Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors

Ming Wang[1,3], *Zheng Yan*[2,3], *Ting Wang*[1], *Pingqiang Cai*[1], *Siyu Gao*[1], *Yi Zeng*[1], *Changjin Wan*[1], *Hong Wang*[1], *Liang Pan*[1], *Jiancan Yu*[1], *Shaowu Pan*[1], *Ke He*[1], *Jie Lu*[2], *Xiaodong Chen*[1]*

[1]Innovative Centre for Flexible Devices (iFLEX), Max Planck – NTU Joint Lab for Artificial Senses, School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798 Singapore.

[2]Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, 15 Broadway, NSW 2007, Australia.

[3]These authors contributed equally: Ming Wang, Zheng Yan.

*Corresponding author. E-mail: chenxd@ntu.edu.sg

Gesture recognition using machine learning methods is valuable in the development of advanced cybernetics, robotics, and healthcare systems, and typically relies on images or videos. To improve recognition accuracy, such visual data can be fused with data from other sensors, but this approach is limited by the quality of the sensor data and the incompatibility of the datasets. Here, we report a bioinspired data fusion architecture that can perform human gesture recognition by integrating visual data with somatosensory data from skin-like stretchable strain sensors. The learning architecture uses a convolutional neural network for visual processing, and then implements a sparse neural network for sensor data fusion and recognition. Our approach can achieve a recognition accuracy of 100%, and maintain recognition accuracy with noisy, under- or over-exposed images. We also show that our architecture can be implemented for robot navigation using hand gestures with a small error, even in the dark.

(150 words)

Human gesture recognition (HGR), which uses mathematical algorithms to interpret human motion, is of value in healthcare[1, 2], human-machine interactions[3-5], and the study of cognitive neuroscience[6]. Sensing and recognition methods often use algorithms that depend on visual images and/or videos. However, the efficiency of these methods is limited by the quality of the images, which are affected by environmental interference such as blocked objects (known as occlusions) and varying light conditions[7, 8].

One approach to overcome these issues is multimodal fusion, which combines visual data with additional sensor information (for example, instantaneous orientation, spatial positions or velocity of human gestures) obtained from wearable inertial[9-12], force[8], and optical oscillator[13] sensors. Multimodal fusion has been shown to improve the recognition accuracy and precision of HGR, but the approach is limited by poor-quality sensor data. Wearable sensors, in particular, are typically bulky, rigid and do not form an intimate contact with the user for high-quality data acquisition[14, 15]. Moreover, integrating visual datasets containing images or videos with wearable sensor datasets (usually recorded as one-dimensional time-series or discrete data) is challenging due to the mismatch in data dimensionality and data density (known as sparseness).

Different machine learning methods have been used to fuse visual and sensor data, including the hidden Markov model[9], support vector machine[16] and K-nearest neighbor[12, 16] classifiers, as well as deep convolutional neural networks (CNNs)[11, 17, 18]. A CNN is a powerful machine learning method because it can automatically learn hierarchical deep spatial features and extract shift-invariant features from original images[19-21]. As a result, CNNs have been applied successfully in visual recognition tasks such as image classification[22, 23] and playing strategic board games such as Go[24, 25]. However, current application of CNN in multimodal (visual-wearable sensor) fusion has been limited to decision-level fusion – that is, the visual and sensor data are first classified independently, and the classification results are merged later – because mismatched dimensionality and sparseness of the datasets remain an issue.

Recent physiological and neuroimaging results based on audiovisual–vocal detection[26, 27] and enhanced interactions with objects[28, 29] show that early interactions of different modalities (visual and somatosensory) in the multisensory neurons area in the brain (Fig. 1a), including the association area (AA), are beneficial for perceptual decision-making. While the fusion process of these early interactions is unclear[30, 31], the results suggest that converging visual and wearable data early in perceptual decision making could potentially improve the accuracy of the recognition tasks. Moreover, biological neural systems have demonstrated that the sparse connectivity between neurons leads to complex sensory data processing with global efficiency and little power[32].

In this Article, we report a bioinspired learning architecture that fuses visual images and somatosensory data from skin-like electronic devices early in the process for human gesture recognition tasks. Our bioinspired somatosensory–visual (BSV) associated architecture consists of three neural networks resembling the somatosensory–visual (SV) fusion hierarchy in the brain (Fig. 1b): a sectional CNN for early visual processing, a multilayer neural network for early somatosensory information processing, and a sparse neural network that fuses early visual and somatosensory information at a high level. The sectional CNN performs convolution operations that resemble the function of the local receptive field in biological nervous systems[33, 34], and thus mimics the initial processing of visual information in the visual primary areas (PA) (Fig. 1a). The sparse neural network represents the early and energy-efficient interactions of visual and somatosensory information in the

multisensory neurons area of the brain. Motivated by the stability theory of linear systems, we developed a pruning strategy based on a Frobenius condition number to obtain the sparse neural network. To capture somatosensory information, we built a stretchable sensor that is transparent, conformable, and adhesive using single-walled carbon nanotubes (SWCNTs).

Our BSV architecture can classify hand gestures against complex backgrounds with an accuracy of 100% using a custom SV dataset. Compared to two unisensory (visual- and somatosensory-based recognition) and three common SV fusion approaches (weighted-average, weighted-attention and weighted-multiplication fusion), our BSV architecture offers superior tolerance to noise, and over- and under-exposures. We also use the BSV architecture to control a quadruped robot with hand gestures, achieving an error of 1.7% under a normal illuminance of 431 lux and 3.3% under a dark illuminance of 10 lux.

Conformable, transparent and adhesive stretchable sensors

We fabricated a transparent, adhesive and stretchable strain sensor that can conformably attach on human skin to accurately capture somatosensory signals from human gestures. The sensor was made transparent to ensure it is inconspicuous in the visual information, and made stretchable up to 75% strain to meet the deformation limit of human parts in most activities[35]. The stretchable sensor is a three-layer stacked structure consisting of SWCNTs as the sensing component, a stretchable polydimethylsiloxane (PDMS) layer and an adhesive poly (acrylic acid) (PAA) hydrogel layer (Fig. 2a, see Methods and Supplementary Fig. 1 for details on

fabrication). Before thermal polymerization of PAA precursors, PDMS was treated with a mixture of argon and acrylic gas[36]. This plasma treatment chemically modifies the PDMS surface with an acrylic acid layer (Fig. 2b, and see Supplementary Fig. 2)[37], allowing PAA hydrogels to bond strongly with PDMS.

