LETTER

https://doi.org/10.1038/s41586-019-1234-z

Learning the signatures of the human grasp using a scalable tactile glove

Subramanian Sundaram[1,2,3,4]*, Petr Kellnhofer[1,2], Yunzhu Li[1,2], Jun-Yan Zhu[1,2], Antonio Torralba[1,2] & Wojciech Matusik[1,2]

Humans can feel, weigh and grasp diverse objects, and simultaneously infer their material properties while applying the right amount of force—a challenging set of tasks for a modern robot[1]. Mechanoreceptor networks that provide sensory feedback and enable the dexterity of the human grasp[2] remain difficult to replicate in robots. Whereas computer-vision-based robot grasping strategies[3–5] have progressed substantially with the abundance of visual data and emerging machine-learning tools, there are as yet no equivalent sensing platforms and large-scale datasets with which to probe the use of the tactile information that humans rely on when grasping objects. Studying the mechanics of how humans grasp objects will complement vision-based robotic object handling. Importantly, the inability to record and analyse tactile signals currently limits our understanding of the role of tactile information in the human grasp itself—for example, how tactile maps are used to identify objects and infer their properties is unknown[6]. Here we use a scalable tactile glove and deep convolutional neural networks to show that sensors uniformly distributed over the hand can be used to identify individual objects, estimate their weight and explore the typical tactile patterns that emerge while grasping objects. The sensor array (548 sensors) is assembled on a knitted glove, and consists of a piezoresistive film connected by a network of conductive thread electrodes that are passively probed. Using a low-cost (about US$10) scalable tactile glove sensor array, we record a large-scale tactile dataset with 135,000 frames, each covering the full hand, while interacting with 26 different objects. This set of interactions with different objects reveals the key correspondences between different regions of a human hand while it is manipulating objects. Insights from the tactile signatures of the human grasp—through the lens of an artificial analogue of the natural mechanoreceptor network—can thus aid the future design of prosthetics[7], robot grasping tools and human–robot interactions[1,8–10].

Humans effortlessly manipulate objects and tools by applying precisely controlled forces[11–13]. To understand the tactile feedback involved in the human grasp, we can use emerging machine learning tools to attempt to distil high-level properties and relationships from high-dimensional tactile data. Such tools require large-scale tactile datasets with high spatial resolution. However, large tactile datasets of human grasps covering the full hand do not exist because densely covering the human hand with tactile sensors is challenging. These tactile sensors come with strict requirements for the form-factor, resolution and mechanical compliance. Whereas electronic skins have made progress on the compliance requirements[14], an electronic tactile glove with dense coverage and capable of collecting large datasets has yet to be demonstrated. The Tekscan Grip system (with its 349 sensors) is the closest high-cost commercially available system, but does not fully cover the hand (details and a comparative list are included in the Supplementary Information). Current high-resolution optical tactile sensors[15,16] and biomimetic multimodal sensor integrations[17,18] have not successfully mapped a full human hand. Broadly, hurdles in creating a scalable tactile feedback network and acquiring large tactile datasets covering the hand have impeded our fundamental understanding of the human grasp.

We first present a simple method of fabricating a low-cost, scalable tactile glove (STAG) covering the full hand with 548 sensors. The STAG can record tactile videos (with frame rate approximately 7.3 Hz), measuring normal forces in the range 30 mN to 0.5 N (with quantization of about 150 levels and a peak hysteresis of about 17.5%). Importantly, the device can be constructed with low-cost materials (around US$10) and be used over long intervals. The STAG can be translated to a variety of different designs (see below). We introduce a large-scale dataset of tactile maps (135,000 frames) recorded using the STAG while manipulating objects with a single hand; see Methods for dataset acquisition conditions. The spatial correlations and correspondence between finger regions that emerge from the dataset represent the tactile signatures of the human grasping strategy (Fig. 1a). Here we observe and learn from successful daily human–object interactions with the long-term goal of aiding the development of robots and prosthetics.

The similarities in the underlying shape perception primitives between the visual and tactile domains are known[19]. We therefore hypothesized, on the basis of visual perception studies (showing that 16×16 pixels were sufficient for face recognition[20] and 32×32 pixels for scene recognition[21] in visual data), that a similar minimal sensor count is suitable for a tactile sensor. The STAG consists of a sensing sleeve with 548 sensors attached on top of a custom knit glove. Figure 1b shows the locations of the 548 sensors and the 64 electrodes (fabrication details are included in the Methods). Fabricated gloves are shown in Fig. 1a and Extended Data Fig. 1a (see Extended Data Fig. 1b for a high-resolution scan of the glove). The sensor array consists of a force-sensitive film (0.1 mm thick) addressed by a network of orthogonal conductive threads (0.34 mm) on each side, insulated by a thin adhesive (0.13 mm) and a low-density polyethylene (LDPE) film (about 13 μm). Each point of overlap between the orthogonal electrodes is sensitive to normal force, modulating the electrical resistance through the force-sensitive film. The force-sensitive film is laser-cut to fit the custom knit glove (yellow) along with holes to guide thread placement, and slots at the finger joints. The sensor laminate is thin and mechanically flexible (three-point bending test results are shown in Supplementary Fig. 1; Supplementary Video 1 demonstrates the compliance visually). The typical force response of a single sensing element (Fig. 1c), measured as the through-film resistance, changes from about 4 kΩ (unloaded) to below 2 kΩ (at a 0.5 N normal load). Each sensing element is sensitive to small forces (starting at about 25 mN; Extended Data Fig. 2a) and saturates beyond 0.8 N. The force response in the working range (30 mN to 0.5 N) is consistent across multiple devices (Extended Data Fig. 2b) and over multiple cycles (1,000 cycle tests in Extended Data Fig. 2c, d). The sensor elements show a stable resistance up to 60 °C and become insulating at temperatures above 80 °C (differential scanning calorimetry and resistance measurements are shown in Extended Data Fig. 2e, f).

We use a modified version of a grounding-based electrical isolation scheme[22] (including charging resistors to improve the readout speed)

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Present address: Biological Design Center, Boston University, Boston, MA, USA. [4]Present address: Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA. *e-mail: subras@csail.mit.edu

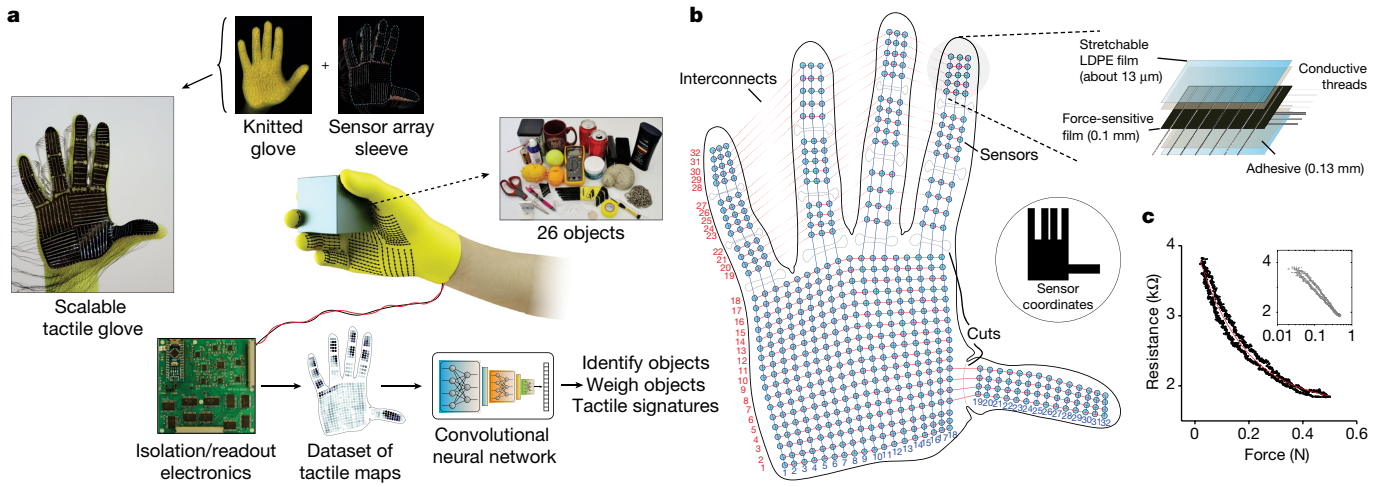


Fig. 1 | The STAG as a platform to learn from the human grasp. a, The STAG consists of a sensor array with 548 elements covering the entire hand, attached to a custom knit glove. An electrical readout circuit is used to acquire the normal force recorded by each sensor at approximately 7.3 fps. Using this setup allows us to record a dataset of 135,187 tactile maps while interacting with 26 different objects. A deep convolutional neural network trained purely on tactile information can be used to identify or weigh objects and explore the tactile signatures of the human grasp. The glove shown at the centre is a rendering. **b**, The design of the STAG architecture shows the individual locations of the 548 sensors, along with the interconnects, slot and 64 electrodes. The piezoresistive sensor array is fabricated by laminating simple materials and can be extended to different architectures easily (Extended Data Figs. 3 and 5). **c**, Each sensor element responds to normal force by exhibiting a change in the through-film resistance. The sensor characteristics are repeatable across multiple devices and reliable over the long term (Extended Data Fig. 2). The inset shows the same characteristics in a logarithmic force scale (axes labels are as for the main plot).

