nature communications

Article

https://doi.org/10.1038/s41467-025-56899-3

A bioinspired in-materia analog photoelectronic reservoir computing for human action processing

Received: 22 July 2024

Accepted: 5 February 2025

Published online: 06 March 2025

Check for updates

Hangyuan Cui[1], Yu Xiao[2], Yang Yang[1], Mengjiao Pei[1], Shuo Ke[1], Xiao Fang[1], Lesheng Qiao[1], Kailu Shi[1], Haotian Long[1], Weigao Xu[3], Pingqiang Cai[4], Peng Lin[2] ✉, Yi Shi[1] ✉, Qing Wan[5] ✉ & Changjin Wan[1,5] ✉

Current computer vision is data-intensive and faces bottlenecks in shrinking computational costs. Incorporating physics into a bioinspired visual system is promising to offer unprecedented energy efficiency, while the mismatch between physical dynamics and bioinspired algorithms makes the processing of real-world samples rather challenging. Here, we report a bioinspired in-materia analogue photoelectronic reservoir computing for dynamic vision processing. Such system is built based on InGaZnO photoelectronic synaptic transistors as the reservoir and a TaO_X-based memristor array as the output layer. A receptive field inspired encoding scheme is implemented, simplifying the feature extraction process. High recognition accuracies (>90%) on four motion recognition datasets are achieved based on such system. Furthermore, falling behaviors recognition is also verified by our system with low energy consumption for processing per action (~45.78 μJ) which outperforms most previous reports on human action processing. Our results are of profound potential for advancing computer vision based on neuromorphic electronics.

Computer vision is important for a wide spectrum of applications, including video retrieval, human-robot interaction, and entertainment, while it is always involved with accessing of a great amount of data in the event streams. Computer vision algorithms run on a von Neumann computing system is always energy-costly due to the efficiency loss with the increase of data scale[1–3]. On the contrary, our human brain has an unerring instinct for processing huge amounts of information in an energy-saving way. The energy efficiency of the biological visual system may rise from the filtering capability and the spike encoding scheme of the biological visual system. The former is achieved through the hierarchical structure and the receptive fields in each layer, which allows only a fraction of incoming visual events to be perceived, remembered or acted on. The latter enables event-driven computing, which consumes energy only when the population of cells are elicited by a sensory event. In this way, we can classify, locate, detect, and segment targets in video inputs that are transmitted by retina at roughly 10 million bits per second (10 $Mb{\cdot}s^{-1}$) with high accuracy and efficiency[4–6].

The neuromorphic visual system (NVS), which emerged recently, is aimed at extending the efficiency and capability of computer vision by replicating biological superiorities from the bottom up[7–10]. For example, optoelectronic graded neurons based on MoS_2 optoelectronic transistors were proposed for in-sensor motion perception, in which motion speeds can be effectively perceived with gate voltage modulation[11]. A van der Waals (vdW) heterostructure array was fabricated with the recognition capability to classify the motion modes

[1]School of Electronic Science and Engineering, National Laboratory of Solid-State Microstructures, Nanjing University, Nanjing 210023, P. R. China. [2]College of Computer Science and Technology, State Key Laboratory of Brain Machine Intelligence, Zhejiang University, Hangzhou 310058, China. [3]Key Laboratory of Mesoscopic Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, China. [4]Jiangsu Key Laboratory of Molecular Medicine, Medical School, Nanjing University, Nanjing 210023, P. R. China. [5]Yongjiang Laboratory (Y-LAB), Ningbo 315202, China.
✉e-mail: penglin@zju.edu.cn; yshi@nju.edu.cn; wanqing@nju.edu.cn; cjwan@nju.edu.cn

(e.g., direction and velocity) of a target, via its gate-tunable electronic and optoelectronic properties[12]. An oxide-based retinomorphic photomemristor-reservoir computing system was proposed for motion modes recognition and prediction[13]. More recently, a fully memristive elementary motion detector has been proposed, achieving a high accuracy and low computational cost in lane-changing maneuvers prediction[14]. However, the identification of the motion and verification on real-world datasets have seldom been reported by previous reports, which greatly limit the applications in real scenarios. This might be due to the complexity of data preprocessing and the difficulty in tailoring the properties of an NVS device to match the requirements of appropriate algorithms.

In this work, we demonstrate a bioinspired in-materia analog photoelectronic reservoir computing (Alpho-RC) system based on indium-gallium-zinc-oxide (IGZO) photoelectronic synaptic transistors as the reservoir and a TaO_X-based memristor array as the output layer for human action processing. The human actions are captured and represented by 3D human skeleton sequences. Such sequences are then encoded into spike trains by several bioinspired Gaussian receptive field (GRF) neurons, and no feature extraction process on skeleton sequence is required. The spike train from each GRF neuron is applied to an IGZO-based photoelectronic synaptic transistor. Such transistor exhibits gate voltage tunable shading memory and nonlinear properties based on its photoelectronic coupling dynamics. In this case, it can effectively map the population encoded spike trains into high-dimension space and can provide abundant states as the virtual nodes for reservoir computing. Human action recognition tasks are stimulated with high accuracies (>90%) based on four standard datasets including UTD-MHAD, MSR Action3D, MSR Action Pairs, and Florence 3D. Furthermore, a one-transistor-one-memristor (1T1R) array used as the readout layer is integrated with the photoelectronic reservoir to construct the Alpho-RC system. Identification of falling behaviors is achieved based on such a system, and an energy consumption of only ~45.78 μJ/action is achieved, at least two orders of magnitude lower than digital processors. Our work can be regarded as a platform for next-generation neuromorphic computing which would prosper the development of high energy-efficient virtual reality, medical care, and visual surveillance.

Results

An analog photoelectronic reservoir computing system

Figure 1a shows the dynamic vision processing in the human visual system. Visual inputs are initially detected by the retina located at the bottom of the eyeballs, which would trigger the action potential by ganglion cells through the transmission of bipolar cells. The encoded information is then transmitted to the thalamus[15–17]. Cells in thalamus with receptive fields are layered and in the form of populations, enabling the feature extraction of visual information. Features like edge, angle, orientation, and direction of motion can be extracted before entering the visual cortex. Through spike precoding by multi-level neuronal populations with receptive fields within the retina, intellectual activities such as recognition and decision-making can be finally realized in the human brain. The powerful and energy-efficient human visual system for dynamic vision processing has inspired us to conceptualize a neuromorphic visual system featuring receptive fields and a spike encoding scheme. Hence, a bioinspired in-materia analog photoelectronic reservoir computing (Alpho-RC) system was built for dynamic vision processing.

Figure 1b shows the schematic diagram of such Alpho-RC system. The data modes of human action are mainly divided into visual modes (RGB, Depth, Skeleton and so on) and non-visual modes (Inertial acceleration data, Wireless-transmission signal and so on)[18–22]. Among them, the visual mode is more in line with human intuitive feelings. Compared with other visual modalities, a 3D skeleton frame as a topological form of joints and skeletons abstracts the human body into a three-dimensional coordinate space, which tends to be more robust in complex environments. Therefore, 3D skeleton-based human action frames are selected as visual input in the system.

The core elements of spike encoding scheme are neurons with GRF without additional filtering process. In biology, stimuli carried essential features are selectively responded by neurons with overlapped and graded receptive fields, and the varied levels of features thus can be extracted in different neural pathways[23]. Stimulus can significantly fire at the neurons with proper receptive field, and might be silence in other neurons. Such biological encoding mechanisms enable only a small size of data being processed by the central nervous system. In contrast, running machine learning algorithms in a digit system always requires complex feature extraction steps, which lead to a huge computational burden[24]. In Alpho-RC system, several GRF neurons are defined as one population encoder, which is corresponding to population transistor consisting of electric-double-layer (EDL) coupled IGZO photoelectronic transistors. EDL coupled IGZO transistors have been demonstrated with photoelectronic synaptic plasticity by gate voltage tunability based on persistent photoconductivity (PPC) effects and proton relaxation dynamics[25–27]. Figure 1c shows the photograph of photoelectronic transistor array and schematic diagram of IGZO transistor structure. Details on transistor array fabrication, basic properties and hardware operation system can be found in "Methods" section and Supplementary Figs. 1–9. The response and relaxation behaviors of the EDL-coupled IGZO photoelectronic transistor in response to a light pulse (38 ms, 2.95 nW·μm^{-2}) under different gate bias conditions (−0.4, −0.2, and 0 V, respectively) has shown in Fig. 1d. The electrons generated by PPC effect would increase the channel conductance transiently. When a light stimulation ends, accumulated electrons would recombine gradually with the triggered oxygen vacancies within a certain time. As a consequence, the channel conductance gradually decreases, exhibiting relaxation characteristics. A negative gate voltage would decrease the channel conductance through the EDL coupling, leading to a decreased decay time. The decrement is positively related to the absolute value of the negative gate voltage (detailed information can be found in Supplementary Note 1). By applying varied gate biases that are corresponded to GRF neurons with different distribution centers, current states of the transistor are tuned in response to the encoding pulses from GRF neurons, mapping the visual input information to the high-dimensional space in turn.