To minimize visual interference of the SWCNT layer, we tested different amounts of SWCNT and found that 40 μL of a 0.1 mg/mL solution in a pattern area of 2 × 0.4 cm^2 had the best combination of high transparency and reliable strain sensing performance (Supplementary Fig. 3). Vacuum filtration produced SWCNT films with optical transmittances of 89% at wavelength of 550 nm (cyan curve in Fig. 2c). After transferring the SWCNT film onto the PDMS substrate, transmittance remained high at 83% at 550 nm (orange curve in Fig. 2c). Upon PAA polymerization, the resultant conformable and adhesive stretchable strain sensor had an optical transmittance above 74% at wavelengths between 400 nm and 900 nm (red curve in Fig. 2c), allowing it to remain inconspicuous in the photographs (Fig. 2d, and inset of Fig. 2f). Furthermore, the polymerized PAA hydrogel layer, which is highly adhesive on human skin (1.27 N/cm^2 versus 0.07 N/cm^2 for PDMS), allowed the strain sensor to conformably attach on the finger (Fig. 2e, and see Supplementary Video 1).

The sensors showed stretchability up to 100% (Fig. 2f), which is enough for monitoring somatosensory signals from a human hand (inset in Fig. 2f). Time-dependent relative change ($\Delta R/R_0$) responses of five successive loading and unloading cycles at peak strains of 5%, 25%, 50%, and 75% show the sensors are

stable and can undergo various dynamic loading tests (Fig. 2g). To further verify the durability and reproducibility, we measured the $\Delta R/R_0$ responses of the strain sensor at a peak strain of 50% over 1000 cycles (Fig. 2h). The resistance response of the strain sensor was stable and regular with an almost constant base resistance. These results show that stretchable strain sensors can reliably collect somatosensory signals without affecting the visual images, making them suitable as somatosensory receptors for the BSV fusion architecture.

Data collection and classification performance

To implement a recognition task based on BSV associated learning, we built a custom SV dataset containing 3000 SV samples distributed into 10 categories of hand gestures (Fig. 3a,b). Each SV sample consists of one image of a hand gesture taken against a complex background, and one group of strain data captured from 5 strain sensors patched over the knuckle of each finger on one hand (Fig. 3a). The sensors extracted curvature information from the fingers, which were relevant to defining hand motions. Fig. 3c shows the flow diagram for preparing the SV dataset (Methods). Due to device variation and hysteresis, the raw strain data were first normalized before being used as the nominal somatosensory information for the hand gesture (Supplementary Note 1 and Supplementary Fig. 4). A commercial off-the-shelf camera sensor was used to capture the hand gesture image as the nominal visual information.

We used t-distributed stochastic neighbor embedding (t-SNE) – a dimensionality

reduction technique – to visualize the group of 3000 strain data (Fig. 3d)[38]. Each point on the plot represents the somatosensory information of one hand gesture projected from the 5-dimensional strain data into two dimensions. The points of the same gesture category (*i.e.*, the same color) clustered together, forming roughly 10 categories of hand gestures (I to X). The slight overlap seen in some categories is due to the similarity in bending states of the fingers in those gestures, which the strain sensors cannot easily distinguish. These results nonetheless show that somatosensory information from a human hand can provide valuable clues for hand gesture recognition.

We used the SV dataset and BSV associated learning architecture for hand gesture recognition. Figure 4a shows the detailed framework of the BSV associated learning architecture, including an AlexNet CNN[39] that was pretrained using the ImageNet dataset (Supplementary Note 2), and a 5-layer sparse neural network (sc8-sc12). Briefly, the pretrained AlexNet is used to learn a visual representation of a given hand gesture in a cost-effective and energy-efficient way. This learned visual output of AlexNet, which can be reviewed as transferable semantic features, is then concatenated with the learned somatosensory representation – a 5-dimensional vector of the collected strain data from one hand gesture – to form a new 53-dimensional vector that serves as an input to the 5-layer sparse neural network for final learning. The sparse connectivity of neural network is to enhance the energy-efficiency and generalization ability of the BSV architecture for scalable sensory data fusion, which has been demonstrated by both biological and computer

fields[32, 40]. Motivated by the stability theory of linear systems[41], we develop a pruning strategy that depends on the Frobenius condition number of global weighting matrix to achieve the sparse neural network (Fig. 4b, and see Supplementary Note 3). Briefly, these connections in a dense neural network are pruned if their removal does not lead to significant numerical increase of the Frobenius condition number of the weighting matrix. In the fusion procedure, the BSV associated learning is firstly trained on a 5-layer dense neural network using backpropagation algorithm and then is pruned via our sparse strategy based on the neural network toolbox of MATLAB. The training, validating, and testing samples were randomly selected with a ratio of 66:14:20 from the 3000 samples within the SV dataset. The final classification performance of the BSV associated learning strategy can reach up to 100% for hand gesture recognition. This classification results can compete with that (99.7%) using a dense neural network, and is superior to that (97.8%) based on other pruning strategy under the same BSV architecture (Supplementary Table 1), which indicates our pruning strategy can learn more general weights for making decision.

As a further comparison, we implemented two unisensory recognition approaches for hand gesture recognition, including visual-based recognition using only visual images based on a CNN, and somatosensory-based recognition using only strain sensor data based on a feedforward neural network (Supplementary Fig. 5). The receiver operating characteristic (ROC) curves were used to illustrate the recognition ability of these approaches as their discrimination threshold was varied. For each threshold, we calculated the sensitivity and specificity, which are defined as,

$$\text{sensitivity} = \frac{\text{true positive}}{\text{positive}}$$

$$\text{specificity} = \frac{\text{true negative}}{\text{negative}}$$

where ‘true positive’ is the number of correctly predicted hand gestures for a given class, ‘positive’ is the number of hand gestures of the given class, ‘true negative’ is the number of correctly predicted hand gestures except the given class, and ‘negative’ is the number of hand gestures except for the given class. The BSV associated learning exhibited the best classification sensitivity and specificity in all 10 categories of hand gestures (Fig. 4c); compared with visual- and somatosensory-based recognition, the area under the ROC curve for BSV associated learning showed maximum value (Fig. 4d,e). Furthermore, the confusion matrix maps for these approaches showed that in a testing dataset containing 600 samples, minimum testing samples were misrecognized in the BSV associated learning (Supplementary Fig. 7a-c). The BSV associated learning achieved the best recognition accuracy (100%) compared with visual-based recognition (89.3%) and somatosensory-based recognition (84.5%) (Fig. 4f). These results demonstrate that pattern recognition in computer vision can be improved when coupled with somatosensory information obtained from skin-like electronic devices.