to extract individual sensor measurements (the readout circuit costs about US$100; see Methods for fabrication details). The circuit topology is shown in Extended Data Fig. 1c along with the fabricated printed circuit-board image (Extended Data Fig. 1d). The sensor response at the output of the amplifier (and the analog-to-digital converter, ADC) is linear with respect to the force (see Supplementary Fig. 2). We note that the STAG design can be simplified to rapidly fabricate regular arrays; Extended Data Fig. 3 shows 1,024-element sensors with sensor spacing of 2.5 mm. Such regular arrays fixed on flat surfaces can record the resting identities of different objects (Extended Data Fig. 4). Furthermore, despite the weak extensibility of the force-sensitive film, we can enhance the achievable stretchability by incorporating auxetic designs[23] into the sensor structure as shown in Extended Data Fig. 5. The auxetic prototype with 10×10 elements can be stretched in multiple directions, as well as folded or crushed (Extended Data Fig. 5e, f; see Supplementary Video 2).

The reliability of our STAG prototype allows us to record tactile videos (and corresponding visual images for illustration) during interactions with a set of 26 objects (Extended Data Fig. 6) with a single hand over many hours (total length of recordings exceeding 5 hours; see Methods for dataset acquisition details). A sample set of interactions from our dataset are shown in Supplementary Videos 3–5. We identify the specific frames in which objects are in contact with the glove (see Methods for filtering procedure). We train a convolutional neural network (CNN) to identify objects using these filtered frames (32×32 arrays in sensor coordinates). We use a ResNet-18-based architecture[24] that takes N input frames (Fig. 2a; see network implementation in the Methods). The classification accuracy improves with the number of inputs and reaches its maximal performance with about seven random input frames (Fig. 2b). This is expected, because multiple contacts with an object help to identify it more accurately. Figure 2c shows eight example tactile frames along with their output classification vectors (the expanded version is shown in Supplementary Fig. 3). Here we observe that it is easy to identify the mug when it is held by the handle but it can be confused with a can or other objects when lifted from the sides. Likewise, the elongated shape of the pen is easier to identify when it is in contact with the palm than when it is held between fingers. Interestingly, when a mug is held by the handle (or while a spray can is being held), the distinct hand pose captured in the tactile map from sensors around the joints (proprioceptive data) may also help in object classification.

The first 3×3 convolution filters learned by our network are shown in Extended Data Fig. 7i. To understand the features at a higher resolution, we scaled the input resolution by three and adapted the network elements appropriately (see Methods). The first layer convolution filters learned by the adapted network are shown in Fig. 2d. The network primarily learns blob-like point detectors, edge detectors and low-frequency filters. The visual domain filters learned by standard ResNet-18 trained on the ImageNet[25] dataset are shown in Extended Data Fig. 7j for comparison. Furthermore, we visualized the features of our trained network (with Network Dissection; see Methods for details) and observed that the early convolution layers are activated in small regions. The higher-layer convolution filters are often activated by more complex grasp-related concepts; Supplementary Fig. 4b shows the activation maps of filters that respond to larger contact patterns, or when specific hand regions are used.

Humans are easily capable of associating similar grasps based on motor movements, and the identification of an object is probably better performed when choosing the most distinct (informative) set of grasps. Motivated by this, instead of choosing N random frames, we identified the most diverse set of N frames for an input recording by k-means clustering (an example with $N = 5$ clusters is shown in Fig. 2e; see an interactive version of the map in Supplementary Data 1). The classification accuracy using N input frames (one from each cluster) shows that clustering provides a marginal improvement in accuracy when a small set of inputs is used ($N < 4$ in Fig. 2b). We note that the results of clustering-based inputs converge with those of randomly chosen inputs for large N because a random selection captures the data well when N is large. The corresponding confusion matrices are shown in Extended Data Fig. 7a–h. We observe that objects with similar shapes, sizes or weights are more likely to be confused with one another. Light objects such as the safety glasses, the plastic spoon or the coin are more easily misclassified, whereas large, heavy objects with distinct signatures, like the tea box, can easily be detected even with a small number of input frames.

The object identification tests described above help to evaluate the capability of the STAG in capturing useful data. We also evaluated the classification performance of lower sensor counts by downsampling the tactile data either uniformly or based on different regions of the hand. The classification accuracy drops considerably as the effective number of sensors is reduced, thereby highlighting the need for a high sensor

