Spike-based beamforming using pMUT arrays for ultra-low power gesture recognition

E. Hardy, B. Fain, T. Mesquida, F. Blard, F. Gardien, F. Rummens, J.C. Bastien, J.R. Chatroux, S. Martin, V. Rat, et al.

Spike-based Beamforming using pMUT Arrays for Ultra-Low Power Gesture Recognition

E. Hardy[1], B. Fain[1], T. Mesquida[2], F. Blard[1], F. Gardien[1], F. Rummens[2], J.C. Bastien[1], J.R. Chatroux[1], S. Martin[1], V. Rat[1] and E. Vianello[1]
[1]CEA-Leti, Univ. Grenoble Alpes, Grenoble, France, email: emmanuel.hardy@cea.fr
[2]CEA-List, Univ. Grenoble Alpes, Grenoble, France

***Abstract*—**Sensor arrays constrain the power budget of battery-powered smart sensor as the analogue front-end (AFE), analogue-to-digital conversion (ADC) and digital signal processing is duplicated for each channel. By converting and processing the relevant information in the spiking domain, the energy consumption can be reduced by several orders of magnitude. We propose the first end-to-end ultra-low power Gesture Recognition (GR) system comprising an array of emitting and receiving piezoelectric micromachined ultrasonic transducers (pMUT), driving/sensing electronics, a novel spike-based beamforming strategy to extract the distance and angle information from incoming echoes without conventional ADCs and a Spiking Recurrent Neural Network (SRNN) for the GR. We experimentally demonstrate a classification accuracy of 86.0% on a dataset of five 3D gestures collected on our experimental setup.

I. Introduction

Recent developments in large-scale neuromorphic computing ICs such as TrueNorth [1], BrainScales [2], Loihi [3], Spinnaker [4] offer a low power, low latency implementation of neural networks through their spike-based approach. However, when interfaced with sensors for real-time applications, the analogue-to-spike conversion can increase significantly the power budget when using conventional ADC with digital signal processing. Efficient spike-based solutions have been proposed in the literature. In [5], an audio signal is separated in frequency bands, with the energy of each band converted to spikes enabling keyword spotting. In [6], a neuromorphic stereo vision system takes spikes from two event-based sensors and detects temporal coincidence between pairs of pixels.

We propose in this paper an end-to-end real-time ultrasonic GR system with a custom pMUT array that takes a full spike-domain approach including beamforming, feature generation and inference. Several ultrasonic GR systems have been published in the literature [7]–[9], but none has been implemented end-to-end, i.e. without a computer for post-processing and classification. The most integrated solution to date presented in [7] for ranging application is a pMUT array along with an ASIC including the emitter drive, channel readout and a bandpass ADC, with an estimated total power of 15.6 µJ/measurement. Our spiking approach coupled with low power design enables a power consumption several orders of magnitude lower.

II. pMUT Array

The pMUTs have been fabricated within the 8'' MEMS platform of CEA-LETI. Each membrane is an 880 µm-wide bimorph structure, made of two 800 nm–thick AlN layers sandwiched between three 200 nm-thick Mo layers and covered by a 200 nm-thick SiN passivation layer, as reported in [10] (Fig. 1). This pMUT is designed to perform efficiently both for transmission (TX) and reception (RX) operations thanks to a four-electrode pairs scheme (Fig. 2) combined with the inherently good sensing properties of AlN [11]. A die is composed of several membranes located 1.5 mm apart, benefitting from MEMS technology to achieve a half-wavelength spatial pitch (Fig. 3). As an emitter, the pMUT membranes typically exhibit a drive sensitivity of 700 nm/V and a surface pressure of 270 Pa/V at 113.6 kHz. The charges arising at the electrodes of each pMUT receiver are converted to a voltage by dedicated AFE electronics. We measured a sensitivity of 0.6 V/Pa at the receiver, which corresponds to a pMUT short-circuit sensitivity of 15 nA/Pa. The noise floor is typically 60 mPa in the current setup, enabling pulse-echo measurement up to 60 cm. The readout electronics also applies a dedicated DC voltage to each pMUT membrane to induce a stress in the piezoelectric layers and therefore tune each pMUT resonant frequency to 113.6 kHz (Fig. 4).