We further compared BSV associated learning with three other known SV fusion architectures – weighted-average fusion (SV-V), weighted-attention fusion (SV-T) and weighted-multiplication fusion (SV-M) – for hand gesture recognition using the same SV dataset, and the same training and testing dataset (Supplementary Fig. 6). In SV-V, the average of two probabilities ($p(x_{\text{visu}})$ and $\text{p}(x_{\text{soma}})$) was taken as the

final output:

$$p(x_{\mathrm{V}}) = 0.5 \times p(x_{\mathrm{visu}}) + 0.5 \times p(x_{\mathrm{soma}})$$

where $p(x_{\mathrm{visu}})$ is the output probability of the last layer of CNN in the visual-based recognition strategy, and $\mathrm{p}(x_{\mathrm{soma}})$ is the output probability of the last layer of feedforward neural network in the somatosensory-based recognition strategy. In SV-T, a weighted integration of these probabilities was used as the final output:

$$p(x_{\mathrm{T}}) = m \times p(x_{\mathrm{visu}}) + n \times p(x_{\mathrm{soma}})$$

where *m* and *n* are obtained from an addition least-square training process. In SV-M, the multiplication of $p(x_{\mathrm{visu}})$ and $\mathrm{p}(x_{\mathrm{soma}})$ was used as the final output:

$$p(x_{\mathrm{M}}) = p(x_{\mathrm{visu}}) \times p(x_{\mathrm{soma}})$$

The error rates in the three common SV fusion recognition strategies (6.3% for SV-V, 4.2% for SV-T, and 3% for SV-M) were significantly higher than BSV associated learning (Fig. 4f, and see Supplementary Fig. 7c-f). Moreover, we compared the BSV associated learning with the state-of-art recognition approaches for hand gesture application (Supplementary Table 2) and also evaluated it on a public hand gesture dataset (Supplementary Table 3). The comparison results demonstrate that the BSV associated learning always maintains the best classification performance (> 99%), indicating the BSV can make best use of the complementary visual and somatosensory information due to its early interactions and rational visual preprocessing.

We also assessed the effect of visual noise on the recognition accuracies of these trained models (visual, SV-V, SV-T, SV-M, and BSV based recognition strategies) by

adding Gaussian white noise into the images in the testing dataset (Supplementary Fig. 8). Increased noise level significantly deteriorated the recognition accuracies of the visual, SV-V, SV-T, SV-M based strategies, while BSV continued to maintain high recognition accuracies (Fig. 4g). These results show our BSV associated learning architecture is tolerant of defects in the visual information, such as noise and motion blur, and is clearly superior to current multimodal recognition approaches due to the biological visual-somatosensory interaction.

Precise HGR for human-machine interactions

As proof-of-concept application for human-machine interactions, we built an auto-recognition and feedback system that allows humans to interact with a robot through hand gestures using our BSV associated learning architecture. This system consists of a data acquisition unit (DAQ) for capturing the somatosensory information of a hand, a built-in camera for capturing the images of hand gestures, a computer for implementing the BSV associated learning, a wireless data transmission module, and a quadruped robot (Fig. 5a, photograph of the system is shown in Supplementary Fig. 9). Each of the 10 categories of hand gestures was assigned a specific motor command that relates to directional movements (Fig. 5b). The different hand gestures were then used to guide the quadruped robot through a labyrinth. Hand gesture recognition powered by our BSV associated learning architecture was able to guide the robot through the labyrinth with zero error, compared to 6 errors in visual-based recognition (Fig. 5c and d, Supplementary Video 2).

Importantly, the robotic system based on our BSV fusion also worked effectively in the dark (Supplementary Video 3). We tested the recognition results of the ten categories of hand gestures in environments with varying illuminances (431, 222, and 10 lux) by adjusting the light condition in a room. The lowest illuminance of ~10 lux resembles an open parking lot at night. For each light condition, 60 trials (6 trials per category of hand gesture) were carried out to control the motion of the robot. Recognition accuracies for all four approaches (visual, SV-V, SV-T, SV-M, and BSV recognition) under illuminance of 431, 222, and 10 lux are shown in Fig. 5f. When the room lights faded, the accuracy of visual, SV-V, SV-T and SV-M approaches decreased dramatically while BSV associated learning maintained high accuracy (> 96.7%). This trend was consistent when testing was done using the same dataset but with the images mathematically processed to have varying brightness (Fig. 5e and Supplementary Fig. 10). Similar to the tolerance of noise defects, the BSV associated learning system is highly accurate and can withstand the harsh environments that cause under- or over-exposure in the images with the brightness times ranging from 0.4 to 2.5 (More scenarios with the brightness times below 0.4 are shown in Supplementary Fig. 11a). A part of the explanation is the complementary effect of somatosensory information that is invariant to the light factors. However, the more important part is the early visual-somatosensory interaction in the BSV associated learning which improves the precision significantly even in the harsh environments (Supplementary Fig. 11), due to the collection of coherent information to reduce the perceptual ambiguity[28]. Such a tolerant learning system with improved

precision in sensing, perception and feedback are critical for various human-machine applications, particularly in healthcare and augmented reality.

Conclusions

We have reported a learning architecture that integrates visual and somatosensory information to achieve high-precision hand gesture recognition. We fabricated an adhesive stretchable strain sensor that can be conformably attached to human skin in order to acquire reliable somatosensory information of hand gestures, and used a commercial camera sensor to obtain images. The transparent (89%) strain sensor was inconspicuous in the images. Compared to unisensory and common SV fusion architectures, our BSV learning that employs a pruning strategy based on a Frobenius condition number, achieved a superior accuracy (100% using a custom dataset) for hand gesture recognition against complex backgrounds. The BSV fusion process can tolerate undesirable features in the visual information, such as noise and underexposure or overexposure. As a result, the approach can achieve a high recognition accuracy (over 96.7%) even in the dark. The learning architecture was also implemented for the control and navigation of a quadruped robot using hand gestures. Our work illustrates that the integration of skin-like electronics with computer vision can significantly improve the precision of HGR, even under harsh environmental scenarios, and is promising for recognition tasks in human–machine interaction applications.