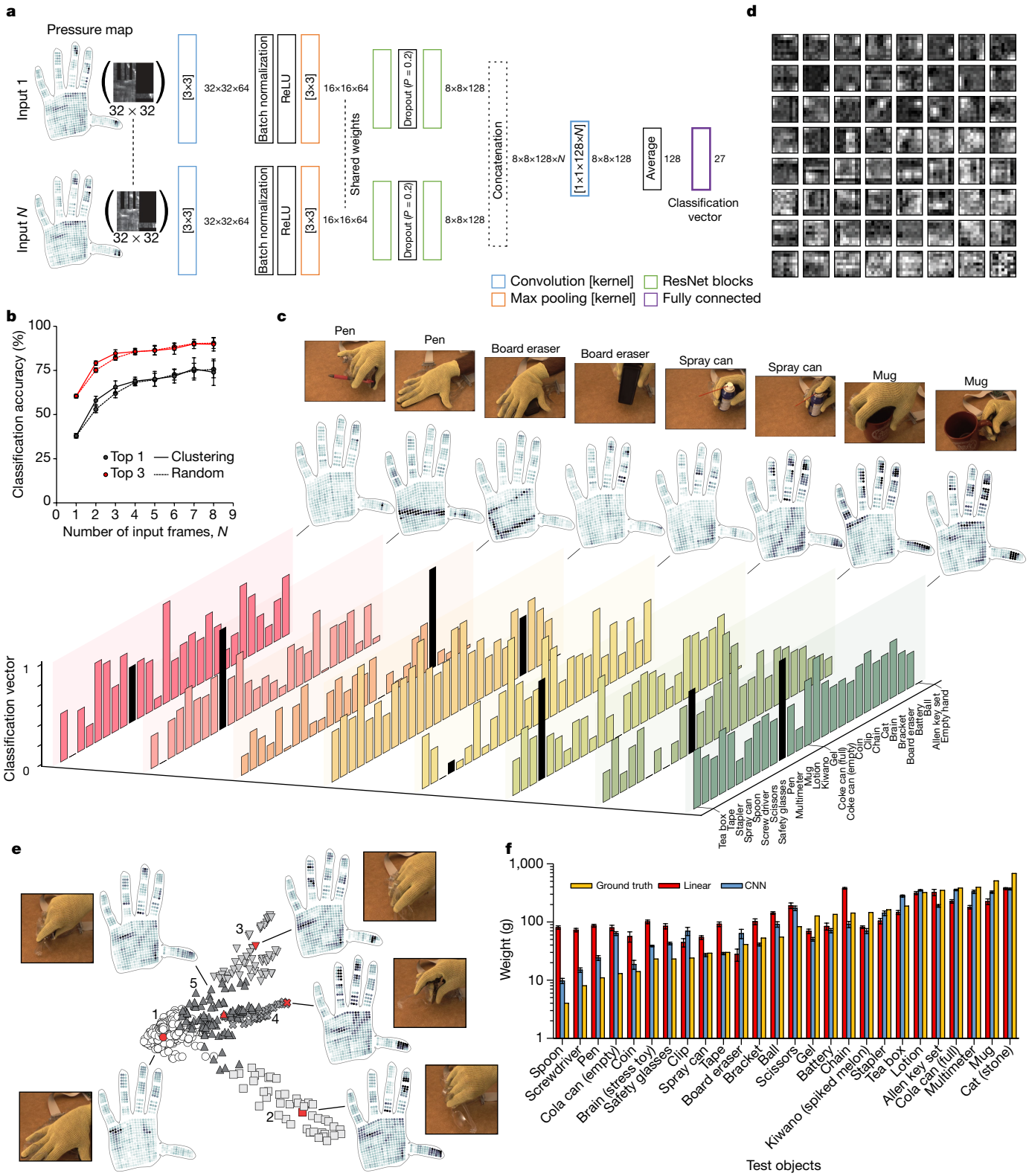


Fig. 2 | Identifying and weighing objects from tactile information. a, The CNN architecture used for identifying objects from tactile information takes as input *N* arrays of tactile data (32 × 32 arrays). Rectified Linear Units (ReLU) are used to introduce non-linearity into the model. 'Dropout' is a regularization technique that randomly drops out nodes of the network to reduce overfitting to the training data. 'Max pooling' is used to reduce dimensionality of the data by passing only the locally highest activations. **b**, The object identification accuracy is enhanced when using a diverse set of tactile maps from *N* distinct clusters as input when compared to a random choice of inputs; results are averaged over ten training runs (mean ± s.d.). Here each distinct cluster (as shown in **e**) is a group of similar grasps. **c**, A representative set of examples during single-hand manipulation of objects. Tactile maps, corresponding visual images and the classification vectors (bottom) from single tactile map inputs are shown—the ground-truth object labels are marked in black (see expanded version in Supplementary Fig. 3). **d**, The convolution filters learned by the scaled version of the network are shown (see Methods). Note that the inputs and the network are scaled to visualize the filters at a higher resolution. The original 3 × 3 convolution filters of the network in **a** and the original ImageNet-trained ResNet filters are shown in Extended Data Fig. 7i, j respectively. **e**, Clustering tactile maps from a single object interaction helps to identify a diverse set of tactile maps that correspond to that object. Five different clusters are used to extract five tactile inputs (*N*) shown in red (see interactive map in Supplementary Data 1). **f**, 'Leave-one-out'-based CNN weight prediction results compared with a linear model (mean ± 95% confidence interval).

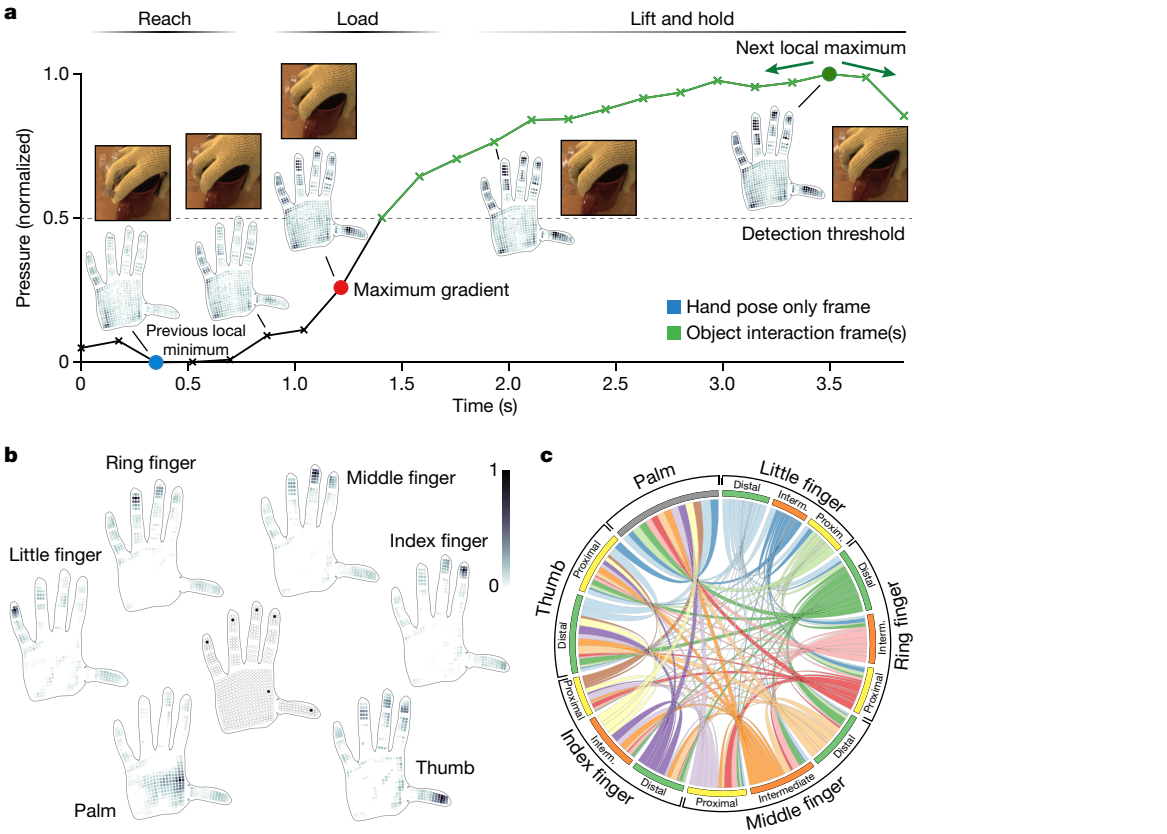


Fig. 3 | Cooperativity among regions of the hand during object manipulation and grasp. a, In a typical interaction sequence, the hand increasingly gets articulated until the point of contact (reach phase) and experiences a sudden rise in tactile forces as the object is held. We use the maximum gradient (red) in the load phase of the action to decompose a tactile map of when the object is held into two parts—the hand pose signal (blue) and the object-related pressure (blue frame subtracted from the green frames). The object interaction frames are automatically detected and marked in green; full signal decomposition details are described in the Methods. **b**, The set of decomposed object-related pressure frames can be used to extract correlations between specific sensors and the full hand. The maps here show the correlations between select pixels (marked in the central hand) and the full hand. Most fingertips are used with other fingers and the thumb (a signature of precision grasp), whereas the regions on the palm are typically used when grasping objects that cause the full palm to come into contact with the object (see Supplementary Data 2). **c**, The circular plot shows the relative correspondences between different parts of the hand (see Methods; Supplementary Data 2 also contains interactive maps of region-level and finger-level correspondences). The distal phalanges of the large fingers are usually used with the thumb to generate forces while gripping an object and result in strong concurrence.

count (classification performance and the effective receptive fields are shown in Supplementary Fig. 5).

In addition to identifying objects, humans can easily estimate the weight of objects from tactile signals. The ability to estimate weights is of practical use in robotics and has been the focus of human perception experiments[26]. To estimate the weights of objects from tactile interactions, we used a restricted dataset of multi-fingered grasps where the object was picked up from above (sample recording in Supplementary Video 6). After an object is picked up, a single frame is used as an input to a CNN to predict its weight. Note that the training and test data have disjoint sets of objects (see Methods). Extended Data Figure 8 shows a representative set of tactile frames and corresponding images. Results in Fig. 2f show that our network performs better than a naive linear model over the entire weight range.