III. Spike-Based Beamforming

Beamforming is a well-known technique to infer the angle of arrival of a signal by combining the outputs of a sensor array. In the conventional Delay-and-Sum (DaS), the time difference of arrival (TDOA) caused by the angle of incidence (Fig. 5) is compensated by a set of delays, building constructive interferences in the direction of interest. In this work, we propose a novel spike-based DaS beamforming that does not rely on ADCs. A bandpass filter is first applied to the analogue signal after a charge amplifier to remove the out-of-band noise and extract the phase information by using a comparator. We keep only the rising edge for further processing (see Fig. 6). At this stage, each channel is a sparse 1-bit stream S_n with zero or one spike per acoustic signal period.

Fig. 7 shows the principle of the spike-based DaS for two different directions: 0° and α. For each direction, a set of pre-calculated delays is applied on the input channels, delayed spikes are summed and a coherence detector count the number of spikes in a specified time window. Delays have a resolution provided by a 4 MHz clock. In our example, the signal comes

from the direction α. In the direction 0°, spikes are not in phase and coherence signal C_0 stays at zero. In the direction α, delays corresponding to the actual TDOA are applied. The spikes are in phase and the coherence signal C_α outputs spikes.

In our GR prototype, the spike DaS takes five channels and calculates the coherence in eleven directions from -50° to +50° for both horizontal and vertical axes. The coherence window is set to 1.5 µs and the threshold to 4 spikes per window. Fig. 8 (a) and (b) compares the conventional and spike-based beam patterns for 0° and 50° and Fig. 8 (c) shows the different beams in the horizontal plane. One may observe that this technique gives narrower beams, with a stable width across angles. Side lobes are eliminated thanks to the signals temporal sparsity and the detector time and amplitude threshold effect. To reduce the classifier complexity, coherence signals are downsampled to 13 distance bins for each pulse-echo measurement or frame, corresponding to a distance between 4.3 cm and 60 cm. We further reduce the feature dimension from a 11×13 matrix per pMUT array to three vectors of 11, 11 and 13 elements per frame corresponding respectively to the x and y axis directional information and the averaged x and y distance information (see Fig. 9).

IV. Neural Network Classifier

Fig. 10 shows a diagram of the proposed GR system including the classification. We selected a Recurrent Neural Network for its ability to classify in streaming with minimal memory footprint and its adequacy with the temporal nature of gestures. We compare the classification performances of a floating-point Gated Recurrent Unit (GRU) baseline with a quantized SRNN (see Fig. 11). Both networks are trained with a max pooling loss function [12] on the same partitions of gestures and non-gestures. The baseline is a 16-units GRU layer followed by a 6-units dense layer with sigmoid activation: one for the presence of gestures and five for each type of supported gesture.

The SRNN is composed of 110 Leaky Integrate & Fire neurons with 6-bits soma potential and 4-bits weights simulated in temporal steps corresponding to frames as shown in Fig. 12 and 13. The fully digital neurons integrate leak, refractory period and synaptic delays to achieve complex spatiotemporal pattern aggregation [13]. The SRNN activity and power consumption is proportional to the input spike rate, making it a good fit for the spike-based beamforming scheme for which spikes are emitted only when an obstacle is present.

V. Experimental Setup and Gesture Dataset

Fig. 14 shows the end-to-end GR prototype we built to demonstrate our concept. A single TX membrane is actuated with a 30 cycles-long sine wave of 7.5 V peak at 113.6 kHz every 40 ms. The RX pMUT array consists of two lines, vertical and horizontal for 3D sensing, of five membranes each. Fig. 15 shows the TX input voltage and five RX voltages. The inner and outer electrode pairs are used for both TX and RX operations. The AFE and spike-based beamforming are emulated via discrete electronics (Fig. 16) and a FPGA.

We used the GR prototype to collect a dataset of 499 examples from 12 participants split into a training and a test set. Five gestures were performed (Fig. 17): four swipes (Right-Left, Left-Right, Upwards and Downwards) and a Push-Pull at distances between 10 and 50 cm. We also collected diverse non-gesture examples, labelled as 'None', to teach the classifier what constitutes a gesture. We used the symmetry of the system, specifically time reversal, x/y flipping and on-axis mirroring, to augment our training set only and reduce overfitting. Fig. 18 shows two examples of feature maps for Left-Right and Push-Pull gestures.