Methods

Fabrication of stretchable strain sensor device. Single-walled carbon nanotubes (SWCNTs) and polydimethylsiloxane (PDMS) were purchased from Carbon Solution and Sigma-Aldrich, respectively. Acrylic acid, N,N'-methylenebisacrylamide (BIS, crosslinker), potassium persulfate (KPS, thermal initiator) and N,N,N',N'-tetramethylethylenediamine (TEMED, co-initiator) were purchased from Sigma-Aldrich. *(1) Fabricated the stretchable strain sensor without the PAA hydrogel layer.* SWCNTs were firstly dispersed in deionized water by sonicating for 2 h, resulting in the SWCNT solution of 0.1 mg/ml. 120 μL of SWCNT solution (total usage amount of 3 strain sensors) was further diluted into a 40 ml deionized water to form the resultant solution for vacuum filtration. The filter membrane was patterned to three rectangle patterns (each pattern is 2×0.4 cm^2). After filtering, the SWCNT layer was transferred to a half-cured PDMS film (first curing) that was used to enhance the bonding strength between the SWCNT and PDMS. This half-cured PDMS was obtained by the PDMS precursors (mixed in a weight ratio of 10:1, and spin-coated onto a hydrophobization-treated Si wafer at 800 rpm for 60 s) being cured in an oven at 60 °C for 40 min. After transferring, the wafer was placed in the oven at 60 °C for 8 h to further cure the PDMS (second curing). Then, we achieved the high transparent stretchable strain sensor with a patterned and uniform SWCNT sensing layer. *(2) PAA polymerization on the PDMS substrate*. Before the PAA polymerization, a plasma polymerization procedure was carried out to chemically modify the PDMS surface (opposite side of SWCNT) by using the treatment with the

argon gas (3 mbar, 1 min), and followed a mixture of argon and acrylic acid gas (4 mbar, 4 min). Then, acrylic acid (1 ml), BIS (2.5 mg), TEMED (35 μL), KPS (175 mg) were subsequently added to 5 mL deionized water. The resultant solution was dropped onto the modification PMDS surface to implement the thermal polymerization onto a hotplate at 70 ˚C for 20 min. After thermal polymerization, a stretchable strain sensor with PAA hydrogel adhesion layer was completed for capturing the somatosensory signals.

Characteristics of stretchable strain sensor. The strain sensing characteristics of stretchable strain sensor devices were performed using an electrical measurement equipment (Keithley 4200 semiconductor device parameter analyzer) and a mechanical measurement equipment (MTS Criterion Model 42). The measurement of optical transmittance is carried out by an UV-2550.

Dataset preparation. Our SV dataset totally contains 3000 SV samples. Each SV sample consists of one hand gesture image with a complex background (corresponding to visual information), and five strain sensor data (corresponding to somatosensory information). The hand gesture images were captured by the commercial-off-the-shelf camera sensors. The corresponding five strain sensor data were simultaneously obtained by our fabricated stretchable strain sensors that were patched on the five fingers of a human hand. In order to guarantee the generalization ability of the dataset, there are totally 10 volunteers and 80 strain sensors employed to collect the somatosensory and visual information in consideration of the individual differences and the device variation of strain sensors. The raw images captured by

cameras were resized to 160 × 120 pixels. The resistance change ($\Delta R/R_0$) of strain sensors was further processed by a normalization step, regarding as the somatosensory information.

Data availability. The data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request. The SV datasets that are used in this study are available at https://github.com/mirwang666-ime/Somato-visual-SV-dataset.

Code availability. The code that supports the plots within this paper and other finding of this study is available at https://github.com/mirwang666-ime/Somato-visual-SV-dataset. The code that supports the human-machine interaction experiment is available from the corresponding author upon reasonable request.

References

1. Yamada, T. et al. A stretchable carbon nanotube strain sensor for human-motion detection. *Nat. Nanotechnol.* **6**, 296-301 (2011).
2. Amjadi, M., Kyung, K.-U., Park, I. & Sitti, M. Stretchable, skin-mountable, and wearable strain sensors and their potential applications: a review. *Adv. Funct. Mater.* **26**, 1678-1698 (2016).
3. Rautaray, S.S. & Agrawal, A. Vision based hand gesture recognition for human computer interaction: a survey. *Artif. Intell. Rev.* **43**, 1-54 (2015).

4. Lim, S. et al. Transparent and stretchable interactive human machine interface based on patterned graphene heterostructures. *Adv. Funct. Mater.* **25**, 375-383 (2015).

5. Pisharady, P.K., Vadakkepat, P. & Loh, A.P. Attention based detection and recognition of hand postures against complex backgrounds. *Int. J. Comput. Vis.* **101**, 403-419 (2013).

6. Giese, M.A. & Poggio, T. Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* **4**, 179-192 (2003).

7. Tan, X. & Triggs, B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* **19**, 1635-1650 (2010).

8. Liu, H., Ju, Z., Ji, X., Chan, C.S. & Khoury, M. *Human motion sensing and recognition*. (Springer, Berlin, Germany, 2017).

9. Liu, K., Chen, C., Jafari, R. & Kehtarnavaz, N. Fusion of inertial and depth sensor data for robust hand gesture recognition. *IEEE Sens. J.* **14**, 1898-1903 (2014).

10. Chen, C., Jafari, R. & Kehtarnavaz, N. A survey of depth and inertial sensor fusion for human action recognition. *Multimed. Tools Appl.* **76**, 4405-4425 (2017).

11. Dawar, N., Ostadabbas, S. & Kehtarnavaz, N. Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. *IEEE Sensors Lett.* **3**, 7101004 (2019).

12. Kwolek, B. & Kepski, M. Improving fall detection by the use of depth sensor and accelerometer. *Neurocomputing* **168**, 637-645 (2015).

13. Tang, D., Yusuf, B., Botzheim, J., Kubota, N. & Chan, C.S. A novel multimodal communication framework using robot partner for aging population. *Expert Syst. Appl.* **42**, 4540-4555 (2015).

14. Wang, C., Wang, C., Huang, Z. & Xu, S. Materials and structures toward soft electronics. *Adv. Mater.* **30**, 1801368 (2018).

15. Kim, D.H. et al. Dissolvable films of silk fibroin for ultrathin conformal bio-integrated electronics. *Nat. Mater.* **9**, 511-517 (2010).

16. Ehatisham-Ul-Haq, M. et al. Robust human activity recognition using multimodal feature-level fusion. *IEEE Access* **7**, 60736 - 60751 (2019).

17. Imran, J. & Raman, B. Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition. *J. Amb. Intel. Hum. Comp.*, 1–20 (2019).

18. Dawar, N. & Kehtarnavaz, N. Action detection and recognition in continuous action streams by deep learning-based sensing fusion. *IEEE Sens. J.* **18**, 9660-9668 (2018).

19. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).

20. Wang, M., Wang, T., Cai, P. & Chen, X. Nanomaterials discovery and design through machine learning. *Small Methods* **3**, 1900025 (2019).

21. Li, S.-Z., Yu, B., Wu, W., Su, S.-Z. & Ji, R.-R. Feature learning based on

SAE–PCA network for human gesture recognition in RGBD images. *Neurocomputing* **151**, 565-573 (2015).

22. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017).

23. Long, E. et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat. Biomed. Eng.* **1**, 0024 (2017).

24. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-489 (2016).

25. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354-359 (2017).

26. Chandrasekaran, C., Lemus, L. & Ghazanfar, A.A. Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. *Proc. Natl. Acad. Sci. USA* **110**, E4668-E4677 (2013).

27. Lakatos, P., Chen, C.M., O'Connell, M.N., Mills, A. & Schroeder, C.E. Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* **53**, 279-292 (2007).

28. Henschke, J.U., Noesselt, T., Scheich, H. & Budinger, E. Possible anatomical pathways for short-latency multisensory integration processes in primary sensory cortices. *Brain Struct. Funct.* **220**, 955-977 (2015).

29. Lee, A.K.C., Wallace, M.T., Coffin, A.B., Popper, A.N. & Fay, R.R. Multisensory processes: The auditory perspective. (Springer, 2019).

30. Bizley, J.K., Jones, G.P. & Town, S.M. Where are multisensory signals combined for perceptual decision-making? *Curr. Opin. Neurobiol.* **40**, 31-37 (2016).

31. Ohyama, T. et al. A multilevel multimodal circuit enhances action selection in Drosophila. *Nature* **520**, 633-639 (2015).

32. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186-198 (2009).

33. Yamins, D.L. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA* **111**, 8619-8624 (2014).

34. Gilbert, C.D. & Li, W. Top-down influences on visual processing. *Nat. Rev. Neurosci.* **14**, 350-363 (2013).

35. Chortos, A., Liu, J. & Bao, Z. Pursuing prosthetic electronic skin. *Nat. Mater.* **15**, 937-950 (2016).

36. Barbier, V. et al. Stable modification of PDMS surface properties by plasma polymerization: application to the formation of double emulsions in microfluidic systems. *Langmuir* **22**, 5230-5232 (2006).

37. Bakarich, S.E. et al. Recovery from applied strain in interpenetrating polymer network hydrogels with ionic and covalent cross-links. *Soft Matter* **8**, 9985 (2012).

38. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

39. Krizhevsky, A., Sutskever, I. & Hinton, G.E. in Adv. Neural Inf. Process. Syst. 1097–1105 (2012).

40. Rockova, V. & Polson, N. in Adv. Neural Inf. Process. Syst. 930-941 (2018).

41. Le, X. & Wang, J. Robust pole assignment for synthesizing feedback control systems using recurrent neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **25**, 383-393 (2013).

Acknowledgements

The project was partially supported by the National Research Foundation (NRF), Prime Minister's office, Singapore, under its NRF Investigatorship (NRF2016NRF-NRF1001-21), Singapore Ministry of Education (MOE2015-T2-2-60), Advanced Manufacturing and Engineering (AME) Programmatic Grant (No. A19A1b0045), and the Australian Research Council (ARC) under Discovery Grant DP200100700. The authors thank all the volunteers for collecting data and also thank Dr. Ai Lin Chun for critically reading and editing the manuscript.

Author contributions

M.W. and X.C. designed the study. M.W. designed and characterized the strain sensor. M.W., T.W. and P.C. fabricated the PAA hydrogels. Z.Y. and M.W. conducted the machine learning algorithms and analyzed the results. M.W., S.G, and Y.Z. collected the SV data. M.W. performed the human-machine interaction experiment. M.W. and X.C. wrote the paper and all authors provided feedback.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available in the online version of the paper.

Reprints and permissions information is available online at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to X.C.

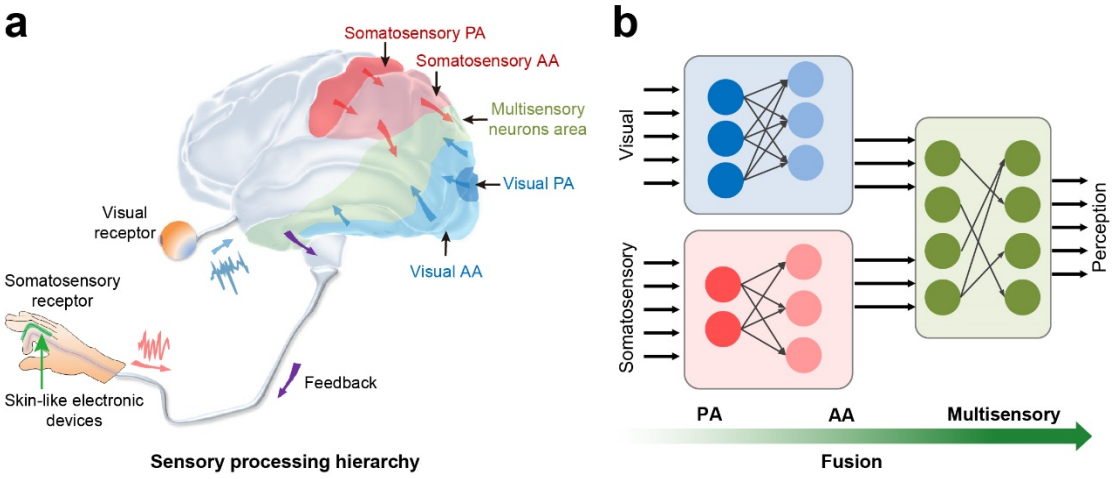

Fig. 1 | BSV associated learning framework. **a,** Schematic showing the processing hierarchy of visual and somatosensory information in the human brain. Red, blue and green areas represent the processing region (PA: primary area; AA: association area) for somatosensory, visual and multimodal information, respectively. Red and blue arrows respectively represent the direction of somatosensory and visual information flow; purple arrows show the feedback information flow after multimodal fusion. Visual and somatosensory information interact early in the multisensory neurons area. Skin-like electronic devices act as somatosensory receptors. **b,** BSV associated learning framework consists of three neural networks that mimic the SV fusion process hierarchy in **a**. Top left: a sectional CNN representing early visual processing; bottom left: a multilayer neural network mimicking the early somatosensory processing; right: a sparse neural network resembling the high level fusion of early interactions in the whole BSV architecture.

