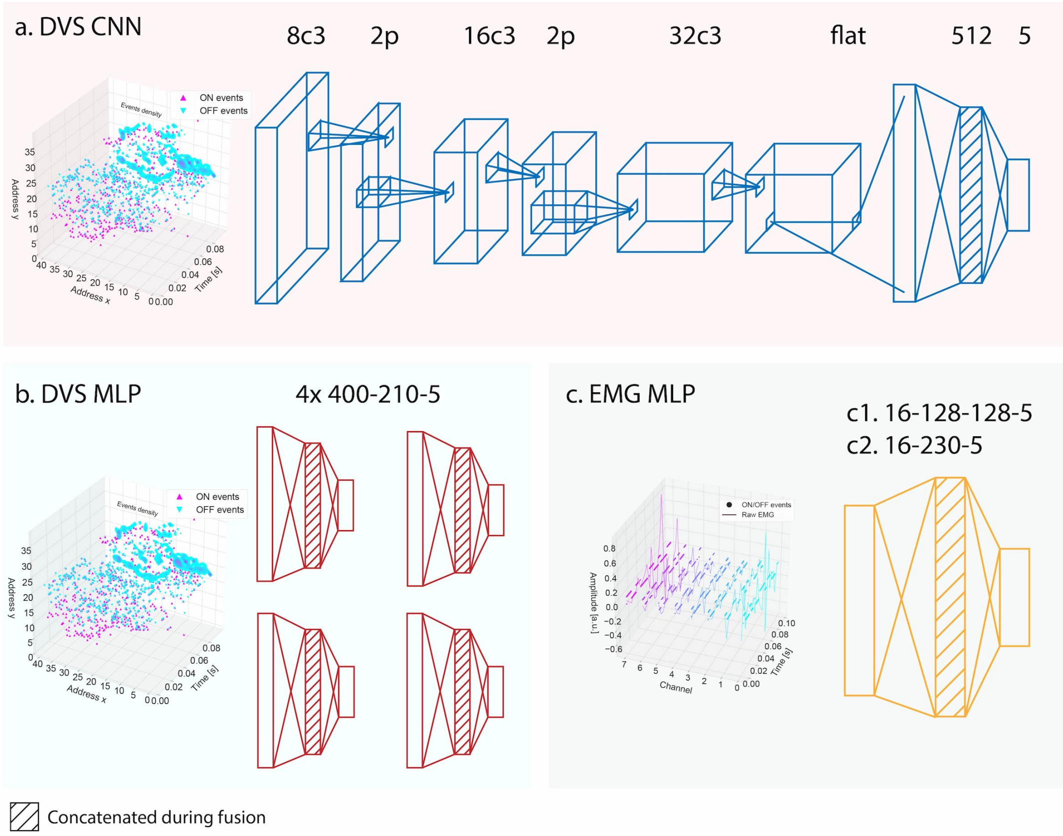

Figure 3. Architectures of the neural networks implemented on the neuromorphic systems and used in the baselines. (a) CNN architecture implemented on Loihi, the corresponding baseline CNN receives APS frames instead of DVS events. (b) subMLP architectures implemented on MorphIC, the corresponding baseline subMLPs receive APS frames instead of DVS events. (c) MLP architecture for the EMG data implemented on Loihi (c1) and on ODIN (c2), the corresponding baseline MLPs receive EMG features instead of EMG events. The shading indicates those layers that are concatenated during the fusion of the networks.