VI. Results and Discussion

Classification results for the SRNN and the GRU baseline are shown on Fig. 19 and 20. The overall accuracy on the test set is 86.0% for the quantized SRNN compared to 87.8% for the GRU baseline. We explain the inconsistent recognition accuracy across gestures by the small dataset size and the large in-class variability. Table I shows a comparison with state-of-the-art GR systems. While the classification accuracy depends greatly on the number of gestures and the difficulty of the dataset, our system is the only fully embedded solution. We estimated the power consumption if implemented on silicon (65 nm process). The always-on power is 195 µW for the TX pulse, 4.22 µW for the 10 AFE [5] and 4.9 µW for the spike processing (estimated by digital synthesis). When applying the effective duty cycle of 0.66% for TX and 7.25% for RX, the sensing power consumption is 1.95 µW or 78.2 nJ/frame. The estimated power for the digital SRNN is 760 nJ/frame with gestures and 330 nJ/frame without.

VII. Conclusion

We presented the first end-to-end GR solution suitable for ultra-low power implementation on silicon with an estimated total power consumption of 408 nJ/frame for a recognition rate of 86.0%. This is achieved by using low-power sensors with pMUTs, extract and process the minimum information with our novel spike-based beamforming and perform classification in the spike domain with an SRNN.

References

[1] F. Akopyan *et al.,* "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," TCAD 2015 **[2]** J. Schemmel *et al.,* "Live demonstration: A scaled-down version of the BrainScaleS wafer-scale neuromorphic system," ISCAS 2012 **[3]** M. Davies *et al.,* "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," IEEE Micro 2018 **[4]** S. B. Furber *et al.,* "The SpiNNaker Project," Proc. IEEE 2014 **[5]** M. Yang *et al.,* "Nanowatt Acoustic Inference Sensing Exploiting Nonlinear Analog Feature Extraction," JSSC 2021 **[6]** N. Risi *et al.,* "A Spike-Based Neuromorphic Architecture of Stereo Vision," Front. Neurorobotics 2022 **[7]** R. J. Przybyla *et al.,* "3D Ultrasonic Rangefinder on a Chip," JSSC 2015 **[8]** A. Das *et al.,* "Ultrasound based gesture recognition," ICASSP 2017 **[9]** F. Zhou *et al.,* "Efficient High Cross-User Recognition Rate Ultrasonic Hand Gesture Recognition System," Sensors Journal 2020 **[10]** B. Fain *et al.,* "Beamforming with AlN-based bimorph piezoelectric micromachined ultrasonic transducers," SSI 2021 **[11]** S. Akhbari *et al.,* "Bimorph Piezoelectric Micromachined Ultrasonic Transducers," J. Microelectromechanical Syst 2016 **[12]** M. Sun *et al.,* "Max-Pooling Loss Training of Long Short-Term Memory Networks for Small-Footprint Keyword Spotting," SLT 2016 **[13]** F. Sandin *et al.,* "Synaptic Delays for Insect-Inspired Temporal Feature Detection in Dynamic Neuromorphic Processors," Front. Neurosci. 2020

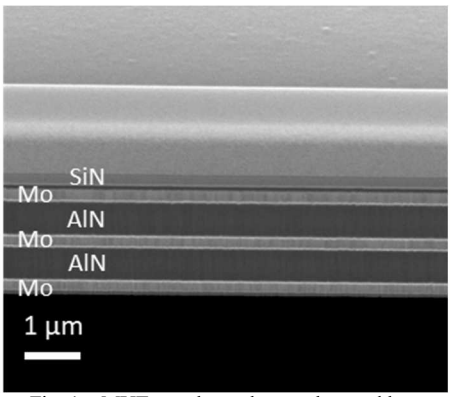

Fig. 1. pMUT membrane layers observed by scanning electron Microscopy (cross-section).

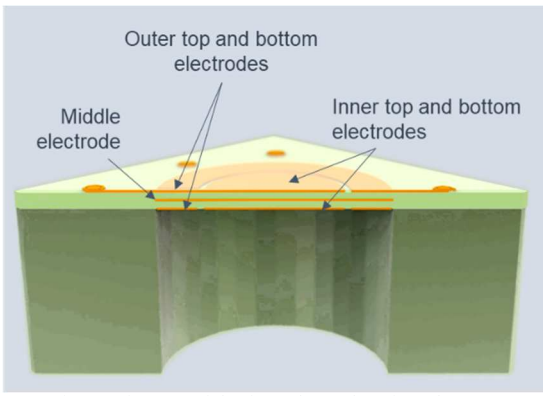

Fig. 2. Diagram of the four-electrode pairs scheme of an AlN-based bimorph pMUT.