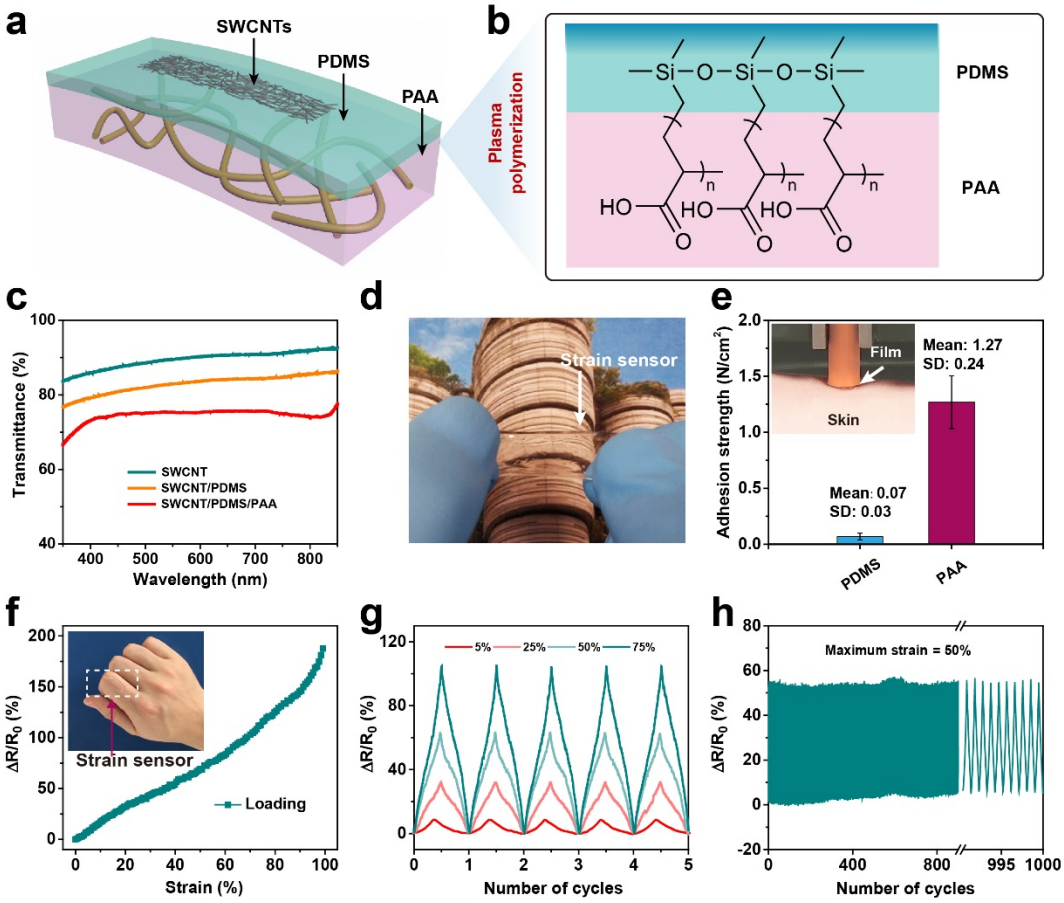

Fig. 2 | Characterization of stretchable strain sensor. **a,** Schematic showing the multilayer stacked structure of the conformable, transparent, and adhesive stretchable strain sensor. SWCNTs, PDMS, PAA hydrogels function as the sensing layer, the stretchable substrate, and the adhesive layer, respectively. **b,** Plasma treatment with argon and acrylic gas chemically modifies the PDMS surface with acrylic acid, allowing PAA hydrogels to bond strongly with PDMS. **c,** Transmittance spectra of the pure SWCNT film, the SWCNT/PDMS device, and the SWCNT/PDMS/PAA device in the visible wavelength range from 350 to 850 nm. **d,** Photograph showing the stretchable strain sensor is transparent. **e,** Adhesion strength of PAA hydrogels (magenta) on human skin was much higher than PDMS (blue). Inset shows the measurement setup. The error bars represent the standard deviation (SD) of

adhesion strength among ten measurements. **f,** Strain-resistance response curves of the stretchable strain sensor under loading. Inset shows a strain sensor patched over the knuckle of an index finger. **g,** Strain-resistance response curves under a triangular strain profile show the sensor is stable and can undergo various dynamic loads. Magnitudes of the respective peak strains are 5%, 25%, 50%, and 75%. **h,** Durability test of 1000 cycles at 50% strain shows the strain sensor response is stable and regular with a nearly constant base resistance.