We looked at the typical sequence of tactile maps immediately before and after an object is grasped (an example is shown in Fig. 3a) to understand the grasp in depth. The hand is increasingly articulated to fit the object closely, during which time the proprioceptive signal in the tactile map increases gradually until contact during the 'reach' phase[2]. When contact is first made with an object ('load' phase), the mean pressure of the frame increases suddenly, resulting in a steep temporal gradient; the red dot shows the detected frame. In brief, we identify the prior local minimum as the frame just before contact (blue dot), which has the maximum hand pose signal (see complete processing details in the Methods). This empty hand pose frame is subtracted from the local maximum frame (green dot; lift and hold phase), which is treated as the frame with the maximal object information. This approach helps in decomposing the tactile map into the hand pose signal and the object-related pressure map (a detailed description of the decomposition is included in the Methods). We analysed the Pearson correlation coefficient between a selected sensor and the remaining sensors in the glove as shown in Fig. 3b. Our correlations are shown in the range from 0 to 1; we did not observe any substantial negative correlations between sensors. We find the largest correlations between the fingertips and the thumb base, where the forces are dominantly applied; this is an expected signature of precision grip in humans. The measured correlations for each sensor can be viewed in our interactive map (Supplementary Data 2). The corresponding correlation between sensors at the fingertips and the full hand, with the decomposed hand pose signal shows little structured correlation, in part demonstrating the effectiveness of our decomposition method (Extended Data Fig. 9). Canonical-correlation analysis on the decomposed object-related tactile map across the different regions of the hand shows the collaborative role between the distal phalanges of the large fingers, which is most often used in generating forces during object grasps (Fig. 3c). In the other phalanges, the distribution is more uniform, corresponding to closed grasps where a large part of the hand surface is in contact with the object at once (see Supplementary Data 2 for an interactive map of the region-level correlations). The correlations between different sensors indicate the collaborations between different hand regions;

the nature of the human grasp is known to be collaborative[27,28]. We have thus empirically and quantitatively observed such collaborations and their spatial extent purely from high-resolution tactile signals. To directly test the proprioceptive content of the tactile signals from the STAG, we articulated specific hand poses in the absence of an object (see G1 to G7 in Extended Data Fig. 10a) based on a standard grasp taxonomy[29]. We observed that the tactile maps related to specific hand poses can be classified with 89.4% accuracy; a visualization of the clustering using *t*-distributed stochastic neighbour embedding (t-SNE) is shown in Extended Data Fig. 10b; confusion matrix from the classification test is shown in Extended Data Fig. 10c; an interactive version of the map is included in Supplementary Data 3. Although hand recognition from visual images has become increasingly robust[30], extracting other meaningful feedback signals (such as establishing contact with an object) remains challenging without a scalable tactile sensing strategy.

Our results demonstrate the broad utility of high-dimensional tactile sensors as well as highlight their enabling potential for future work. The current study focuses mainly on the spatial relationships of tactile signals; the dataset also presents important relationships between sensors that are temporally linked together. These temporal relationships spotlight the dynamics of actions performed by humans. Linking these temporal relationships along with the spatial correspondences between tactile signals would greatly enhance our understanding of the basic principles of dexterous manipulation. Likewise, the dataset presented here also contains synchronized visual information along with the tactile data. In this regard, the STAG is a useful testbed for multimodal learning across visual and tactile domains, which is potentially useful for robotics applications. Finally, the STAG hardware platform itself can be augmented, for example, the STAG could be fitted with diverse sensors that mimic the different sets of mechanoreceptors in the human hand. In addition, transmitting data wirelessly from a wearable module and more compact packaging will extend its utility in manipulation tasks that require considerable mobility.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1234-z.

Received: 5 November 2018; Accepted: 9 April 2019;
Published online 29 May 2019.

1. Bartolozzi, C., Natale, L., Nori, F. & Metta, G. Robots with a sense of touch. *Nat. Mater.* **15**, 921–925 (2016).
2. Johansson, R. & Flanagan, J. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nat. Rev. Neurosci.* **10**, 345–359 (2009).
3. Mahler, J., Matl, M., Satish, V., Danielczuk, M., DeRose, B., McKinley, S. & Goldberg, K. Learning ambidextrous robot grasping policies. *Sci. Robot.* **4**, eaau4984 (2019).
4. Levine, S., Finn, C., Darrell, T. & Abbeel, P. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.* **17**, 1334–1373 (2016).
5. Morrison, D., Corke, P. & Leitner, J. Closing the loop for robotic grasping: a real-time, generative grasp synthesis approach. In *Proc. Robotics: Science and Systems* https://doi.org/10.15607/RSS.2018.XIV.021 (RSS Foundation, 2018).
6. Saal, H., Delhaye, B., Rayhaun, B. & Bensmaia, S. Simulating tactile signals from the whole hand with millisecond precision. *Proc. Natl Acad. Sci. USA* **114**, E5693–E5702 (2017).
7. Osborn, L. et al. Prosthesis with neuromorphic multilayered e-dermis perceives touch and pain. *Sci. Robot.* **3**, eaat3818 (2018).
8. Okamura, A. M., Smaby, N. & Cutkosky, M. R. An overview of dexterous manipulation. In *Proc. IEEE International Conference on Robotics and Automation (ICRA'00)* 255–262 https://doi.org/10.1109/ROBOT.2000.844067 (2000).
9. Cannata, G., Maggiali, M., Metta, G. & Sandini, G. (2008). An embedded artificial skin for humanoid robots. In *Proc. International Conference on Multisensor Fusion and Integration for Intelligent Systems* 434–438 https://doi.org/10.1109/MFI.2008.4648033 (2008).
10. Romano, J., Hsiao, K., Niemeyer, G., Chitta, S. & Kuchenbecker, K. Human-inspired robotic grasp control with tactile sensing. *IEEE Trans. Robot.* **27**, 1067–1079 (2011).
11. Marzke, M. Precision grips, hand morphology, and tools. *Am. J. Phys. Anthropol.* **102**, 91–110 (1997).
12. Niewoehner, W., Bergstrom, A., Eichele, D., Zuroff, M. & Clark, J. Manual dexterity in Neanderthals. *Nature* **422**, 395 (2003).
13. Feix, T., Kivell, T., Pouydebat, E. & Dollar, A. Estimating thumb-index finger precision grip and manipulation potential in extant and fossil primates. *J. R. Soc. Interf.* **12**, https://doi.org/10.1098/rsif.2015.0176 (2015).
14. Chortos, A., Liu, J. & Bao, Z. Pursuing prosthetic electronic skin. *Nat. Mater.* **15**, 937–950 (2016).
15. Li, R. et al. Localization and manipulation of small parts using GelSight tactile sensing. In *Proc. International Conference Intelligent Robots and Systems* 3988–3993 https://doi.org/10.1109/IROS.2014.6943123 (IEEE/RSJ, 2014).
16. Yamaguchi, A. & Atkeson, C. G. Combining finger vision and optical tactile sensing: reducing and handling errors while cutting vegetables. In *Proc. IEEE 16th International Conference on Humanoid Robots (Humanoids)* 1045–1051 https://doi.org/10.1109/HUMANOIDS.2016.7803400 (IEEE-RAS, 2016).
17. Wettels, N. & Loeb, G. E. Haptic feature extraction from a biomimetic tactile sensor: force, contact location and curvature. In *Proc. International Conference on Robotics and Biomimetics* 2471–2478 (IEEE, 2011).
18. Park, J., Kim, M., Lee, Y., Lee, H. & Ko, H. Fingertip skin-inspired microstructured ferroelectric skins discriminate static/dynamic pressure and temperature stimuli. *Sci. Adv.* **1**, e1500661 (2015).
19. Yau, J., Kim, S., Thakur, P. & Bensmaia, S. Feeling form: the neural basis of haptic shape perception. *J. Neurophysiol.* **115**, 631–642 (2016).
20. Bachmann, T. Identification of spatially quantised tachistoscopic images of faces: how many pixels does it take to carry identity? *Eur. J. Cogn. Psychol.* **3**, 87–103 (1991).
21. Torralba, A., Fergus, R. & Freeman, W. 80 million tiny images: a large dataset for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 1958–1970 (2008).
22. D'Alessio, T. Measurement errors in the scanning of piezoresistive sensors arrays. *Sens. Actuators A* **72**, 71–76 (1999).
23. Ko, J., Bhullar, S., Cho, Y., Lee, P. & Byung-Guk Jun, M. Design and fabrication of auxetic stretchable force sensor for hand rehabilitation. *Smart Mater. Struct.* **24**, 075027 (2015).
24. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 https://doi.org/10.1109/CVPR.2016.90 (IEEE, 2016).
25. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
26. Brodie, E. & Ross, H. Sensorimotor mechanisms in weight discrimination. *Percept. Psychophys.* **36**, 477–481 (1984).
27. Napier, J. The prehensile movements of the human hand. *J. Bone Joint Surg. Br.* **38-B**, 902–913 (1956).
28. Lederman, S. & Klatzky, R. Hand movements: a window into haptic object recognition. *Cognit. Psychol.* **19**, 342–368 (1987).
29. Feix, T., Romero, J., Schmiedmayer, H., Dollar, A. & Kragic, D. The GRASP taxonomy of human grasp types. *IEEE Trans. Hum. Mach. Syst.* **46**, 66–77 (2016).
30. Simon, T., Joo, H., Matthews, I. & Sheikh, Y. Hand keypoint detection in single images using multiview bootstrapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 4645–4653 https://doi.org/10.1109/CVPR.2017.494 (IEEE, 2017).