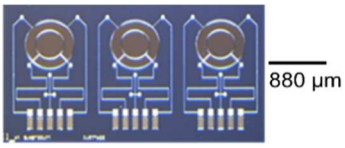

Fig 3. Micrograph of three adjacent pMUT membranes.

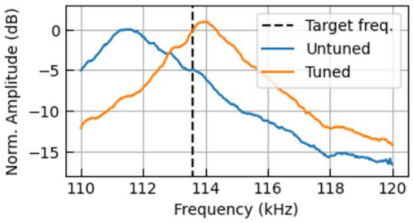

Fig. 4. pMUT resonant frequency tuning by applying a DC voltage. TX acoustic measurement.

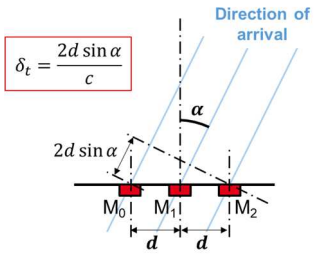

Fig. 5. Time difference of arrival vs direction of arrival α.

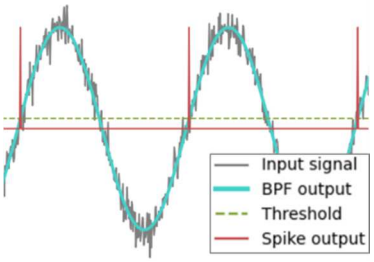

Fig. 6. Analogue-to-spike conversion using bandpass filtering and comparator.

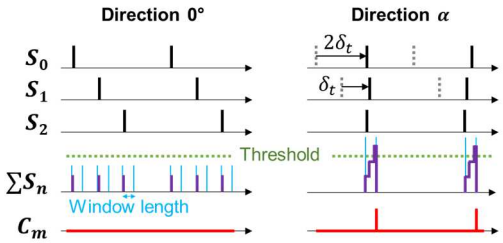

Fig. 7. Spike based Delay-and-Sum beamforming principle with coherence detection. α is the actual angle of incidence.

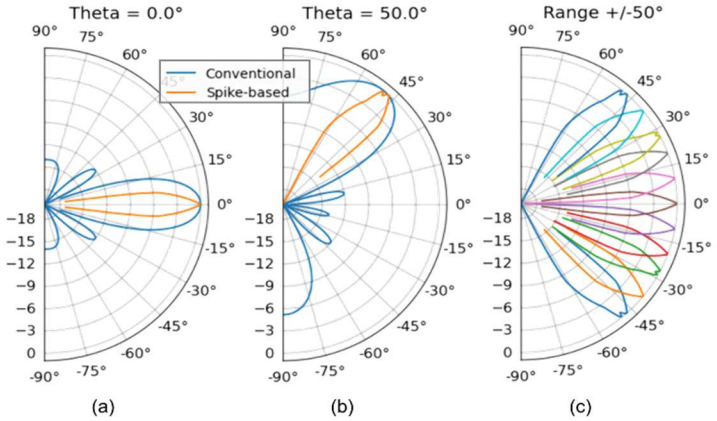

Fig. 8. (a) Spike-based vs conventional beam pattern at 0° and (b) 50°, (c) Eleven beams of the spike-based beamforming on the +/-50° range.

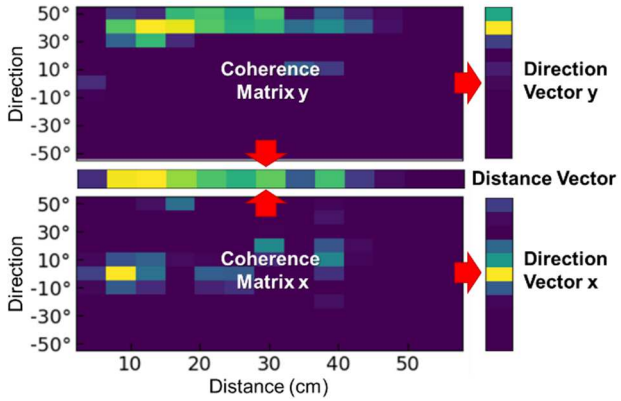

Fig. 9. Feature dimension reduction of the coherence detector output from axes x and y to three vectors: distance, direction x and direction y.