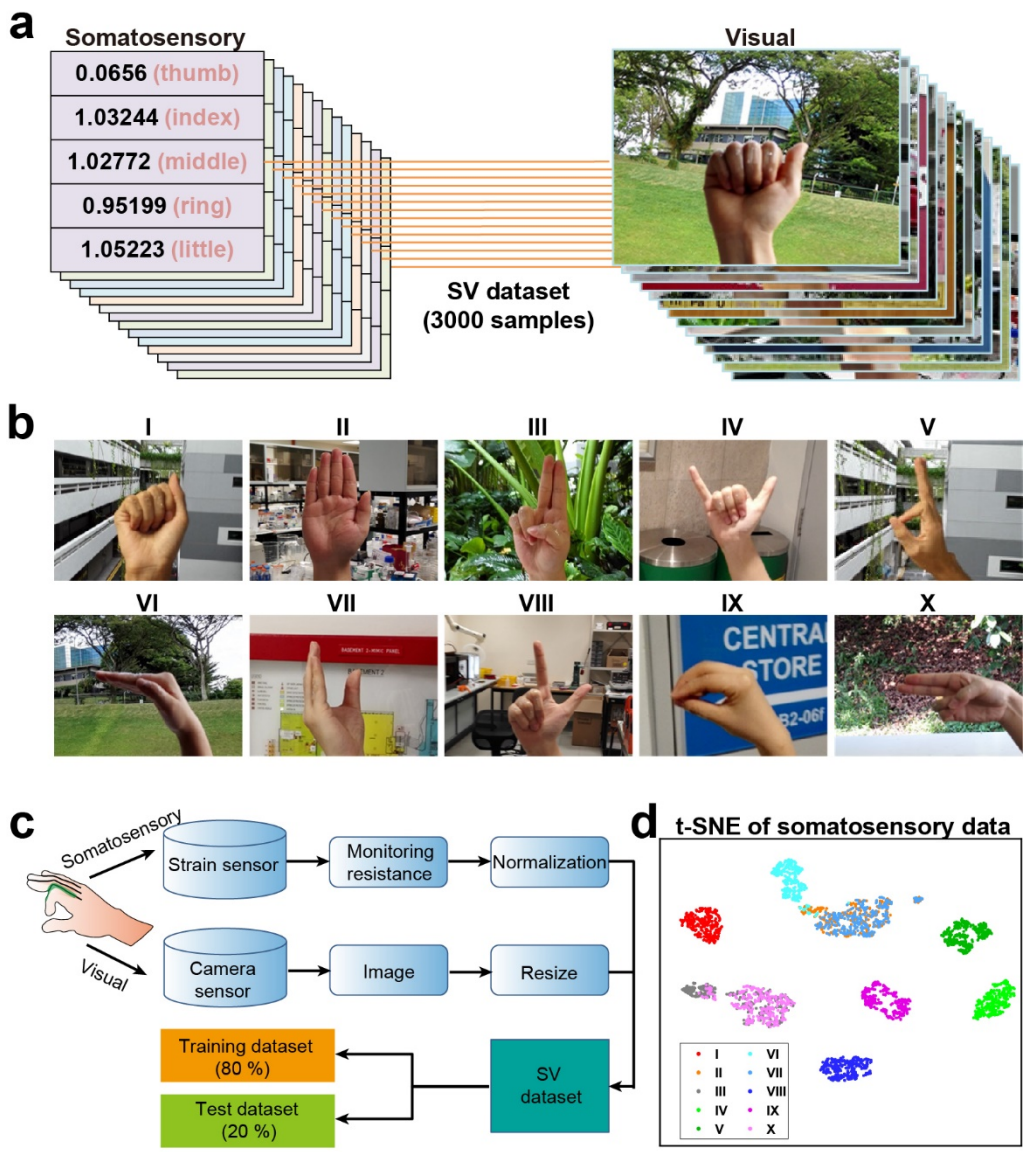

Fig. 3 | Dataset preparation for BSV associated learning. **a,** Illustration of the SV dataset containing 3000 SV samples. Each SV sample consists of one image of a hand gesture and the corresponding somatosensory information represented by strain data captured from 5 strain sensors patched over the knuckle of the thumb, index, middle, ring and little finger. **b**, Photographs showing the 10 categories (I to X) of hand gestures in the SV dataset. **c,** Flow diagram for the SV data collection. Normalization of strain data and resizing the images are required to structure the somatosensory and visual signals. **d,** Visualizing the somatosensory information

within the 3000 samples in the SV dataset using t-SNE dimensionality reduction. Each point represents somatosensory information of one hand gesture projected from the 5-dimensional strain data into two dimensions. Similar gestures clustered together, forming 10 categories of hand gestures in the entire SV dataset.

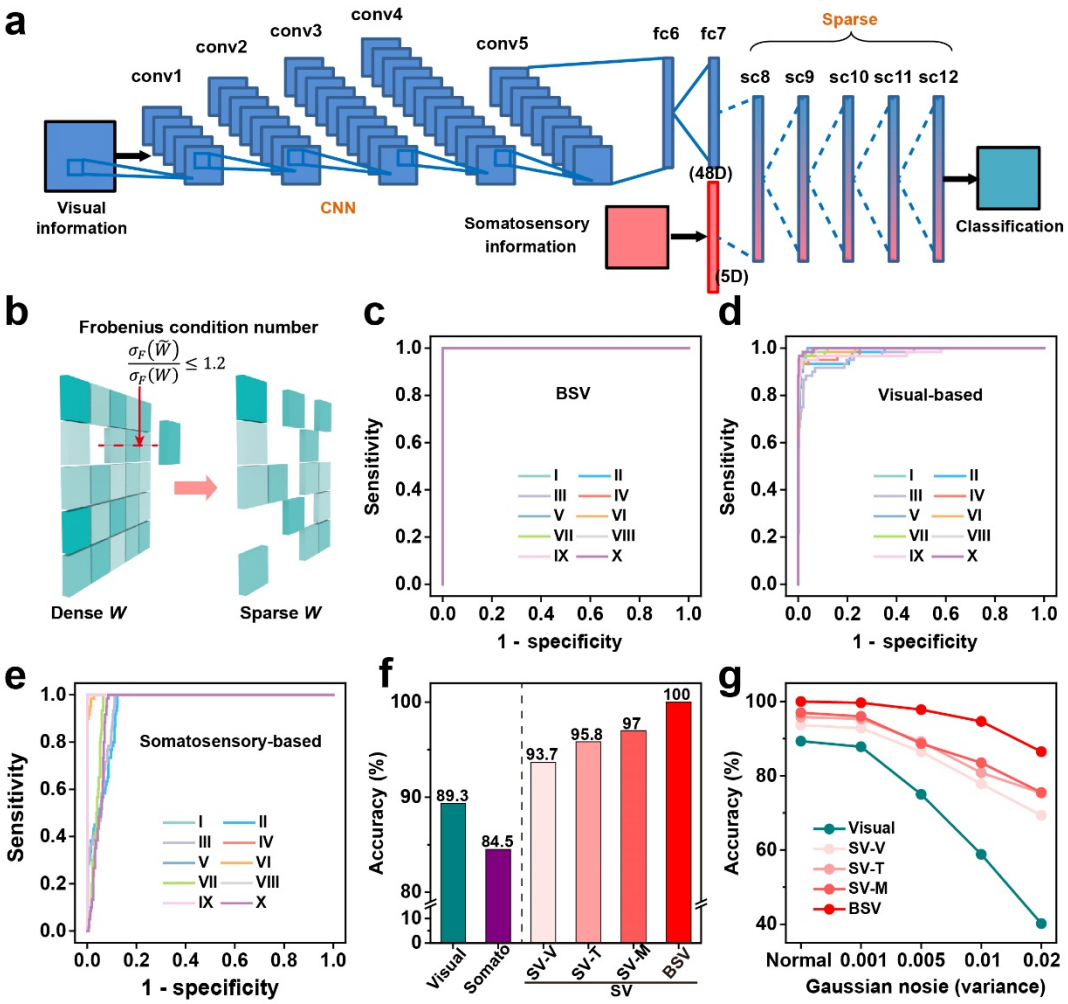

Fig. 4 | BSV associated learning for classification. **a**, Schematic showing how visual and somatosensory information is processed and fused in the BSV associated learning architecture. conv.: convolutional layer; fc: fully connected neural network; sc: sparsely connected neural network; 48D: 48-dimensional visual vector; 5D: 5-dimensional somatosensory vector. Blue boxes represent the AlexNet CNN, which was pretrained using the ImageNet dataset. The 5-layer sparse neural network (sc8 to sc12) was used to fuse the SV information. **b**, Schematic of Frobenius condition number-dependent pruning strategy. W: Weighting matrix; $\widetilde{W}$: W after pruning; $\sigma_F(W)$: Frobenius condition number of W. Color lightness represents the value of weights. Dense W (Left) and sparse W (Right) between two adjacent neuron