Acknowledgements S.S. thanks M. Baldo, V. Bulovic and J. Lang for their comments and discussions. P.K. and S.S. thank K. Myszkowski for discussions. We gratefully acknowledge support from the Toyota Research Institute.

Reviewer information *Nature* thanks Giulia Pasquale and Alexander Schmitz for their contribution to the peer review of this work.

Author contributions S.S. conceived the sensor and hardware designs, performed experiments, was involved in all aspects of the work and led the project. P.K. performed all data analysis with input from all authors. Y.L. performed network dissection. S.S. and P.K. generated the results. A.T. and W.M. supervised the work. All authors discussed ideas and results and contributed to the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1234-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1234-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

STAG sensor fabrication. The STAG consists of a sensing laminate attached to a light, custom-knit glove designed not to interfere with hand movements (loosely knitted at half-gauge on a Shima Seiki SWG091N2 15-gauge v-bed knitting machine). The sensing laminate is made by first laser-cutting (Universal PLS 6.150D CO_2 laser cutter; Universal Laser Systems) the force-sensitive film (FSF) (3M Velostat electrically conductive copolymer 0.1 mm thick; Adafruit Industries) to fit a hand. To prevent movement of the FSF during the laser cutting process, we attach the film to an acrylic board with a thin layer of water. The laser-cut film is washed to remove debris from the surface. The laser-cut pattern includes holes to route conductive thread electrodes (3-ply stainless steel conductive threads of 12 μm fibre diameter and about 0.34 mm overall diameter; SparkFun Electronics) and slots at the finger joints to allow unrestricted movement (pattern in Fig. 1b; original design is available from the corresponding author). The conductive threads are sewn with a needle on either side of the FSF, ensuring that there are no points of direct contact between electrodes. Regions of overlap between the 64 row and column electrodes respond to forces through a change in the through-film resistance. The electrodes are first held taut and then subsequently held in place by attaching a thin, stretchable, two-sided acrylic adhesive tape (3M 468MP 200MP adhesive 0.13 mm thick) on both sides. The two exposed sides of the adhesive are insulated with a thin, stretchable polybutylene-coated LDPE film (about 13 μm; Saran; S.C. Johnson & Son). The laminate architecture is shown in Fig. 1b and Extended Data Fig. 3b. The exposed conductive threads are coated with a polydimethylsiloxane mixture (PDMS, 1:10 ratio of crosslinker to pre-polymer; SYLGARD 184, DOW Corning). The PDMS-coated conductive thread electrodes are separately cured on a hot plate at 60 °C for 2 h and left overnight at room temperature. The PDMS-coated conductive thread electrodes are attached to insulation-displacement connectors that can be connected to the readout circuit. Forming robust electrical connections with the conductive thread electrodes is typically challenging. Using PDMS to insulate the conductive threads and subsequently using insulation-displacement connectors resulted in robust electrical contacts, which was critical for the long-term use of the STAG. Likewise, the robustness of connections between fingers is critical to the long-term stability of the STAG. Furthermore, they are required to allow free movement of individual fingers without hindering motion and object interactions. We used a single conductive thread reaching through all the fingers for each row and insulated the region between fingers using the acrylic adhesive tape and LDPE film to form insulated ribbons routed along the sides of fingers (seen between fingers in Extended Data Fig. 1a, b).

The regular 32 × 32 array version of the sensor laminate was fabricated using a process similar to that described above. In this case, a square of FSF was cut and assembled into a laminate in two separate versions. Extended Data Figure 3b shows a design that is identical to the version used in the STAG. Extended Data Figure 3a shows a simplified version of the above design by replacing the adhesive film and LDPE film with a 25.4-μm-thick polyimide film with adhesive on one side. Although it is simpler in construction, since the polyimide film is inextensible, we observed that this design (Extended Data Fig. 3a) is less flexible than the STAG laminate architecture, and is therefore better suited for use in fixed conditions. The spacing between the electrodes was set to 2.5 mm using a specially designed thread layout tool (Extended Data Fig. 3c). The layout tool consists of two pieces of acrylic, one of which has laser-etched grooves to hold the conductive threads in place with the correct spacing. The two designs in Extended Data Fig. 3a, b were used to fabricate the square 1,024-element sensor array shown in Extended Data Fig. 3d, e respectively.

Auxetic designs of the sensor (10 × 10 elements) were designed by first patterning the FSF with the design shown in Extended Data Fig. 5b. The holes allow the conducting threads to be routed (red and blue traces) to enable stretching. The cuts in the FSF allow individual sensor squares to rotate and enable extensibility in all directions. A close-up of the FSF after routeing the conductive threads is shown in Extended Data Fig. 5c. The double-sided adhesive and LDPE film (as in Extended Data Fig. 3b) are then added to both sides and the slots identical to those on the FSF are cut with a scalpel. The conductive thread electrodes are subsequently insulated with PDMS, and then connected to the insulation-displacement connector (finished design in Extended Data Fig. 5d). The auxetic sensor array can be folded, crushed and stretched in different directions as shown in Extended Data Fig. 5e, f.

STAG sensor characterization. To characterize individual sensor responses, we cut approximately one-inch squares of the FSF and attached orthogonal electrodes on either side as in the STAG architecture (inset of Extended Data Fig. 2f). The typical sensor force response was measured by applying controlled normal forces using a table-top mechanical tester (Instron 5944; Instron) and simultaneously recording the electrical resistance using a Sourcemeter (2611B Keithley Instruments). The loading rate was controlled at a specific strain rate in all tests (0.05–0.1 mm min^{-1}) until a maximum load of 0.5 N or 5 N was reached for the results shown in Fig. 1c and Extended Data Fig. 2a, b. For the long-term tests in Extended Data Fig. 2c, d, we cycled the force between 20 mN and 0.5 N for 1,000 cycles at a controlled strain rate of 2.5 mm min^{-1}. The differential scanning calorimetry (DSC) measurements were performed using a TA Instruments Q100 DSC (TA Instruments) at a temperature ramp rate of 10 °C min^{-1}. The differential scanning calorimetry measurements show that the material is probably a two-polymer blend that softens at temperatures around 100 °C. FSF is considered to be thermoformable, but to directly check the feasibility of thermoforming our sensor and studying the temperature stability, we placed fabricated single-element prototypes in a convection oven at controlled temperatures for 10 min and measured the resistance through the film after removal from the oven (Extended Data Fig. 2f). The sensors show large changes in the resistance beyond 60 °C.