Physical reservoir computing (PRC) based on material intrinsic dynamics has been demonstrated with excellent time signal processing capabilities, which is selected as the calculation method in Alpho-RC system[28–30]. Obtained response currents by population transistors are used as reservoir states for training output weights. Compared with the processing path by single device in traditional PRC systems, such receptive field-enhanced population encoding mechanism increases the information processing capacity by providing multiple parallel information processing ports. Meanwhile, no additional feature extraction is required to apply on skeleton sequences, through such bioinspired in-materia reservoir computing framework. However, the feature extraction is still required for skeleton frames in vast number of reported algorithms. For example, while facing the skeleton-based action recognition tasks, feature set including spatial-domain-feature (relative position, distances between joints, distances between joints and lines) and temporal-domain-feature (joint distances map, joint trajectories map) needs to be extracted prior to the training of neural networks[31].

The output layer was also implemented in a 32 × 32 1T1R array for fully hardware implementation of the bioinspired in-materia reservoir computing system, as shown in Fig. 1e (the detailed device structure and fabrication can be found in Supplementary Fig. 10 and "Methods" section). TaO_X-based memristors were integrated on top of a foundry-made complementary-metal-oxide-semiconductor (CMOS) chip with

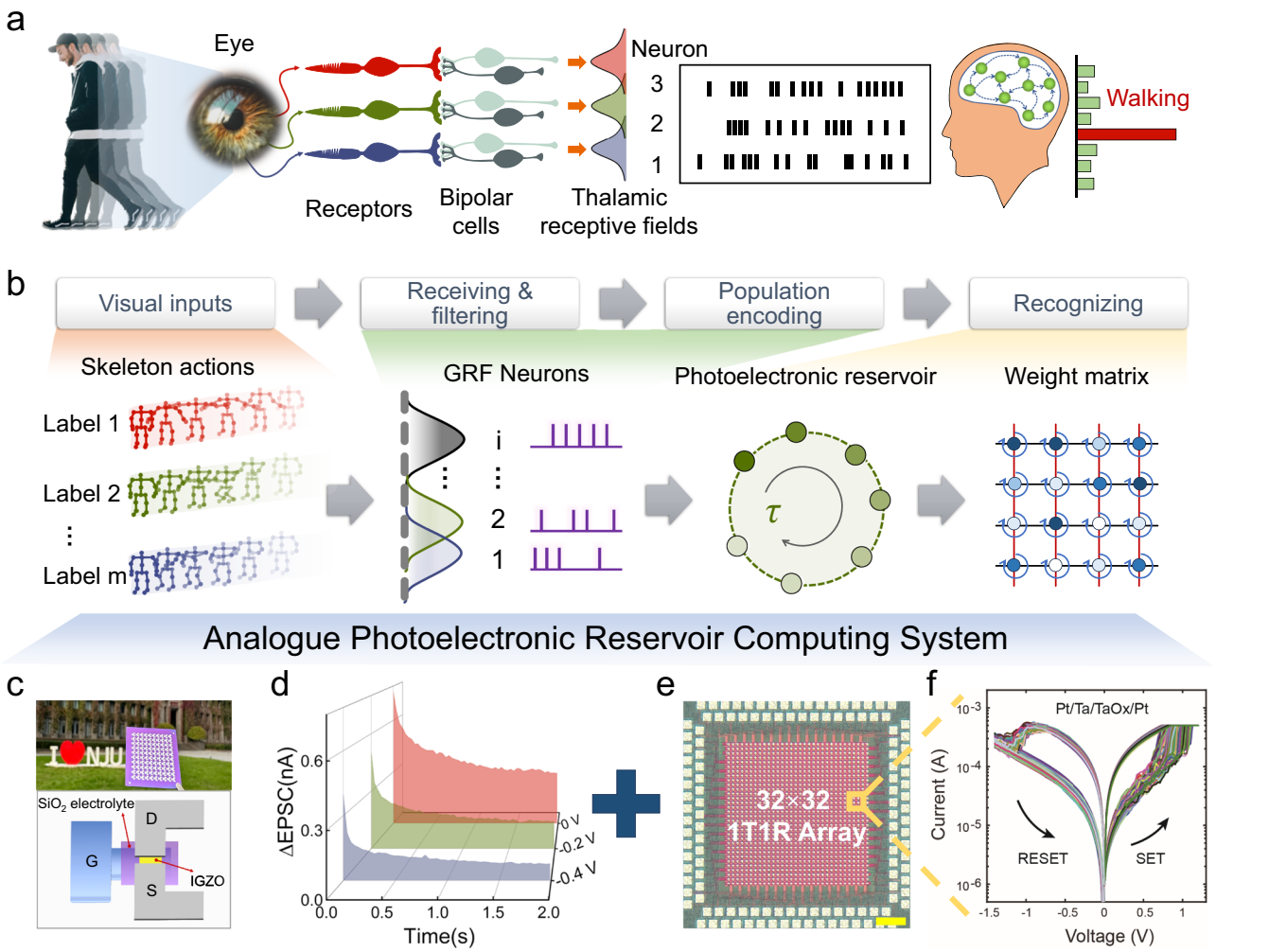

Fig. 1 | A bioinspired in-materia analog photoelectronic reservoir computing (Alpho-RC) system. a Conceptual diagram of dynamic vision processing in human visual system. The walking man and eye are reproduced with permission from Pixabay. **b** Schematic diagram of calculation processes in Alpho-RC system. **c** Photograph of EDL-coupled IGZO photoelectronic transistor array and schematic diagram of IGZO transistor structure. **d** The response and relaxation behaviors of EDL-coupled IGZO photoelectronic transistor in response to a light pulse (38 ms, 2.95 nW·μm^{-2}) under different gate bias conditions (−0.4, −0.2, and 0 V, respectively). **e** A micrograph image of a 32 × 32 1T1R array (scale bar: 200 μm). **f** 100 cycles of I-V sweeps of a Pt/Ta/TaOx/Pt memristor in the array.

silicon-based selecting transistors and fan-outs (the detailed characterizations and the setup of hardware operation system can be found in Supplementary Figs. 11–14). The memristors demonstrated good switching uniformity and stability for reliable programming of offline-trained weights for human action processing (Fig. 1f). Our analog photoelectronic reservoir computing system is built based on the IGZO transistors for reservoir states collection and the TaOx-based memristors for matrix operations.

The bioinspired in-materia reservoir computing framework

The dimensionality enhancement process of the photoelectronic reservoir in Alpho-RC system contains time multiplexing and GRF encoding. Time multiplexing with mask matrixes is a common method in PRC systems, which solves the difficulty of interconnecting physical device nodes based on the concept of delayed nodes[30]. Figure 2a shows the calculation flow of time multiplexing on original action skeleton information. Firstly, K body joint points with coordinate information ($\mathbf{x}_i$, $\mathbf{y}_i$, $\mathbf{z}_i$) are collected from a subject, which maps the human body's action posture into a three-dimensional coordinate space. Therefore, temporal changes of the K joint coordinate positions from frame-to-frame can describe the dynamic change of an action. 3D joint coordinates of each frame are written as one-dimensional sequence vector $[\mathbf{x_1}\ \mathbf{y_1}\ \mathbf{z_1}\ \mathbf{x_2}\ \mathbf{y_2}\ \mathbf{z_2}\ \ldots\ldots\ \mathbf{x_K}\ \mathbf{y_K}\ \mathbf{z_K}]$ or $[\mathbf{J_1}\ \mathbf{J_2}\ \ldots\ldots\ \mathbf{J_K}]$ according to the order of the skeleton joints. In this case, the dynamic coordinates of a complete the action (Act, with a size of $3{\cdot}K{\times}N$) can be represented by:

$$Act = [\boldsymbol{J_{11}J_{12}}\ldots\boldsymbol{J_{1k}};\ldots\ldots;\boldsymbol{J_{n1}J_{n2}}\ldots\boldsymbol{J_{nk}};\ldots\ldots;\boldsymbol{J_{N1}J_{N2}}\ldots\boldsymbol{J_{Nk}}] \quad (1)$$