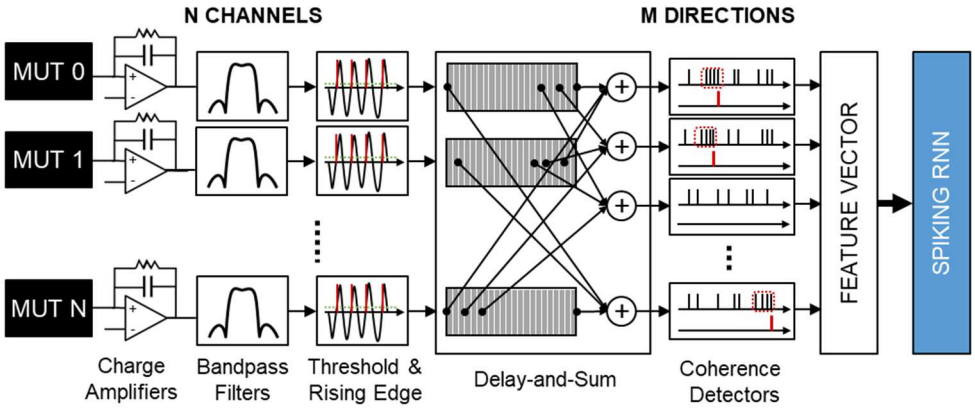

Fig. 10. Block diagram of the GR system and specifically the spike-based beamforming.

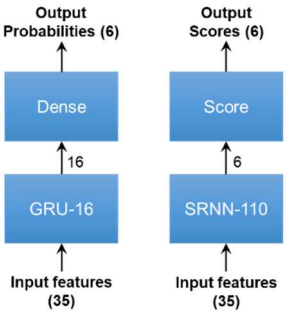

Fig. 11. GRU vs SRNN topology.

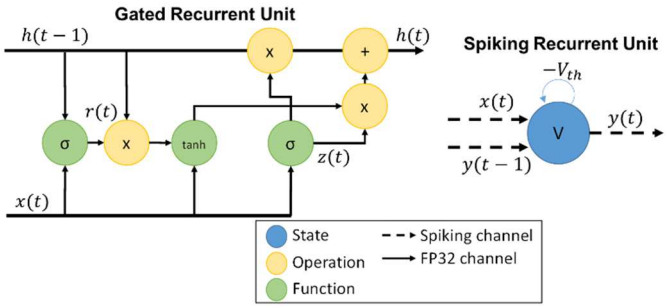

Fig. 12. GRU vs SRNN Unit. The SRNN unit membrane potential is updated at each time step as shown on Fig. 13.

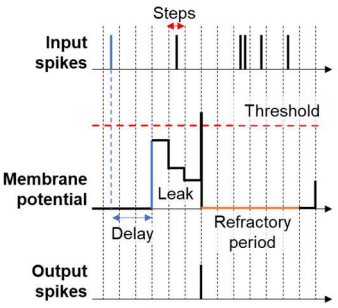

Fig. 13. SRNN unit temporal update mechanisms.

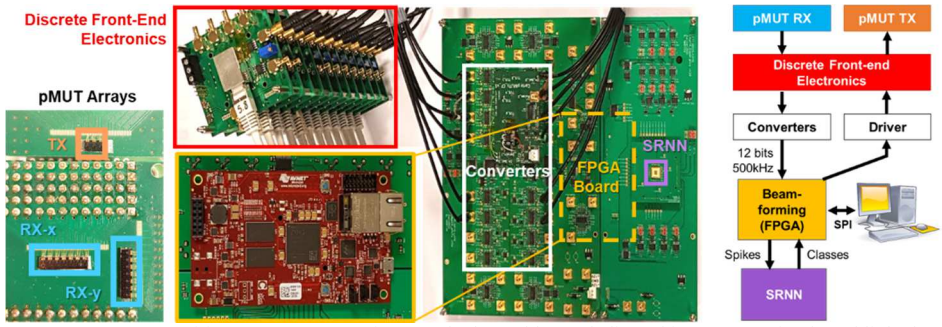

Fig. 14. Gesture Recognition prototype overview. Each channel has a dedicated instrumentation board linked to an ADC. Filtering and Spike-Based Beamforming is emulated in the FPGA and interfaced to the SRNN.

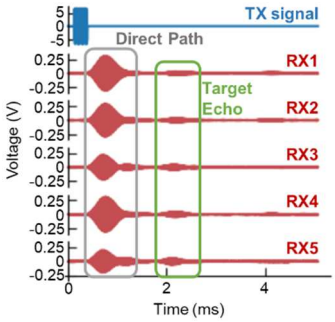

Fig. 15. TX signal and measured RX acoustic response on the x axis.