Circuit architecture and design. The passive matrix design of the STAG makes the fabrication simple and the design easily extensible to multiple platforms. However, immediately after fabrication, any two electrodes will appear to be electrically shorted together owing to a large number of spurious current paths; this is well known in passive arrays. A signal isolation circuit is required to overcome the extensive crosstalk between sensors[31]. It is possible to eliminate a majority of crosstalk and parasitic effects using a balanced readout scheme. Here we used an improved version of an electrical-grounding-based readout architecture[22] where one row of the sensor currently being read is grounded while all other rows are maintained at the reference voltage V_{ref} (2.5 V in our design; see Extended Data Fig. 1c, d; active row is indicated by a grey arrow). The current between all resistors in other rows is ideally 0 A since the voltage difference across all the resistors is 0 V. During this state, a 32:1 analog switch is used as an analog demultiplexer to raster through the row and read individual resistances one by one. After completing measurements at the active row, the 32 single-pole double throw (SPDT) switches are used to ground the next row while returning the currently active row to V_{ref}. We added charging resistors, R_c, to each column to charge the inverting node of each amplifier back to V_{ref} quickly and to reduce the input noise. We observed that this led to a more stable readout of the resistances without affecting the actual potential at the output of the amplifier. Note that the R_c arm of the circuit in each column is analogous to an adder circuit where one input is always 0, that is, the input to the R_c arm is V_{ref} which is the same as the voltage at the non-inverting terminal of the amplifiers. Rastering through the readout resistors in the matrix is controlled by an Arduino Nano by switching the 32 SPDT switches and the 32:1 analog switch. The single sensor measurement at the output of the analog switch is converted to a digital signal (10-bit resolution; 0–1,023 corresponding to 0–5 V) and transmitted serially to a computer. An image of the fabricated printed circuit board is shown in Extended Data Fig. 1d, with the insulation-displacement connector cables inserted to connect the sensor array (seen at the top-right and bottom).

Dataset acquisition methods. Previous psychophysics studies on human tactile performance have used carefully designed experimental conditions (objects, awareness and tasks) that are each motivated by the purpose of the study. For instance, to demonstrate the orientation dependence of human tactile performance, previous studies have used protocols with blindfolded humans interacting with artificial objects of similar properties that are fixed in space[32]. Likewise, careful experimental designs have also been implemented in tactile interaction studies in robotics[33]. Our general objective here is to learn from successful human interactions with objects, which are typical of daily interactions. Our data acquisition methods were designed with this in mind, and motivated by a few cues from seminal human tactile perception studies of the past three decades[28,34,35]. In particular, the use of everyday objects is better as opposed to artificial objects (unless critically required for the task[32]) since human performance in object interaction and identification is underestimated when unfamiliar objects are used[34]. Therefore, the STAG prototype was used to record single hand manipulation of 26 different common objects with a few different sizes, weights and materials (Extended Data Fig. 6).

Visually aware and blindfolded conditions. The task setup influences the general hand movements and interactions in haptics and tactile recognition[28,35]. Therefore, the level of awareness of objects during interactions is a critical part of experimental design. In this regard, human perception studies have shown that blindfolded subjects can identify common objects within 2–3 s almost perfectly; this interval increases to about 16 s when wearing a glove but with no loss in accuracy (summarized in a review[35]). In blindfolded interactions, the level of awareness increases during these intervals, and the interactions become fully aware once the object is identified. Our large-scale dataset was captured with complete visual access to the object (visually aware); it allows us to record human interactions with a constant level of awareness of the object at each frame. Each object was manipulated for 3–5 min at a time and included several different grasps and touch sequences. However, given that the objects are in sight, some of these grasps are more discriminative (that is, object-dependent). To test the generality of this visually aware dataset, we also performed a blindfolded study where each interaction was terminated as soon as the object was identified (at the moment the subject is aware of the object). Our original CNNs that were trained purely on the visually aware dataset were used in evaluating the blindfolded test set.

The blindfolded tests were performed by the subject (S.S.), where the task was to identify the object. In all these tests, the objects were placed on a soft foam surface to reduce sounds and the subject was made to listen to white noise through headphones. The tactile recording was started and the shoulder of the subject was tapped. The subject was asked to interact with the object, identify it and hold the object still. The tests were stopped as soon as the object was identified. Typically, the objects were identified by the subject in 6.16 ± 2.65 s (Supplementary Fig. 6a). The tasks were performed continuously 104 times (4 times for each of the 26 objects) in a randomized order. One additional task was performed without any object as an empty hand control. The objects were identified correctly by the subject in 103 out of the 104 tests. The last 5 frames from each identification attempt (total 20 frames per object from 4 identification attempts) were used for a classification test using a CNN trained only on the visually aware data. The confusion matrix for this test is shown in Supplementary Fig. 6b. We observe that the lighter objects are harder to identify more accurately using a single input frame, and the single-input ($N = 1$) top-1 classification accuracy is 28.19% (top-3 classification accuracy of 49.91%). Overall, the performance is slightly worse than the single-input classification tests of the frames from the visually aware dataset ($N = 1$, top-1 classification accuracy about 37.97% and top-3 classification accuracy about 60.43% in Fig. 2b). The ability to identify objects using tactile frames from blindfolded tests using the original CNN trained only on the visually aware dataset demonstrates that a general set of object interactions is probably captured in the visually aware dataset.

Dataset acquisition metrics. In addition to recording interactions with 26 objects, we recorded the empty hand data while articulating the hand without interacting with any object. In all cases, we also recorded corresponding visual images from the experiment using a FLIR GS2-GE-20S4C-C camera for illustration. Overall we recorded 135,187 frames for the visually aware dataset used for the object identification task. All experiments (visually aware and blindfolded tests) were performed by S.S. using a STAG worn on the right hand. We recorded the tactile maps from the glove along with timestamps (and corresponding visual frames) at an average frame rate of 7.3 frames per second. Each object sequence was recorded three times over different days and in a randomized order for the visually aware dataset.

A similar procedure was followed with the same set of objects in recording the dataset for the weight estimation task. To ensure that the weight of an object is not trivially associated with a particular grasp type, we standardized the grasp used in the recording to be identical for all objects. Each object was picked up from above using a multi-finger grasp (see Extended Data Fig. 8a for example images) where the weight of the object was supported by the fingers and the thumb. During each recording, the selected object was grasped, lifted, held and dropped to a flat table multiple times. The procedure was repeated in 10-s intervals for a total duration of 1 min. Each object was recorded in multiple recording sessions. In total, we recorded 11,682 frames for the weight estimation dataset.

Finally, we recorded a dataset of different articulated hand poses (empty hand and G1 to G7; Extended Data Fig. 10) based on a standard grasp taxonomy[29] to analyse the proprioceptive content of the tactile information recorded with the STAG. We recorded each articulated hand pose in random order over 7 different recording sessions and collected 24,037 frames in total. The processing details are described in the 'Hand pose' section.

Object identification. *Processing and network design.* The overall object identification scheme relies on the use of CNNs to extract meaningful information from tactile signals and classify objects. There are many examples of the use of CNNs with tactile sensors, especially in the context of robotics[36–39]. This section describes the full details of our tactile data processing and network architecture.

The recorded tactile dataset (135,187 frames) contained useful signals in the range 500–650 (0–1,023 corresponds to 0–5 V at the amplifier output using a 10-bit ADC). The tactile map, transmitted as a 32×32 map from the readout circuit, is first normalized from 500–650 to 0–1. We discard all frames with any sensor reading over 950, which results from shorted electrodes; this happens when the STAG is punctured by a sharp metal object, and is therefore rare and is easy to detect. We next remove frames without any useful signal when the hand is not in contact with the object. We use the empty hand recording as a reference and detect the maximum response for each sensor over time. We then consider a frame with an object to be valid if at least one sensor response is above the maximum response found in the empty hand recording. After filtering we obtain 88,269 valid frames.

To estimate the accuracy of the above frame-classification method, we manually inspected a random sequence of 150 frames from five different recordings (allen key set, multimeter, tape, mug and the foam model of a brain). In our inspection, we used the temporal sequence of both the tactile signals and the synchronized visual frames to determine the state of the grasp and to validate the thresholding-based algorithm. This allows us to notice changes in the pressure, object position and hand position, consequently making it easier to check for contact, and partially alleviates difficulties with occlusion. We observe that out of these 750 frames, there were only 57 false negatives (7.6%) where an object contact was omitted and only 3 false positives (0.4%) where a fake contact was detected. The main sources of false negatives are the weak contacts at the onset of the grasp where the pressure signal is weak; soft objects are more prone to this issue (24 for the foam model of a brain versus 1 for the mug). The main sources of false positives are the occasionally pronounced hand poses. This issue is rare because we threshold our data based on the empty-hand dataset, which covers a large range of possible hand movements. Overall, our automatic frame classification method is conservative and presents clean data.

We then split the dataset into training and test subsets; out of the three sets of trials in the visually aware dataset, we used two sets of trials for training, and one set of trials for testing. Each subset was randomly subsampled to contain a balanced number of valid frames for each of the object classes (26 objects and empty hand recording). The training set has a total of 36,531 frames (1,353 per class), and the test set has 16,119 frames (597 per class).