where n ($n = 1, 2, \ldots, N$) is the frame order of action and N is the total number of frames consisting of an action. The dynamic change of the 3D joint coordinates (ΔAct, with a size of $3{\cdot}K \times N$) can be represented by the subtraction of the action matrix and the first sequence of the action:

$$\Delta Act = [\mathbf{0}\ \mathbf{0}\ldots\mathbf{0};\ldots\ldots;\boldsymbol{\Delta J_{n1}\Delta J_{n2}}\ldots\boldsymbol{\Delta J_{nk}};\ldots\ldots;\boldsymbol{\Delta J_{N1}\Delta J_{N2}}\ldots\boldsymbol{\Delta J_{Nk}}] \quad (2)$$

where ΔJ_{nj} is the change of coordinates (ΔJ_{nj}=J_{nj}-J_{1j}), *and* j is the number of joint ($j = 1, 2, \ldots, K$). Figure 2b shows examples of ΔAct_n ($\Delta Act_n = [\boldsymbol{\Delta J_{n1}\Delta J_{n2}}\ldots\ \boldsymbol{\Delta J_{nK}}]$) corresponding to the action "openarm" shown in Fig. 2a. After that, ΔAct is mapped into high dimensional space by multiplying mask matrixes. The mask matrix is a randomly generated two-dimensional matrix consisting of 1 and -1, with a size of $3{\cdot}K \times M$ (M is the mask length). At nth frame of the action, the ΔAct_n ($1 \times 3{\cdot}K$) is multiplied by the mask matrix, generating virtual nodes

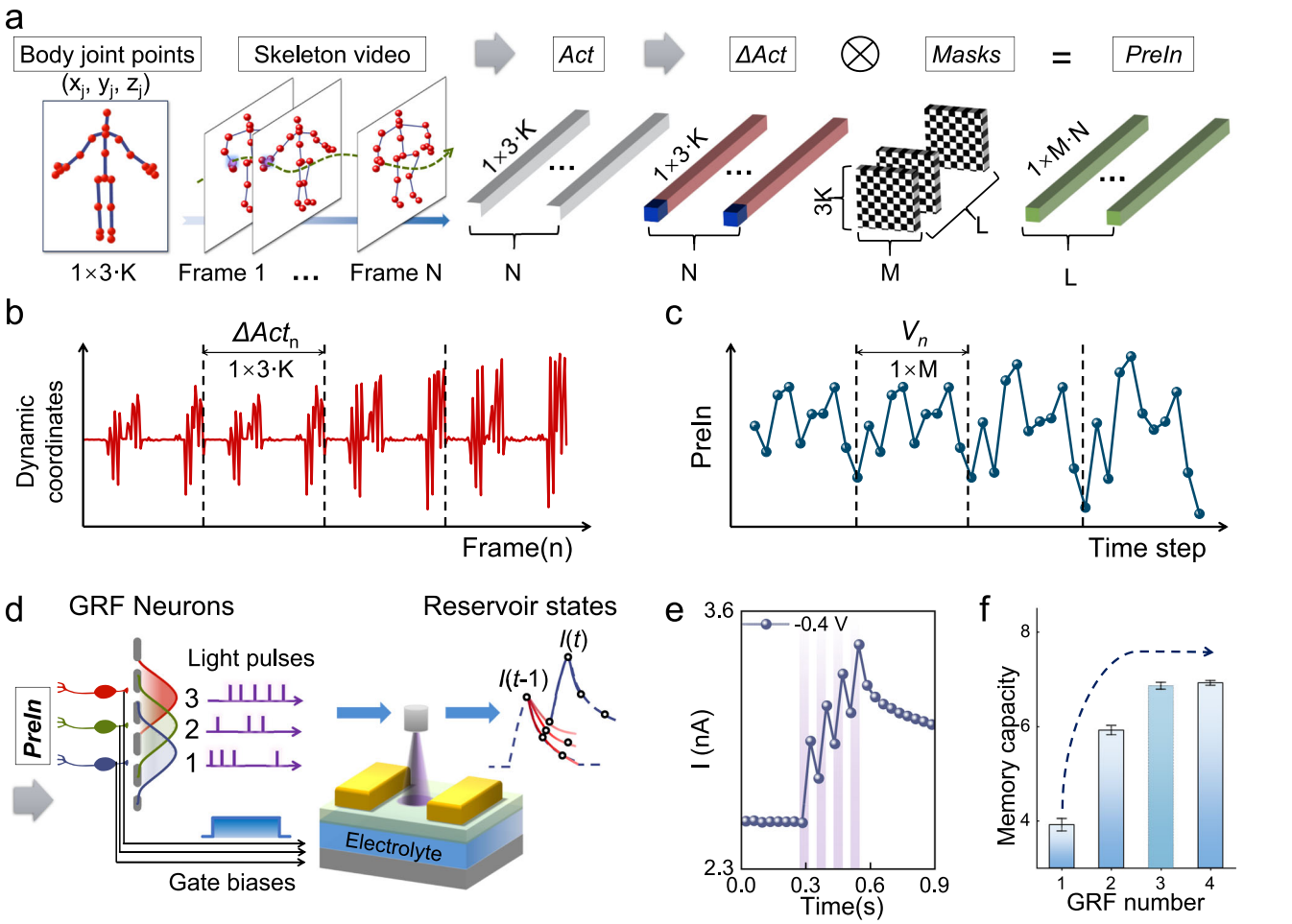

Fig. 2 | The Gaussian receptive field based encoding and photoelectronic reservoir computing. a The calculation flow of time multiplexing on original action skeleton information in photoelectronic reservoir computing. **b** Examples of ΔAct_n corresponding to the action "openarm" in (**a**). **c** An example of PreIn corresponding to ΔAct_n in (**b**) with a mask matrix ($M = 10$). **d** The converted light pulse trains from spike trains of Neuron #1, #2, and #3 and are applied on one population transistor consists of three IGZO-based photoelectronic synaptic transistors with different modulated biases. **e** The response current of transistor with gate bias of −0.4 V to four consecutive light pulses (38 ms ON and 38 ms OFF). **f** MC values with different numbers (1, 2, 3, and 4) of Gaussian receptive fields. The error bars represent the standard deviation of five independent tests.

($V_n = [\mathbf{V_{n1}}\ \mathbf{V_{n2}}\ \ldots\ \mathbf{V_{nM}}]$) with a size of $1{\times}M$. In this case, the ΔAct is mapped to a higher dimensional matrix noted as PreIn matrix ($1{\times}M{\cdot}N$):

$$PreIn = [\mathbf{V_1}\mathbf{V_2}\ldots\mathbf{V_N}] \tag{3}$$

Figure 2c shows one example of PreIn corresponding to ΔAct_n in Fig. 2b with a mask matrix ($M = 10$).

Then, the PreIn corresponding to a completed action is encoded to spikes through GRF neurons. The Gaussian receptive fields of the Alpho-RC system are set based on the intensity of PreIn, which is inspired from encoding process of biological systems and spiking neural networks (SNN)[32]. For every input, the output through i_{th} receptive field yields the Gaussian distribution (G_i) with a mean value of μ_i and deviation of σ_i. So, the output ($G_i(A)$) is dependent on the distance between input intensity (A) and μ_i:

$$G_i(A) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(A-\mu_i)^2}{2\sigma_i^2}\right) \tag{4}$$