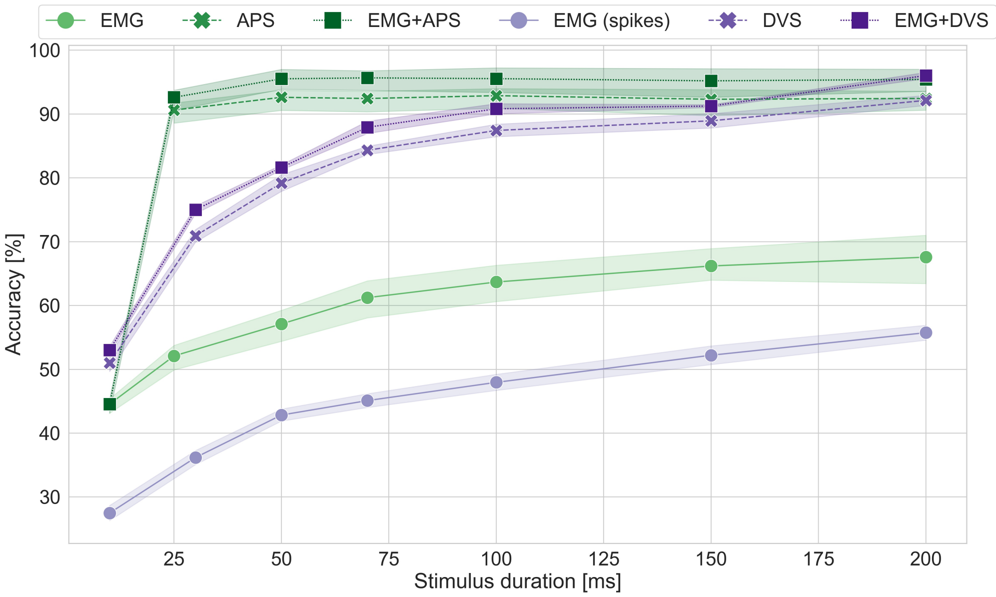

Figure 4. Accuracy vs stimulus duration for the Loihi system and its software baseline counterpart. In green the results for the CNN (GPU), in purple the results for the spiking CNN (Loihi). No classification is present for APS frames before 50 ms since the frame rate is 20 fps.

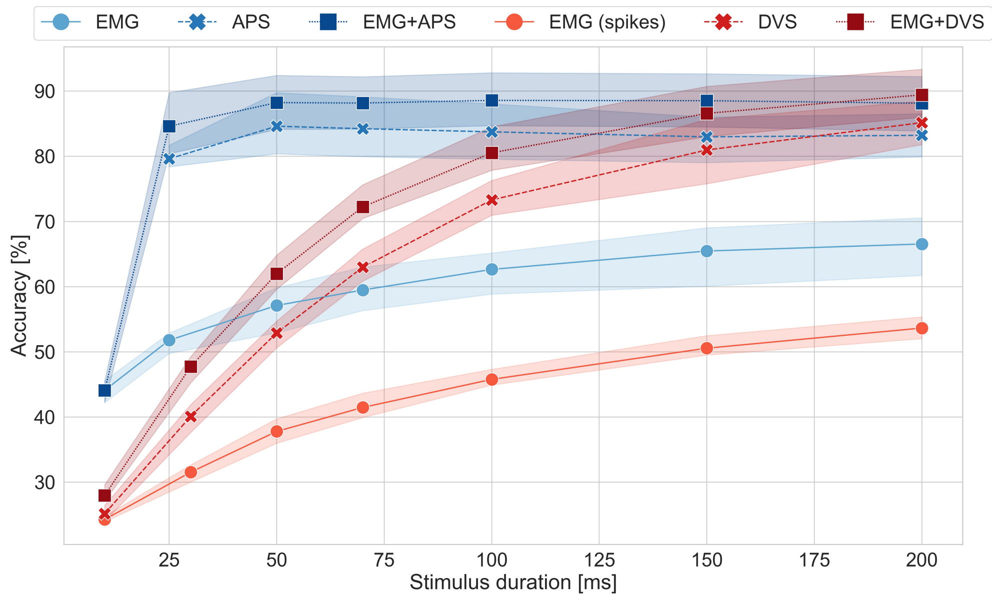

Figure 5. Accuracy vs stimulus duration for the ODIN + MorphIC system and its software baseline counterpart. In blue the results for the MLP (GPU), in red the results for the spiking MLP (ODIN+MorphIC). No classification is present for APS frames before 50 ms since the frame rate is 20 fps.

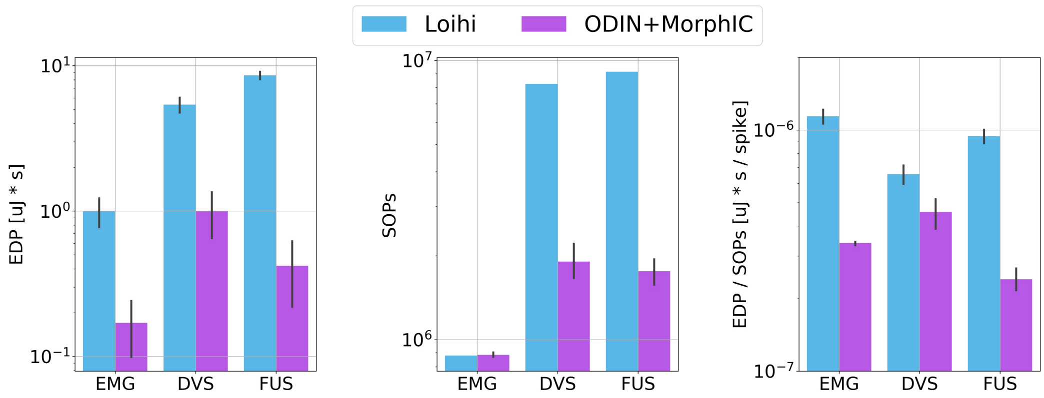

Figure 6. Comparison between the two neuromorphic system with respect to (a) energy delay product (EDP) (see Section 1), (b) number of synaptic operations (SOPs) (see Section 2.3.1), (c) EDP normalized by the number of SOPs.