With the aim of predicting an object's identity based on its tactile signature, we treat each recording trial of a given object as a single instance of the problem, which simulates an agent exploring an object using multiple grasps. We feed $N = 1...8$ frames from the recording to the deep neural network to accommodate information from different grasp configurations and provide more varied sensory data. When we use $N > 1$ input frames to our network during evaluation we consider two different strategies for their selection. The first one is a simple random choice of N frames from the recording. The second is aimed at minimizing the redundancy of data between the N frames by maximizing their variance. For this, we use principal component analysis to reduce the dimensionality of the tactile signal to 8. We then find N clusters via k-means clustering (Fig. 2e); that is, we complement every randomly selected input frame with $N - 1$ other frames, each of which belongs to a different cluster.

We use a modified version of the ResNet-18 architecture[24] as the base of our network (Fig. 2a). We reduce the filter size of the initial convolution layer from 7 to 3 and the stride from 2 to 1. This allows the entire filter to fit within the smallest features in our sensor data (finger width is 3 pixels). Since our inputs are 32×32 pixels, we remove the upper two of the four ResNet layer groups leaving the final feature vector size to be 128 values. To reduce overfitting to our training, set we introduce a spatial dropout layer[40] with 20% drop probability between the two remaining block layers. Additionally, we also augment our training data by additive Gaussian noise with zero mean and a standard deviation of 0.015 during training. To incorporate multiple input frames, we apply the same network with shared weights to each input. The outputs of all network branches are then concatenated and reduced to 128 dimensions by a per-pixel convolution. After spatial averaging, a final fully connected layer computes the classification vector. We implemented this network in the PyTorch (https://pytorch.org/) deep learning framework[41]. We use Adam solver implemented in PyTorch to train our model and minimize the cross-entropy loss. We apply an initial learning rate of 10^{-3}, which we decrease by a factor of 10 every 100 epochs. We train the network using our training dataset for 200 epochs with batches of 32 samples. We report the average results over 10 training runs. Similar methods were used for the training and classification tests used to evaluate different sensor resolutions (results in Supplementary Fig. 5). In all cases, the tactile data was downsampled (by averaging) and resized to an input size of 32×32 in order to use the same network. These inputs were used for training and subsequent classification tests using methods identical to those described above.

The 3×3 filter of the first convolution layer is not easy to interpret visually owing to the low resolution. Therefore, we trained a scaled version of the model with the first convolution increased to 9×9 and the stride to 3, for visualization. Max pooling layers reduce the dimensionality of the model by spatially selecting the locally highest activation values and are used in both versions of the model. We adapted the filter size of the first max pooling layer to 7 and its stride to 4 to make the resulting features similar to the original model. We also scaled the resolution of the inputs identically (by 3) using bilinear upsampling. The remainder of the model and the training procedure stayed the same. The learned convolution filters are shown in Fig. 2d; the resulting features and performance are similar to the original model. See Extended Data Fig. 7i for the corresponding 3×3 convolution filters learned by the original network. The first convolution filters of the standard ResNet-18 architecture pre-trained on ImageNet[25] are shown in Extended Data Fig. 7j for comparison.

To further understand the internal representation of the CNN, we perform Network Dissection[42], a widely used technique for analysing the deeper layers of a network. Specifically, for any convolution filter, the method first ranks the tactile maps according to the highest activation value. We subsequently select the tactile frames that rank first with the highest activations from each object category. This, in essence, shows the top candidates that different convolution filters have learned to look out for. Supplementary Figure 4 highlights the activations of a few representative convolution filters in both the first and the second ResNet blocks. In each pressure map, the yellow contours highlight the regions where these convolution filters are most activated. Typically, a threshold of 0.2–0.5 times the peak activation is used to generate these contours; here we use 0.3 times the peak activation for

these visualizations. We observe that the convolution filters in the first of the two ResNet blocks are most activated in smaller, spatially confined regions around the tactile signal peaks. However, interesting grasp-related concepts emerge in the second ResNet block. For instance, these filters are activated by different object patterns or by specific regions of the hand such as the palm or the thumb.

Weight prediction. *Dataset.* We performed identical pre-processing of frames as for the classification task to obtain normalized 32×32 tactile maps. The collection sequence of the grasping procedure for weight estimation was predetermined; we used the frames between 4 s and 6 s of each sequence when the object was held (2,301 frames). To prevent the problem from being reduced to a simple classification task where each weight is associated with an object, we used a 'leave-one-out' approach where the object whose weight is predicted is not part of the training data.

Regression. Our goal is to predict weights (in grams) for a previously unseen object based on a single tactile frame; unlike object recognition, weight estimation is possible as soon as an object is held. Here we use a similar ResNet-based network architecture. Because this task is based on a single frame input, we remove the input branch concatenation (retaining a single branch of the CNN alone) and directly apply a fully connected layer to regress the weight. Owing to the wide range of object weights in our dataset, we perform the prediction in the logarithmic space. We optimize the parameters of our model by minimizing the mean squared error between the predicted and measured ground truth weights (in logarithmic space). We used the same optimizer and learning parameters as before and train the network for ten epochs.

We construct a linear baseline for the weight estimation. This is motivated by the fact that the weight of an object is directly linked to the sum of all forces it applies when the object rests on a horizontal surface (that is, the naive estimate of weight is $aX + b$, where X is the sum of the tactile map). This baseline is not valid in practice, however, because the tactile response is affected by the articulation of the hand (grasp), as well as by the surface friction and the additional force used to avoid slipping. We used the same 'leave-one-out' training procedure to find the best choice of a and b that minimizes the mean squared error between the prediction and ground-truth weight in logarithmic space.

Figure 2f compares the results of our network with the linear baseline model. The error for each object corresponds to a different instance of our model trained with that particular object left out from the training dataset. We also computed a mean predicted error across all objects expressed in grams in linear space. We found that the average prediction error of our model is 56.88 g, which is less than the 89.68 g of the naive linear model. This result shows that nonlinear behaviours connected with grasps and the physical properties of the objects cannot be omitted and that the neural network can compensate these to some extent. The weight estimation errors in different weight ranges are listed in the table in Extended Data Fig. 8b. We observe that the CNN outperforms the linear baseline in all cases. Furthermore, the estimation errors are also listed as relative errors (normalized by object weights); this is analogous to the Weber fraction used in the tactile weight perception literature[26]. We also tested an additional modification to the linear baseline by removing the hand pose component from the tactile signal using the methods outlined in the section 'Decomposing signal and sensor correlations' and Fig. 3a. We observed that the modified linear baseline did not present any noticeable improvement over the naive baseline approach, and the performance of the CNN was better than both linear methods. We believe that this is due to the complex relationship between the tactile signals and the weight estimates and this is borne out by previous weight perception studies[43].

It is noteworthy that humans rely both on cutaneous and kinaesthetic senses (and their inertial responses during object interactions) to gauge weight effectively. The Weber fraction is known to be about 1/3 when using the cutaneous sensors[26]. Our estimated metric analogous to the Weber fraction is similar for moderately heavy to heavy objects (Extended Data Fig. 8b); this performance is comparably good considering that the human hands possess additional types of mechanoreceptors that are not used here.

Hand pose. *Dataset.* To evaluate whether hand pose (proprioceptive) information can be retrieved from the glove even when no objects are being manipulated, we picked seven distinct grasps from the grasp taxonomy[29] (along with a neutral hand pose for reference). We chose these specific grasps because they are often used and involved object interactions at the palmar side of the hand, which is covered with sensors in the STAG. Furthermore, we articulate each hand pose back and forth from the neutral hand pose; this covers a larger set of grasps in the taxonomy and increases the intra-class variance. We removed ambiguous articulations of the hand (close to the neutral pose) by thresholding. Frames with a mean pressure signal higher than 75% of a local dynamic range in a symmetric window with a 6-s temporal radius were considered as reliably valid hand poses. This filtering step reduced the number of frames to 7,697.

t-SNE embedding. We applied t-SNE[44] to the tactile frames to discover structure in the signal and to evaluate its dependency on the performed hand pose. We first reduce the dimensionality of our pressure readings from the original 548 active sensors to 50 using principal component analysis and then applied t-SNE to obtain the final two-dimensional projection presented in Extended Data Fig. 10. The network described in the 'Object identification' section was also used to train and classify the hand poses from single tactile maps with 89.4% accuracy (average of 10 runs; 3,080 training frames and 1,256 distinct test frames). The confusion matrix corresponding to these tests is shown in Extended Data Fig. 10c. Most grasps can be identified correctly except for G1 and G6, which are occasionally confused.