In this case, each input can trigger an output through all receptive fields, while the outputs are different. The GRF neuron would trigger a firing spike only if the output from the corresponding receptive field is the largest. Here, several neurons (Neuron #1, #2, ... #i) corresponding to receptive fields (GRF #1, #2, ... #i) are defined as one population encoder which is used for spike encoding. The σ of neurons is uniformly set to the standard value of 1, and the centers ($\mu_1,\mu_2, \ldots\mu_i$) of GRFs are set by:

$$\mu_i = A_{\min} + \frac{A_{\max} - A_{\min}}{m+1} \times i \tag{5}$$

where $A_{\min}$ and $A_{\max}$ represent the minimum value and the maximum value of input intensity (A), m is the number of GRF neurons. To facilitate data processing, we perform a normalization process, so that the amplitude of PreIn is limited between -1 and 1. Accordingly, Gaussian centers are set by:

$$\mu_i = \frac{2}{m+1} \times i - 1 \tag{6}$$

The normalized PreIn of a completed action is then converted into spike trains (Spike-input) through the aforementioned GRF-based population encoding scheme. The spike trains from Neuron #1, #2, ... #i are converted into light pulse trains (405 nm, 2.95 nW·μm^{-2}) and

applied on one population transistor consisting of EDL-coupled IGZO photoelectronic transistors under different gate bias conditions (as shown in Fig. 2d). Each light pulse corresponding to firing spike is set to 38 ms ON, and stationary state without firing spike is set to 38 ms OFF. Thus, the time step between each node of PreIn is fixed, and the triggered channel currents of population transistor at each time step are collected as reservoir states. Relaxation characteristic curves are used as the activation function. As mentioned before, the characteristics are varied in transistors with different gate biases, which increase the richness of reservoir states. Figure 2e shows the response current of transistor with gate bias of -0.4 V to four consecutive light pulses (38 ms ON and 38 ms OFF). Supplementary Figs. 15, 16 show the device responses under different gate bias voltages to light pulse trains of different frequencies. Specific model fitting and calculation details can be found in Supplementary Fig. 17, Fig. 18, Table 1 and Note 2. L population encoders are connected in parallel to build a large parallel RC system. Each population encoder is responsible for encoding different PreIn, which is generated with L different mask matrixes parallelly. Thus, L population transistors with $3\times L$ devices in arrays are used to obtain parallel reservoir states and $3\times L\times M$ reservoir states are recorded per frame. Supplementary Video I shows the aforementioned spike encoding and reservoir states collection processes.

Finally, the reservoir states are collected for training and testing. During the training process, only the output weights (W_{out}) connected to the output layer are needed to be trained. The collected reservoir states are subjected to a one-step linear regression along with labels to obtain the desired weights[33]. The resulting output weights are multiplied with reservoir states collected from the testing process, and the output label is obtained based on the winner-take-all method[34]. Supplementary Fig. 19 shows the specific procedures, and detailed calculation procedures can be found in "Methods".

Memory capacity (MC) is a task-independent evaluation index of reservoir computing, which represents the ability of reservoir states to preserve input information previously[29,35]. We evaluated the MC metrics of the bioinspired reservoir with different numbers (1, 2, 3, and 4) of GRF neurons. During encoding, the Gaussian centers were also sequentially set according to Eq. (6). Population encoders with different GRF neuron numbers (1, 2, 3, and 4) are corresponding to population transistor consisting of different amounts of EDL-coupled IGZO photoelectronic transistors with fixed V_{GS} of 0 V, −0.2 V/0 V, −0.4 V/−0.2 V/0 V, −0.8 V/−0.4 V/−0.2 V/0 V, respectively. Specific calculation process of MC can be found in Supplementary Note. 3. As shown in Fig. 2f, high MC value (~6.9) is obtained by three receptive fields with filtering and encoding. This should be due to the enriched reservoir states rendered by multiple GRF neurons. When encoding number is more than three, the MC value is almost remained. However, the greater number of GRF neurons, the larger the data scale is required, which would increase the computational burden on the following steps. Therefore, three GRF neurons are fixed as one population encoder and one population transistor is fixed of three EDL-coupled IGZO photoelectronic transistors with fixed V_{GS} of −0.4 V, −0.2 V, 0 V, respectively in the bioinspired reservoir.

Human action recognition tasks based on standard datasets

The UTD-MHAD dataset contains 27 classes of human actions collected from 8 subjects[36]. Each skeleton frame consists of dynamic coordinates of 20 body joints. Figure 3a shows examples of color images and skeleton frames corresponding to a home-made action "High throw" (left panel) and an action "Basketball shoot" (right panel) in UTD-MHAD dataset. Color images and skeleton frames of home-made action were collected by authors through a mobile phone camera and a Kinect camera, respectively. In the UTD-MHAD standard dataset test, we randomly selected 30 samples of each action class, and divided them into training set and testing set at a ratio of 9:1. Specific encoding processes can be found in Supplementary Fig. 20. Figure 3b shows partial time sections of reservoir states triggered by the light pulse trains and collected from the IGZO synaptic transistors. Different mask matrixes were used and the optimized performance was achieved with $M = 30$ and $L = 30$. After training and testing processes, the recognition rate was calculated by counting the recognition results of the entire data in the testing set. Repeated random subsampling validation method was used by ten times to enhance the reliability of results[37]. The average of ten results was used as the final system recognition result.

Figure 3c shows the recognition results with respect to UTD-MHAD dataset. The recognition rate based on all action classes reaches 93.58%. As can be seen, most actions can be well recognized and only a few actions like "Wave" are with relatively low recognition accuracy. Among all action classes, the system achieves an excellent recognition rate of 100% on 14 classes, and high accuracy (≥ 90%) on 21 classes. This shows that our system can well distinguish multiple types of complex action processes.

The recognition tasks were also verified with different numbers (2, 3, and 4) of Gaussian receptive fields as shown in Fig. 3d. The Gaussian centers were also sequentially set according to Eq. (6). As shown in Fig. 3d, high recognition accuracy (>90%) on UTD-MHAD dataset is obtained by three receptive fields. This is consistent with the results of MC metrics in Fig. 2f. We further investigated the effect of standard deviation of Gaussian receptive fields on the recognition results. Our results indicate that σ can have a slight effect on recognition accuracy. By varying σ from 0.1 to 10, the maximum variation on recognition accuracies is only 1.11% (Supplementary Fig. 21). Computation complexity can be reduced based on such GRF encoding scheme. On the one hand, the scale of our network parameters are almost 2 orders of magnitudes smaller than previous machine learning algorithms (e.g., ResNet18) regarding UTD-MHAD dataset (detailed illustration can be found in Supplementary Fig. 22)[38–40]. On the other hand, multiple training iterations are not required in our training process, and only one-step linear regression is needed, which also reduces the computation complexity. Our results indicate that GRF-based preprocessing is vital important for implementing the human action recognition by reservoir computing, and the very limited number of GRF neurons can have a significant improvement in recognition accuracy.

We further verified this bioinspired reservoir computing paradigm on human action recognition tasks by three datasets of MSR Action3D, Florence 3D, and MSR Action Pairs. The MSR Action3D dataset contains 20 classes of skeleton-based action frames, and 20 skeleton joints' coordinates are recorded per frame[41]. The dataset was used for verification, where 90% samples were chosen randomly for training and the rest for testing. A high recognition rate of 90.50% is obtained (Supplementary Fig. 23). Florence 3D dataset contains 9 action classes, and 15 skeleton joints' 3D coordinates are recorded in each frame[42]. Our system can identify 91.11% of the testing samples successfully (Supplementary Fig. 24). The actions in MSR Action pairs dataset are in paired with similar action trajectories[43]. This makes it difficult to recognize actions in a pairwise relationship. As shown in Supplementary Fig. 25, 12 action classes can be well distinguished with an average recognition accuracy of 90.67%. Specific parameters of bioinspired reservoir on three dataset standard tests can be found in Supplementary Table 2. Repeated random subsampling validation method was applied by five times in the three tasks to enhance the reliability of results. In addition, such bioinspired reservoir computing paradigm can be also applied for high-precision and multi-type action recognition task based on depth images in videos (achieving an accuracy of 92.35% on 27 labels, and the calculation process can be found in Supplementary Note 4). As high recognition accuracies (>90%) can be achieved on different validation datasets as shown in Fig. 3e and Supplementary Fig. 26, our system exhibits remarkable versatility and fault tolerance for human action recognition.