layers. **c-e**, ROC curves of the BSV, visual-, and somatosensory-based recognition for 10 categories of hand gestures (I to X). **f**, BSV associated learning showed the best accuracy among the unimodal (visual- and somatosensory-) and multimodal fusion strategies (SV-V, SV-T, SV-M, BSV). Unimodal strategies: visual-based recognition using only visual images and somatosensory-based recognition using only strain sensor data. Multimodal fusion strategies using both visual images and strain sensor data: weighted-average fusion (SV-V), weighted-attention fusion (SV-T), weighted-multiplication fusion (SV-M), and BSV associated learning fusion. The final recognition accuracies are 89.3%, 84.5%, 93.7%, 95.8%, 97% and 100% for visual, somato, SV-V, SV-T, SV-M, and BSV based strategies, respectively. **g**, Testing results of visual, SV-V, SV-T, SV-M, and BSV based strategies under defective visual information with various Gaussian noises (0.001, 0.005, 0.01, 0.02) show only BSV can maintain its high recognition accuracy with increased noise level.

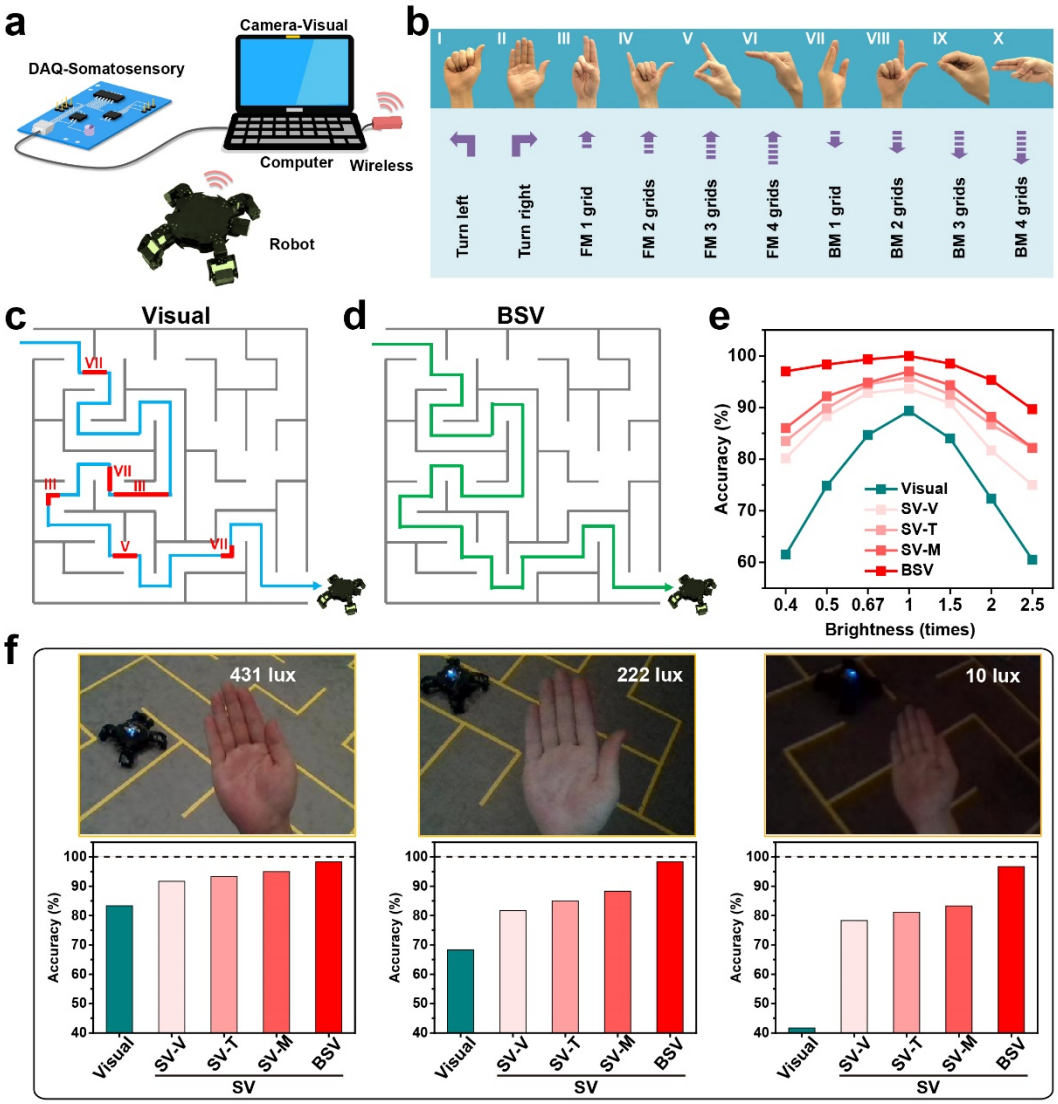

Fig. 5 | Precise HGR based on BSV for human-machine interaction. **a**, Schematic shows the system consists of a somatosensory-data acquisition unit (DAQ), a camera for capturing visual image, a computer, a wireless data transmission module, and a quadruped robot. **b**, Each of the 10 categories (I to X) of hand gestures was assigned a specific motor command to guide the movement of the quadruped robot. Forward move: FM; Back move: BM. **c,d,** Scenarios of the robot walking through the labyrinth based on visual-based recognition (**c**) and BSV associated learning recognition (**d**). Red sectors represent errors in the recognition while Roman

numerals show the predicted hand gesture categories. Visual-based recognition made more errors than BSV recognition. **e**, Testing results using the previous testing dataset, now with images mathematically processed to adjust the brightness. Visual images were processed for underexposure (brightness < 1) or overexposure (brightness > 1). No treatment (brightness = 1). **f**, Performance accuracy of the robot using different recognition architectures under different illuminances (431, 222, and 10 lux). Below each image of the three light conditions are subplots showing the statistics for 60 testing trials for the five approaches (Visual, SV-V, SV-T, SV-M, and BSV). Robots using BSV architecture maintained high accuracy even in the dark.