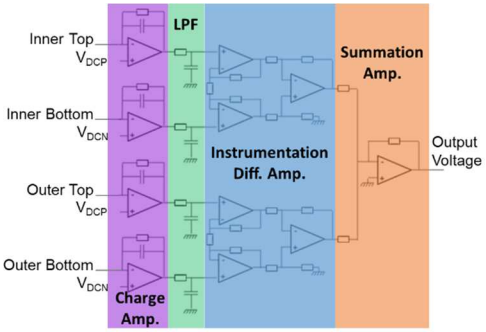

Fig. 16. GR Prototype Instrumentation electronics. V_{DCP} and V_{DCn} are resp. the positive and negative tuning DC voltages.

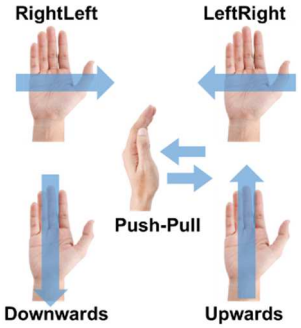

Fig. 17. Supported types of gestures.

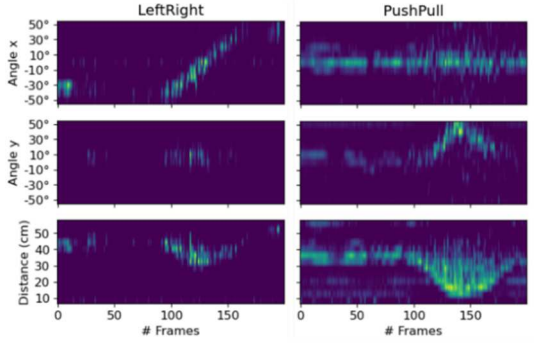

Fig. 18. Examples of gestures: LeftRight and PushPull. Negative angles correspond to Left and Bottom for resp. x and y axes.

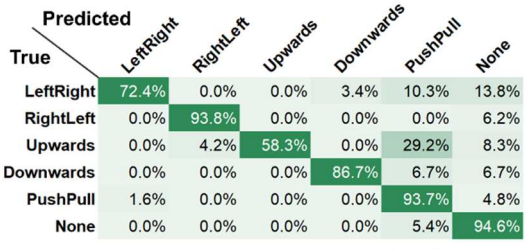

Fig. 19. Confusion matrix on the SRNN classification.

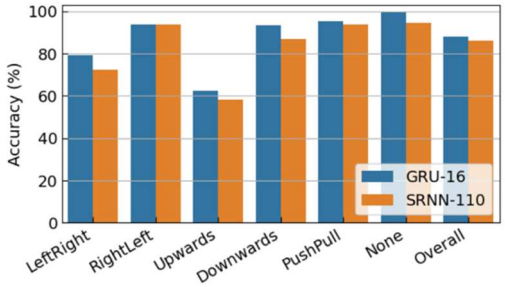

Fig. 20. Classification precision between the GRU baseline and the SRNN.

	This work	[7]	[8]	[9]
Type of transducers	TX-RX: AlN pMUT	TX-RX: AlN pMUT	TX: COTS RX: MEMS	TX-RX: COTS
$f_{acoustic}$	113.6 kHz	217 kHz	40 kHz	40 kHz
Method/ beamforming	Time of Flight/ Spike DaS	Time of Flight/ DaS	Time of Flight/ MVDR	Doppler/ N/A
# RX - Pattern	10 - 2 Lines X/Y	7 - Zigzag	8 - Square	1
Classif. type	SRNN	N/A	CNN-LSTM	Random Forest
Classifier performance	86.0% (5 gest.)	N/A	64.5% (5 gest.)	93.9 % (7 gest.)
Meas. period	40 ms	5.9 ms	20 ms	1.5 ms
Max. range	60 cm	100 cm	100 cm	4 cm
Hardware integration	COTS	ASIC+ Post-process	COTS+ Post-process	COTS+ Post-process
Estimated sensing energy (ASIC)	78.1 nJ/meas.	15.6 µJ/meas.	Not measured	Not measured
Estimated inference energy (ASIC)	330/760 nJ/meas. (None/Gesture)	N/A	Not measured	Not measured

Table I. Comparison of the proposed system with state-of-the-art Ranging and GR systems.