https://doi.org/10.1038/s41467-025-56899-3

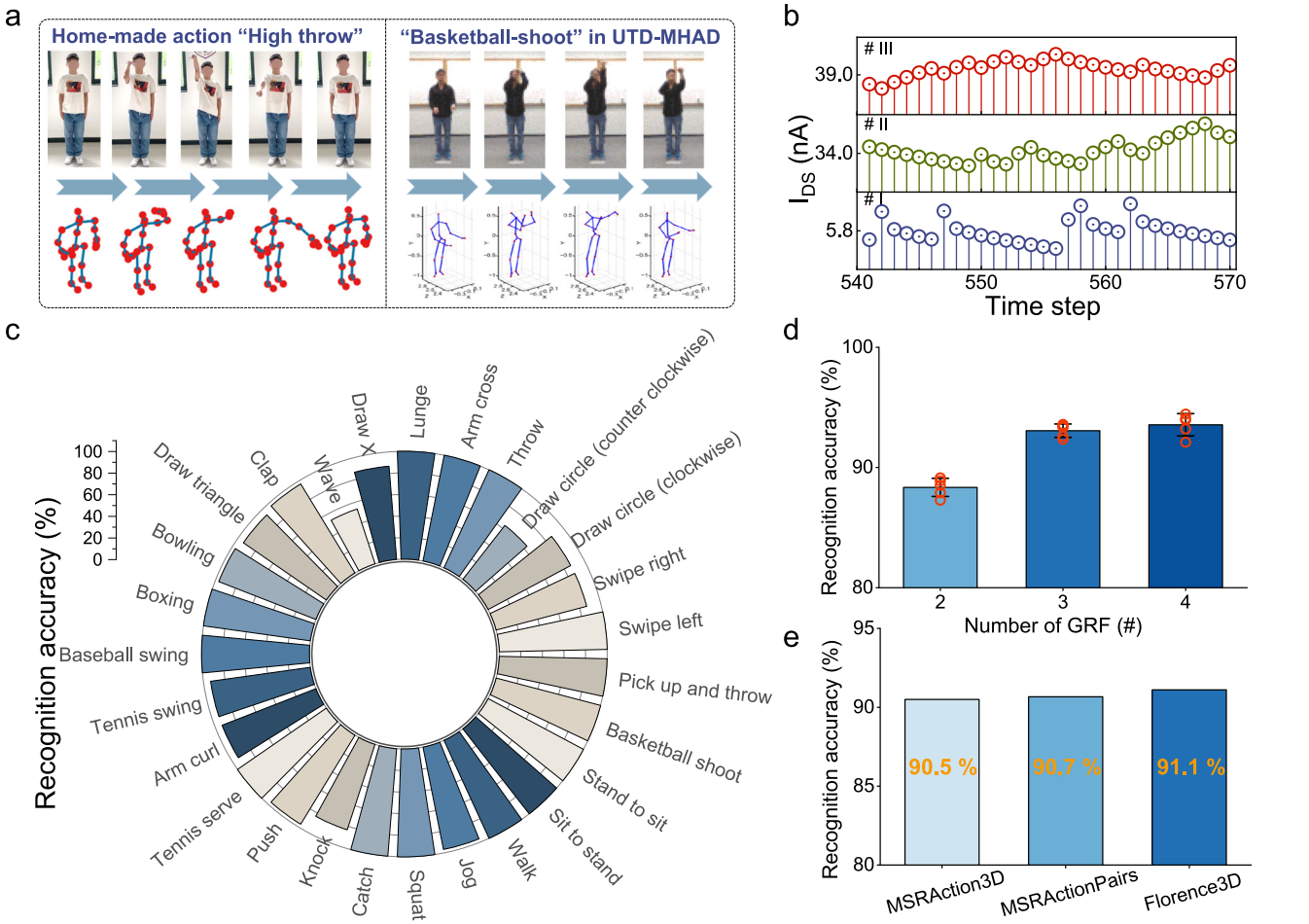

Fig. 3 | Recognition results and analysis on standard dataset tests based on bioinspired reservoir computing paradigm. a Examples of digital images and skeleton frames corresponding to a home-made action "High throw" and an action "Basketball shoot" in UTD-MHAD dataset, where the digital images and skeleton frames of home-made action were collected by authors through a mobile phone camera and a Kinect camera, respectively. Digital images and skeleton data in UTD-MHAD dataset is reprinted with permission[36], Copyright 2015, IEEE. **b** The partial time sections of reservoir states triggered by the light pulse trains and collected from the IGZO synaptic transistors. **c** Specific recognition results with bioinspired reservoir computing on UTD-MHAD dataset. **d** The recognition accuracies of UTD-MHAD dataset standard test with different numbers (2, 3, and 4) of Gaussian receptive fields. The error bars represent the standard deviation of five independent tests. **e** Recognition accuracies achieved on different validation datasets with bioinspired reservoir computing.

Recognition of falling behaviors based on analog system

The conceptual diagram of our Alpho-RC system for real-world human action processing tasks is shown in Fig. 4a. Skeleton frames of daily activities collected by the Microsoft Kinect camera are the original input, and the Gaussian receptive field based encoding mechanism is implemented. During the training process, currents obtained by population transistors in response to encoding pulse trains are used as reservoir states for training output weights. During the testing process, resulting output weights are mapped by 1T1R array, and output currents of testing action sample are obtained. The video processing tasks are mostly involved in human action recognition and prediction. Human action recognition tasks usually focus on completed actions, and achieve classification purposes by learning the entire processes of actions. However, in actual scenarios, it is often necessary to achieve early classification and prediction before action ends, such as predicting and alerting before falling behaviors are completed[44]. Purpose of the prediction task is to output the corresponding action label by using only the early frame sequences for inferencing.

Falling behaviors as common events in real life often involves potential health risks. For example, elderly people living alone and children are more likely to suffer physical injuries if they fall, thus affecting their health. Effectively identify falling behaviors from normal behaviors is crucial to the safety of both young and old[45,46]. A home-made 3D skeleton-based falling dataset was used for verifying the efficacy of such system for dealing with real-world tasks. Such dataset including two normal actions (squat and stretch) and three falling actions (fall down, fall to the left and fall to the right) was collected and built by both digital and Kinect camera. Figure 4b shows examples of the five actions recorded by a mobile phone camera. Figure 4c shows examples of action skeleton frames recorded by a Kinect camera correspondingly. Specific demonstration of skeleton frames can be found in Supplementary Video II. The collection details can be found in the "Methods". Here, the recognition on such home-made falling dataset was implemented by our Alpho-RC system. The 90% of action samples in the dataset were randomly selected as the training set, and the output weights were obtained by noise-aware linear regression training method ($M = 6$, $L = 5$). The detailed training process can be found in Supplementary Note. 5. The offline-trained weights were then programmed into the memristor cells in the 1T1R array. The numerical weight values were firstly mapped to device conductance from 200-1300 μS (Fig. 4d) and then programmed into the array using a write-and-verify programming scheme. Detailed reliability data of the hardware setup can be found in Supplementary Figs. 27, 28. Good programming accuracies were achieved with write error tolerance of ±1% (Fig. 4e). The memristor-based output layer successfully classified five actions with recognition accuracy of 96.67%

https://doi.org/10.1038/s41467-025-56899-3

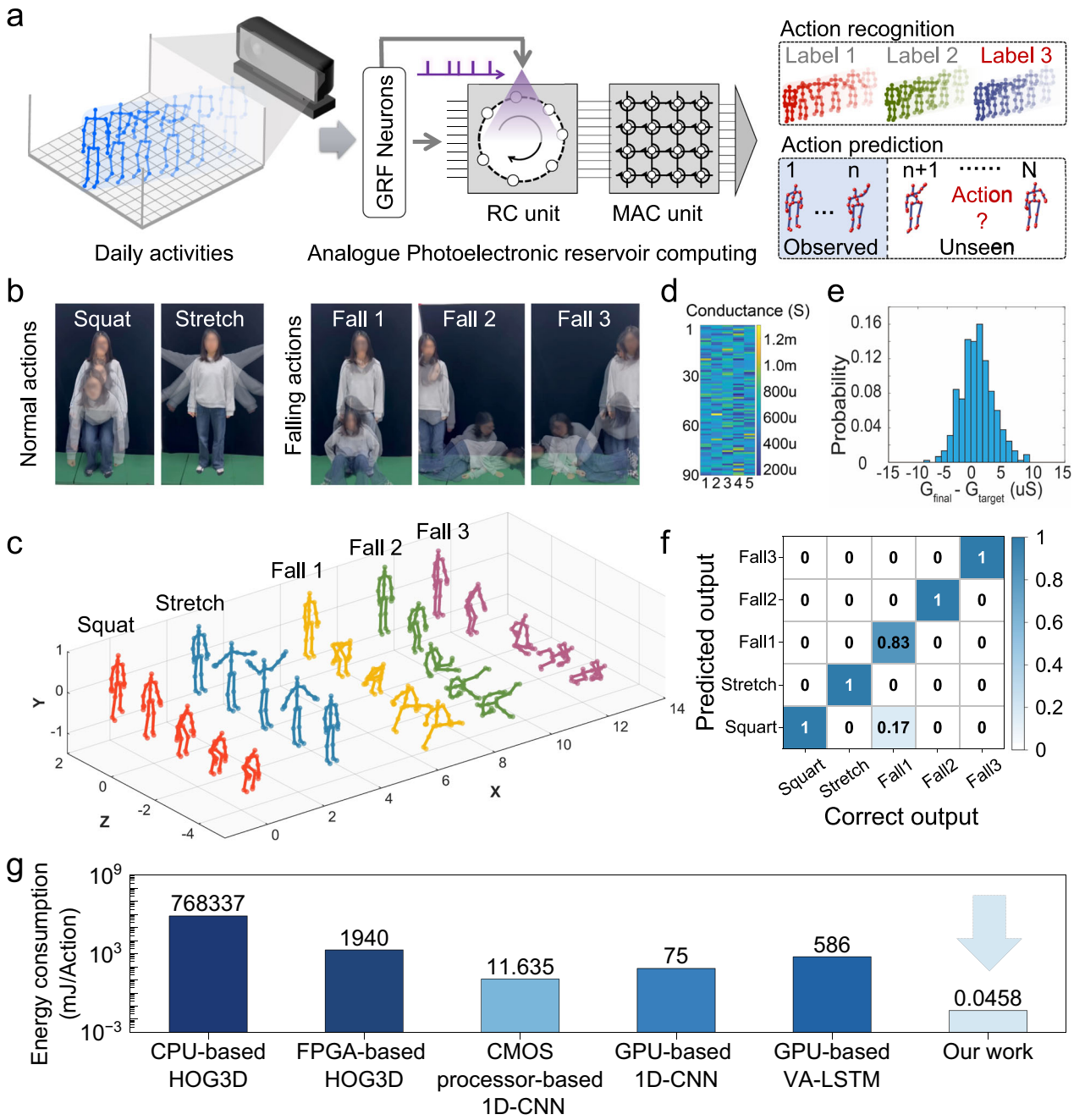

Fig. 4 | Recognition of falling behaviors based on Alpho-RC system. a The computational architecture schematic diagram of Alpho-RC system for human action recognition and prediction. **b** Examples of five actions including two normal actions (squat and stretch) and three falling actions (fall down, fall to the left and fall to the right) recorded by a mobile phone camera**. c** Examples of action skeleton frames corresponding to five actions. **d** Device conductance from 200–1300 μS mapped from the numerical weight values. **e** The write error tolerance of ±1% during mapping program. **f** Recognition accuracy of 96.67% obtained by memristor-based output layer. **g** Comparison of energy consumption per action among Alpho-RC system and algorithms based on various advanced processors.

(as shown in Fig. 4f), which is close to software simulations of 98.33% (as shown in Supplementary Fig. 29). The successful demonstration of human action recognition in the hardware based bioinspired in-materia reservoir computing paves the way for energy-efficient edge computing applications. Our Alpho-RC system was also applied for human action prediction. Only the observed frames were fed into the photoelectronic reservoir for predication. Here, the ratio between the number of observed frames and the frames of the completed action is defined as observation ratio. As shown in Supplementary Fig. 30, the prediction accuracies are plotted as function of the observation ratios with respect to home-made falling dataset. The prediction accuracies are larger than 80% with the observation ratio higher than 50% and larger than 90% with the observation ratio higher than 70%. Such results indicate our system can achieve excellent prediction results, verifying the capability of both human action recognition and prediction.

Finally, the energy efficiency of such system is evaluated by calculating the energy consumption corresponding to processing one completed action. As mentioned before, no complex filtering process is required for the feature extraction of every action. The most energy consumption before the training of weights connected to the output layer should be attributed to encoding process and the current flow through transistors. The energy consumption for processing an action by our system is only ~45.78 μJ (detailed estimation can be found in

experimental section). However, feature extraction is required in most previous software-based methods, which dramatically increases the floating-point operations and the energy consumption in turn. The comparison of energy consumption is summarized in Fig. 4g. Our Alpho-RC system is at least two orders of magnitude lower than that of CMOS-based processors including CPU, FPGA, and GPU[47–49]. We also compared the reported physical reservoir computing architectures for human action recognition (Supplementary Table 3), achieving largest type number of human action classification[13,50–52].

Discussion

In conclusion, we developed a biologically plausible Alpho-RC architecture for implementing high energy efficient human action processing. Such architecture provides an advanced bioinspired in-materia reservoir computing framework for processing dynamic coordinates of skeleton joints without prior feature extraction process, which would facilitate the development of full-hardware implementation of reservoir computing. In such framework, the feature extraction could be incorporated in the population encoding by the GRF neurons and the nonlinear mapping by the photoelectronic synaptic transistors. In this case, the computation complexity is greatly reduced, in comparison with the previous methods that require complex feature extraction algorithms and multiple iterations in most feedforward networks. In many previous reports, not only prior feature extraction process but also feature enhancement process are required. Feature enhancement process is always implemented in complex computational modules like CAG (Coordinate Aware Grouping) and VAG (Virtual-part Aware Grouping) modules[53]. In this work, the frames of dynamic coordinates were encoded into only three spike trains, greatly simplifying the complexity of the RC system. The richness of reservoir states can be originated from the electron/proton electrostatic coupling at the IGZO/electrolyte interface. The introduction of light stimulation can further enrich the reservoir states based on the persistent photoconductivity effects in the IGZO channel. More importantly, the light can convey the spike encoding information in a fast, noise-robust, and high bandwidth manner which might promote the throughput and reliability of system[54,55]. For example, the usage of light can alleviate the congestion of electrical operations and broaden the data capacity range in computing. This could improve the efficiency for processing multiple floating-point operations, such as human action recognition[56,57]. Such biologically plausible architecture is very efficient for both human action recognition and prediction, indicating the remarkable feasibility and versatility.

More importantly, a 1T1R array as readout layer is integrated with the photoelectronic reservoir to construct an analog photoelectronic reservoir computing system. High accuracy of over 96% on real-world falling action recognition demonstrates the Alpho-RC system's application potential in the field of smart healthcare. Meanwhile, the estimated energy consumption for processing per action is two orders of magnitude lower than the reported digital methods. Comparison of PRC devices for human action processing also illustrates the advantages of our Alpho-RC system (Supplementary Table 3). Our devices showed the capability of recognizing human actions with the largest number and a simple training process. Additionally, our devices have been verified by homemade datasets, and the potential for fully analog reservoir computing has been verified by using memristor array as the readout layer. Future improvements may be devoted to the optimizing of photoelectronic synaptic transistor with respect to operating speed, and the development of large-scale system for high throughput processing. Previously, a $SrTiO_3$-based memristor with multiple synaptic functions emulations has been proposed. A modified short-term plasticity neuron (m-STPN) was built based on such memristor. Such m-STPN has been applied on a bio-inspired deep neural networks (DNN), and an estimated gain in energy efficiency between 96× and 966× were achieved[58]. This concept can be borrowed for further optimizing our Alpho-RC system. Our Alpho-RC system provides a bionic computing paradigm with low energy consumption for the applications in virtual reality system, human-computer interaction system, smart medical care and many more.

Methods

Device fabrication of photoelectronic transistor array

The synapse array integrates 10×10 transistors on a silicon substrate (Si/SiO_2). Firstly, Al films were deposited by thermal evaporation through shadow masks as gate electrodes on substrate. Then, SiO_2 electrolyte films with loose and porous structure were deposited by plasma-enhanced chemical vapor deposition (PECVD) on the patterned gate electrodes under low temperature condition through shadow masks or photolithography process (Litho-ACA Pro, TuoTuo Technology). During the deposition process, SiH_4/N_2 mixture (5% SiH_4 + 95% N_2) and N_2O were used as reactive gases with flow rate of 40 sccm and 120 sccm at 140 °C. Next, IGZO films were sputtered as channel layers by RF magnetron sputtering process through shadow masks on the surface of SiO_2 electrolyte films. During sputtering process, the flow rate of Ar gas was 30 sccm and the working pressure was 0.3 Pa. Finally, Al films were deposited as source electrodes and drain electrodes by thermal evaporation through shadow masks. The IGZO channel length and width of all devices in the array were 80 and 320 μm, respectively. The surface roughness of SiO_2 film was characterized by atomic force microscopy (Asylum Cypher S system). The photoelectric performances of transistors were tested by a semiconductor parameter analyzer (Keithley 2636B). The laser used in the photoelectric test was the fiber coupled laser module (CNI Laser PGL-FC-405nm). All device tests were performed at room temperature with relative humidity of 50%.

Device fabrication of 1T1R array

The 1T1R arrays were fabricated by integrating the memristors on the top of CMOS chips in house, while CMOS chips were fabricated in a commercial foundry with 180 nm technology node. Firstly, via 5 for memristors and transistors connections were exposed after dry etching treatment to remove the passivation layer on the chip surface using SF_6/Ar and C_4F_8 plasma, respectively. Then, 20/3 nm-thick Pt/Ti bottom electrodes were patterned by photolithography and deposited by DC sputtering. Next, 8nm-thick TaO_x films were fabricated by photolithography and reactive sputtering using a Ta target with Ar/O_2 (2:1) ambient. Finally, 25 nm-thick Ta top electrodes were patterned by photolithography and deposited by DC sputtering, followed by sputtering of 20 nm-thick Pt protection films.

The electric characteristics of memristors and CMOS transistors in the 1T1R array were evaluated using semiconductor parameter analyzers (Keithley B1500A and B1531A).

Detailed procedures of training and testing processes

In the training process, obtained reservoir states are used to train the output weights (W_{out}) with target by one-step linear regression. Recorded reservoir states per frame are regarded as a $(3\times L\times M)$-dimensional vector, and target is a c-dimensional vector, where c is the number of action classes. The label value is corresponded to the index position of the c-dimensional vector, which is set to 1. The remaining index positions are set to 0. By combining the state vectors and target vectors of training data, matrixes of X and Y_{target} are arranged for output weights (W_{out}) training, which are calculated by:

$$W_{out} = Y_{target} X^{T} (X X^{T})^{\dagger}$$

where symbol "†" and "T" represent matrix operations of Moore–Penrose pseudo-inverse and transpose, respectively[33]. In the testing process, the trained weights are multiplied by the reservoir state vector per frame of one testing data. All frames' ouput vectors are

summed up as a one-dimensional vector. Among it, the index position of the maximum value is treated as the output result label by the winner-take-all method. The output label would be compared with the target label. If they are consistent, it is judged that the test data recognition is successful.

Collection of human action skeleton frames

The home-made falling dataset contains five types of actions in total, including two normal actions (squat and stretch) and three falling actions (fall down (fall 1), fall to the left (fall 2), and fall to the right (fall 3)). Twelve adults participated in the action collection, and each person performed each action 10 times, so that the dataset contains 600 samples. The human action skeleton information was collected using one Microsoft Kinect V2 camera. The person stood at a distance of two meters from the camera to act. During the process of executing an action, Kinect SDK software package was called through Matlab in real time to track skeleton information of each frame. The 3D coordinates of 25 skeleton joints were recorded in each frame. The acquisition frequency of skeleton frames was approximately 10 frames·s^{-1}. The skeleton frame number of action samples in dataset is between 19 and 82.

Energy consumption estimation

In Alpho-RC system, energy consumption mainly comes from three steps: GRF-based encoding process of population encoders, the generation process of reservoir states and the weighted computations in the output layer. We use the home-made falling dataset as an evaluation benchmark. Among it, the longest skeleton sequence contains 82 frames, which represents the maximum energy consumption of the Alpho-RC system processing an action. During GRF-based encoding process, the operation steps of GRF neurons were implemented based on simulated method and can be implemented by electronic components. Here, we evaluate the hardware overhead of GRF neurons based on FPGA (ZYNQXC7Z020). In the recognition task, 82×6 = 492 time steps were generated by time multiplexing ($M = 6$) and L ($L = 5$) population encoders (each containing three GRF neurons) were used for GRF encoding. Therefore, 492 × 5 = 2460 operations are needed to be executed on GRF neurons. Each operation contains the calculation process of output values under three Gaussian distributions and the comparison process of output values (i.e., WTA rules). The total time of 2460 operations based on FPGA was estimated to 26560 ns. The working power of FPGA during operations was 1.691 W. Hence, the energy consumption of GRF neurons for whole process was estimated as 26560 ns × 1.691 W ≈ 44.91 μJ. During the generation process of reservoir states, the average energy consumption per action can be estimated by: $P = I \times T \times V_{DS}$, where T is the total time of pulse sequences encoded for an action. Parameters I and V_{DS} represent the average response currents and the source-to-drain voltage. By time multiplexing ($M = 6$), 82 × 6 = 492 time steps were generated. The source-to-drain voltage was 0.05 V. The average response currents corresponding to the action sample with 82 frames was statistically 18.39 nA. Hence the average energy consumption of one population encoder was 18.39 nA×492×3×38 ms×0.05 V ≈ 0.052 μJ. L ($L = 5$) population encoders were used for GRF encoding, so that the energy consumption of the generation process of reservoir states can be estimated as 0.052 μJ×5 ≈ 0.26 μJ. In the memristor-based output layer, the input reservoir states were encoded to 11-bit square pulses with pulse width of T = 1 μs and pulse amplitude of 0.1 V (average amplitude of $V_{avr} = 0.05$ V). The total time steps of computations were $N = 82$ frames × 11 bits = 902 steps. The average energy consumption per action can then be estimated as $V_{avr}^2 \times G \times T \times N$, where G was the average conductance of the memristors (600 μS). Therefore, the average energy consumption per action was 902×(0.05 V)^2×600 μS×1 μs = 1.353 nJ per memristor and the total energy consumption of the output layer was 1.353 × 90 × 5 ≈ 0.61 μJ. Therefore, the energy consumption of Alpho-RC system for recognizing one action can be estimated as 44.91 + 0.26 + 0.61 ≈ 45.78 μJ.

We compared system energy consumption with machine learning algorithms based on different microprocessors (CPU/FPGA/CMOS-based processor/GPU). Previously, a HOG3D feature extraction method was implemented for human action recognition[47]. This work provided the energy consumption per frame based on CPU and FPGA respectively, where the UCF50 dataset was used as test benchmark. In this dataset, each video contains 97 frames. Therefore, the energy consumption per action of CPU-based and FPGA-based HOG3D is estimated as 7.921 J × 97 = 768337 mJ, 0.020 J × 97 = 1940 mJ, respectively.

Statistical analysis

The calculation processes of computational architecture were implemented in Matlab and Python. The data analysis and displaying were realized in Origin.

Data availability

The data that support the plots within this paper and other findings of this study are available from the corresponding author on request. Source data are provided with this paper.

Code availability

The code that supports this study is available from Code Ocean at https://doi.org/10.24433/CO.7958352.v1 or from the corresponding author on request.

References

1. Tye, N. J., Hofmann, S. & Stanley-Marbell, P. Materials and devices as solutions to computational problems in machine learning. *Nat. Electron.* **6**, 479–490 (2023).
2. Yang, C. et al. Photoelectric memristor-based computer vision for artificial intelligence applications. *ACS Mater. Lett.* **5**, 504–526 (2023).
3. Chai, Y. In-sensor computing for computer vision. *Nature* **579**, 32–33 (2020).
4. DeAngelis, G., Ohzawa, I. & Freeman, R. Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature* **352**, 156–159 (1991).
5. Tanabe, S. Population codes in the visual cortex. *Neurosci. Res.* **76**, 101–105 (2013).
6. Sabarigiri, B. & Suganyadevi, D. The possibilities of establishing an innovative approach with biometrics using the brain signals and iris features. *Res. J. Recent Sci.* **2277**, 2502 (2013).
7. Hong, X. T. et al. Two-dimensional perovskite-gated AlGaN/GaN high-electron-mobility-transistor for neuromorphic vision sensor. *Adv. Sci.* **9**, 2202019 (2022).
8. Zhou, F. et al. Optoelectronic resistive random access memory for neuromorphic vision sensors. *Nat. Nanotechnol.* **14**, 776–782 (2019).
9. Han, J.-K. et al. Bioinspired photoresponsive single transistor neuron for a neuromorphic visual system. *Nano Lett.* **20**, 8781–8788 (2020).
10. Lee, J. et al. Light-enhanced molecular polarity enabling multi-spectral color-cognitive memristor for neuromorphic visual system. *Nat. Commun.* **14**, 5775 (2023).
11. Chen, J. et al. Optoelectronic graded neurons for bioinspired in-sensor motion perception. *Nat. Nanotechnol.* **18**, 882–888 (2023).
12. Pan, Xuan et al. Parallel perception of visual motion using light-tunable memory matrix. *Sci. Adv.* **9**, eadi4083 (2023).
13. Tan, H. & van Dijken, S. Dynamic computer vision with retino-morphic photomemristor-reservoir computing. *Nat. Commun.* **14**, 2169 (2023).
14. Song, H. et al. Fully memristive elementary motion detectors for a maneuver prediction. *Adv. Mater.* **36**, 2309708 (2024).
15. Demb, J. B. & Singer, J. H. Functional circuitry of the retina. *Annu Rev. Vis. Sci.* **1**, 263–289 (2015).

16. Grill-Spector, K. & Malach, R. The human visual cortex. *Annu Rev. Neurosci.* **27**, 649–677 (2004).
17. Balasubramanian, V. & Sterling, P. Receptive fields and functional architecture in the retina. *J. Physiol.* **587**, 2753–2767 (2009).
18. Ji, S., Xu, W., Yang, M. & Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 221–231 (2013).
19. Chen, C., Liu, K. & Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. *J. Real.-Time Image Process.* **12**, 155–163 (2013).
20. Song, S., Lan, C., Xing, J., Zeng, W. & Liu, J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 31 (AAAI Press, 2017).
21. Bruno, B., Mastrogiovanni, F., Sgorbissa, A., Vernazza, T. & Zaccaria, R. Analysis of human behavior recognition algorithms based on acceleration data. In *2013 IEEE International Conference on Robotics and Automation* 1602–1607 (IEEE, 2013).
22. Wang, W., Liu, A. X., Shahzad, M., Ling, K. & Lu, S. Understanding and modeling of wifi signal based human activity recognition. In *Proc. 21st Annual International Conference on Mobile Computing and Networking* 65–1676 (Paris, France, 2015).
23. Sharpee, T. O. Computational identification of receptive fields. *Annu. Rev. Neurosci.* **36**, 103–120 (2013).
24. Sze, V., Chen, Y.-H., Emer, J., Suleiman, A. & Zhang, Z. Hardware for machine learning: challenges and opportunities. In *2017 IEEE Custom Integrated Circuits Conference (CICC)* 1–8 (IEEE, 2017).
25. Yang, Y. et al. Reservoir computing based on electric-double-layer coupled InGaZnO artificial synapse. *Appl. Phys. Lett.* **122**, 043508 (2023).
26. Lee, M. et al. Brain-inspired photonic neuromorphic devices using photodynamic amorphous oxide semiconductors and their persistent photoconductivity. *Adv. Mater.* **29**, 1700951 (2017).
27. Ke, S. et al. Indium-gallium-zinc-oxide based photoelectric neuromorphic transistors for modulable photoexcited corneal nociceptor emulation. *Adv. Electron. Mater.* **2100487**, 1–9 (2021).
28. Zhong, Y. et al. Dynamic memristor-based reservoir computing for high-efficiency temporal signal processing. *Nat. Commun.* **12**, 408 (2021).
29. Liang, X. et al. Rotating neurons for all-analog implementation of cyclic reservoir computing. *Nat. Commun.* **13**, 1549 (2022).
30. Torrejon, J. et al. Neuromorphic computing with nanoscale spintronic oscillators. *Nature* **547**, 428–431 (2017).
31. Li, Chuankun. et al. Skeleton-based action recognition using LSTM and CNN. In 2017 IEEE International conference on multimedia & expo workshops (ICMEW) 585-590 (IEEE, 2017).
32. Bohte, S. M. et al. Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer RBF networks. *IEEE Trans. neural Netw.* **13**, 426–435 (2002).
33. Lukosevicius, M. & Jaeger, H. Survey: reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.* **3**, 127–149 (2009).
34. Appeltant, L. et al. Information processing using a single dynamical node as complex system. *Nat. Commun.* **2**, 468 (2011).
35. Lee, O. et al. Task-adaptive physical reservoir computing. *Nat. Mater.* **23**, 79–87 (2024).
36. Chen, C., Jafari, R. & Kehtarnavaz, N. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)* 168–172 (IEEE, 2015).
37. Pei, M. et al. Power-efficient multisensory reservoir computing based on Zr-Doped HfO_2 memcapacitive synapse arrays. *Adv. Mater.* **35**, 2305609 (2023).
38. Tasnim, N., Islam, M. K. & Baek, J. H. (2021). Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints. *Appl. Sci.* **11**, 2675 (2021).
39. Usmani, A., Siddiqui, N. & Islam, S. Skeleton joint trajectories based human activity recognition using deep RNN. *Multimed. Tools Appl.* **82**, 46845–46869 (2023).
40. Tasnim, N. & Baek, J. H. Dynamic edge convolutional neural network for skeleton-based human action recognition. *Sensors* **23**, 778 (2023).
41. Li, W., Zhang, Z. & Liu, Z. Action recognition based on a bag of 3D points. In *2010 IEEE Conference on Computer Vision and Pattern Recognition* 9–14 (IEEE, 2010).
42. Seidenari, L., Varano, V., Berretti, S., Del Bimbo, A. & Pala, P. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *2013 IEEE Conference on Computer Vision and Pattern Recognition* 479-485 (IEEE, 2013).
43. Oreifej, O. & Liu, Z. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *2013 IEEE Conference on Computer Vision and Pattern Recognition* 716–723 (IEEE, 2013).
44. Kong, Y., Kit, D. & Fu, Y. A discriminative model with multiple temporal scales for action prediction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland,* September 6-12, 2014, Proceedings, Part *V 13* 596-611 (Springer, 2014).
45. Rubenstein, L. Z. Falls in older people: epidemiology, risk factors and strategies for prevention. *Age Ageing* **35**, ii37–ii41 (2006).
46. Haarbauer-Krupa, J. et al. Fall-related traumatic brain injury in children ages 0–4 years. *J. Saf. Res.* **70**, 127–133 (2019).
47. Ma, X., Borbon, J. R., Najjar, W. & Roy-Chowdhury, A. K. Optimizing hardware design for human action recognition. In *2016 26th International Conference on Field Programmable Logic and Applications (FPL)* 1-11 (IEEE, 2016).
48. Zhang, B., Han, J., Huang, Z., Yang, J. & Zeng, X. A real-time and hardware-efficient processor for skeleton-based action recognition with lightweight convolutional neural network. *IEEE Trans. Circuits Syst. II: Express Briefs.* **66**, 2052–2056 (2019).
49. Zhang, P. et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision* 2117–2126 (IEEE, 2017).
50. Chen, R. et al. Thin-film transistor for temporal self-adaptive reservoir computing with closed-loop architecture. *Sci. Adv.* **10**, eadl1299 (2024).
51. Jang, Y. H. et al. A high-dimensional in-sensor reservoir computing system with optoelectronic memristors for high-performance neuromorphic machine vision. *Mater. Horiz.* **11**, 499–509 (2024).
52. Sun, Y. et al. In-sensor reservoir computing based on optoelectronic synapse. *Adv. Intell. Syst.* **5**, 2200196 (2023).
53. Xu, K., Ye, F., Zhong, Q. & Xie, D. Topology-aware convolutional neural network for efficient skeleton-based action recognition. *Proc. AAAI Conf. Artif. Intell.* **36**, 2866–2874 (2022).
54. Xu, X. Y. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
55. Nahmias, M. A. et al. Photonic multiply-accumulate operations for neural networks. *IEEE J. Sel. Top. Quantum Electron.* **26**, 7701518 (2020).
56. Feldmann, J. et al. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
57. Miller, D. A. B. Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
58. Weilenmann, C. et al. Single neuromorphic memristor closely emulates multiple synaptic mechanisms for energy efficient neural networks. *Nat. Commun.* **15**, 6898 (2024).

Acknowledgements
This work was supported by National Natural Science Foundation of China (Grant no. 92364106 and 92364204), the Fundamental Research Funds for the Central Universities (Grant no. 2024300393), the National Key Research and Development Program of China (Grant no. 2023YFE0208600), Major Program of Natural Science Foundation of Zhejiang Province in China (Grant no. LDQ23F040001), Nanjing Science and Technology Plan Project (Grant no. 202305001), and the Natural Science Foundation of Jiangsu Province (Grant no. BK20220121). We also thank for support from Ministry of Education Engineering Research Center for Optoelectronic Materials and Chip Technology.

Author contributions
C.W., Q.W., Y.S., P.L. and H.C. conceived and designed the experimental protocol. H.C., Y.X., M.P. designed the computational architecture and performed verifications. H.C., Y.Y. fabricated the device equipment and conducted testing. H.C., X.F., K.S., H.L. collected and produced the home-made dataset. Y.Y. performed the material characterization. S.K., L.Q., W.X., P.C. contributed to material analysis. H.C. and Y.X. contributed equally to this work. All authors contributed to the preparation of the manuscript and commented on it.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-56899-3.

Correspondence and requests for materials should be addressed to Peng Lin, Yi Shi, Qing Wan or Changjin Wan.

Peer review information *Nature Communications* thanks Kyung Min Kim and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.