Decomposing signal and sensor correlations. We computed the mutual correlation of the sensors over our entire object identification and classification dataset. Each recorded frame contains two different signals: the hand pose signal and the object-related pressure due to forces between an object and the hand. The hand pose signal denotes the articulation of the hand and would be present even in the absence of an object. We assume the hand pose signal to be saturated once the hand reaches full articulation just before contact (reach phase in Fig. 3a). This simplification holds in a majority of cases where the hand articulation does not change much after initial contact. In this condition, the additional force due to an object contact can be superimposed on the articulated hand. By design, the sensors respond to normal compressive force components between the upper and the lower electrodes. Therefore, the force components picked up by the sensors can be treated as additive. Furthermore, the output of the sensors as seen at the output of the amplifier circuit or the ADC is linear with respect to the applied force in the working range (see Supplementary Fig. 2).

To extract the hand pose signal and the object pressure signal from the tactile frame, we implement the following decomposition procedure. We first compute the mean pressure response for every frame of each recording and filter it over time with a symmetric Gaussian kernel with the three-frame standard deviation (around 400 ms) to remove noise. The gradient of this signal (as a forward difference) can show possible object contact points as local maxima of the gradient (symmetrical window of five frames; about 700 ms). Considering that the presence of two distinct contacts in this window is unlikely, we first locate the nearest preceding minima (up to 20 frames away) of the filtered mean pressure signal. To ensure that this minimum is reliable and to avoid detecting random pressure fluctuations, we require the mean pressure value to be below 20% of the recording dynamic range. Similarly, we find the next local maxima and ensure it is larger than 20% of the dynamic range of the recording. The minimum and maximum are marked as the empty hand pose frame (blue dot in the reach phase; Fig. 3a) and as the object manipulation frame (green dot in the lift and hold phase). To recover a larger amount of useful data, we further explore the frames surrounding the local maximum frame and include every frame up to a symmetrical 80-frame radius to the object pressure set until a frame that does not appear to be valid; here, valid frames are those that lie above the detection threshold (0.5) in this normalized scale. The empty hand pose was subtracted from each of the object manipulation frames to recover the pure object-related tactile data.

To evaluate the effectiveness of this decomposition process, we manually analysed the first 60 s of recordings with five random objects of different sizes and weights (full cola can, mug, multimeter, pen, and tape). We observe that 64 grasps were detected out of the 93 while only producing 4 false positive detections. Typically, more detections were missed for light objects compared to heavier objects (for example, 12 for the tape versus 2 for the full cola can). Furthermore, the location of the before-contact point (blue) was on average 1.7 frames before the true location. The first detected object-contact point (first of the green frames) was on average 3.5 frames after the true location. Our algorithm is therefore conservative as it does not produce data from false locations and places the empty hand sample safely in the pre-contact space, as well as uses reliable post-contact frames as a source for extracting the object signal.

Pearson correlation coefficient and canonical-correlation analysis were used to analyse the correlations between sensors and sensor groups separately on the decomposed hand pose data and the decomposed object-related tactile data. Pearson correlations coefficients based on single sensor correlations uncover local relations between neighbouring sensors but are sensitive to the selection of specific sensors and do not present overall trends. To extract phalange level correlations, we used canonical-correlation analysis; results are shown in a circular plot[45] in Fig. 3c. The three finger phalanges, starting at the palm and going outward, are proximal, intermediate and distal. Note that the thumb does not have an intermediate phalanx by standard convention. Since canonical-correlation analysis always results in larger correlations than those achieved at the individual sensor level, we show the differences in the relations by subtracting the minimum correlation; the intermediate phalanx of the little finger and the proximal phalanx of the index finger show minimum correlation.

Hardware designs and firmware availability. Printed circuit board designs and firmware files, and STAG design files are available from the corresponding author.

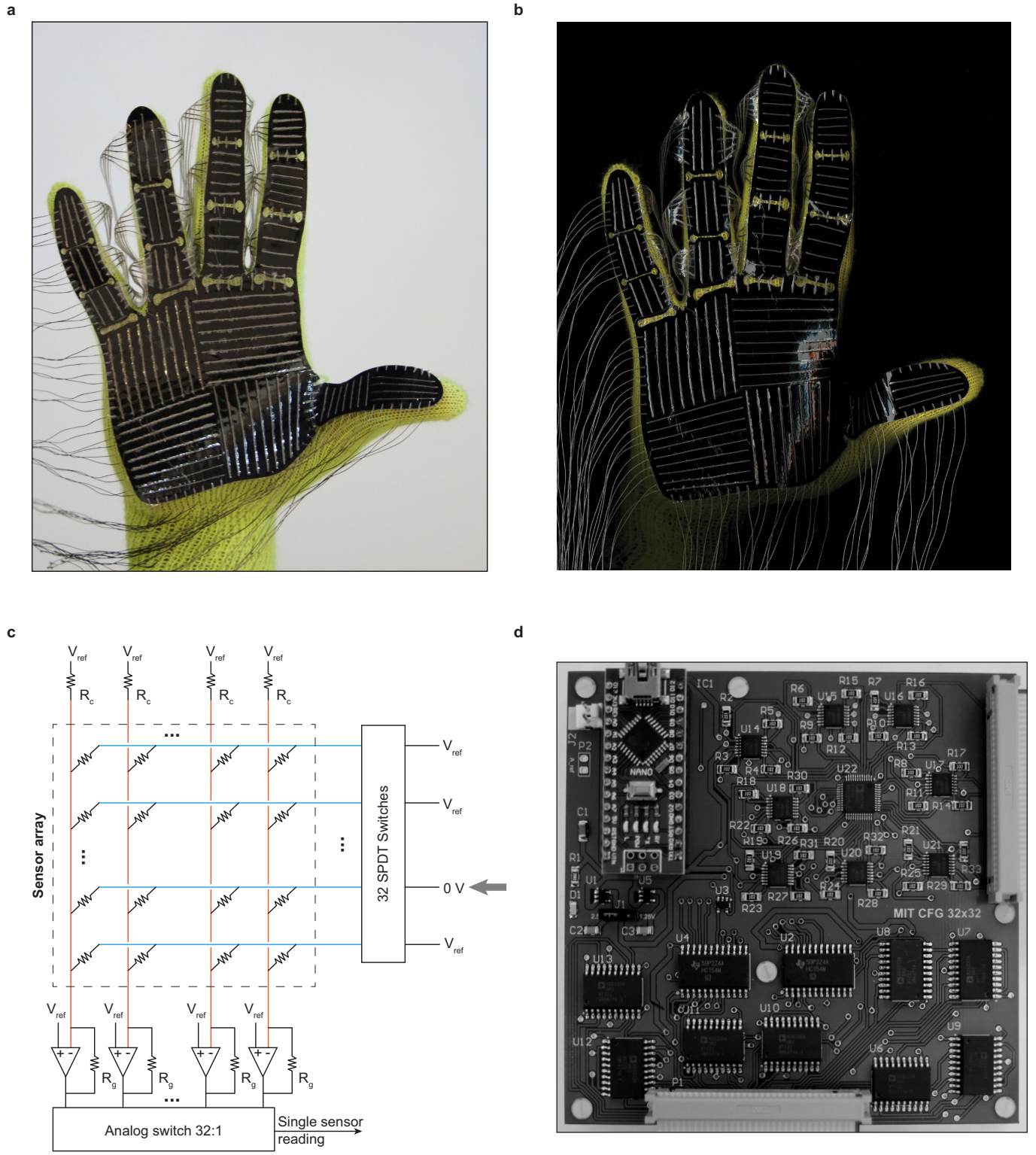
Code availability

Custom code used in the current study is available from the corresponding author on request.

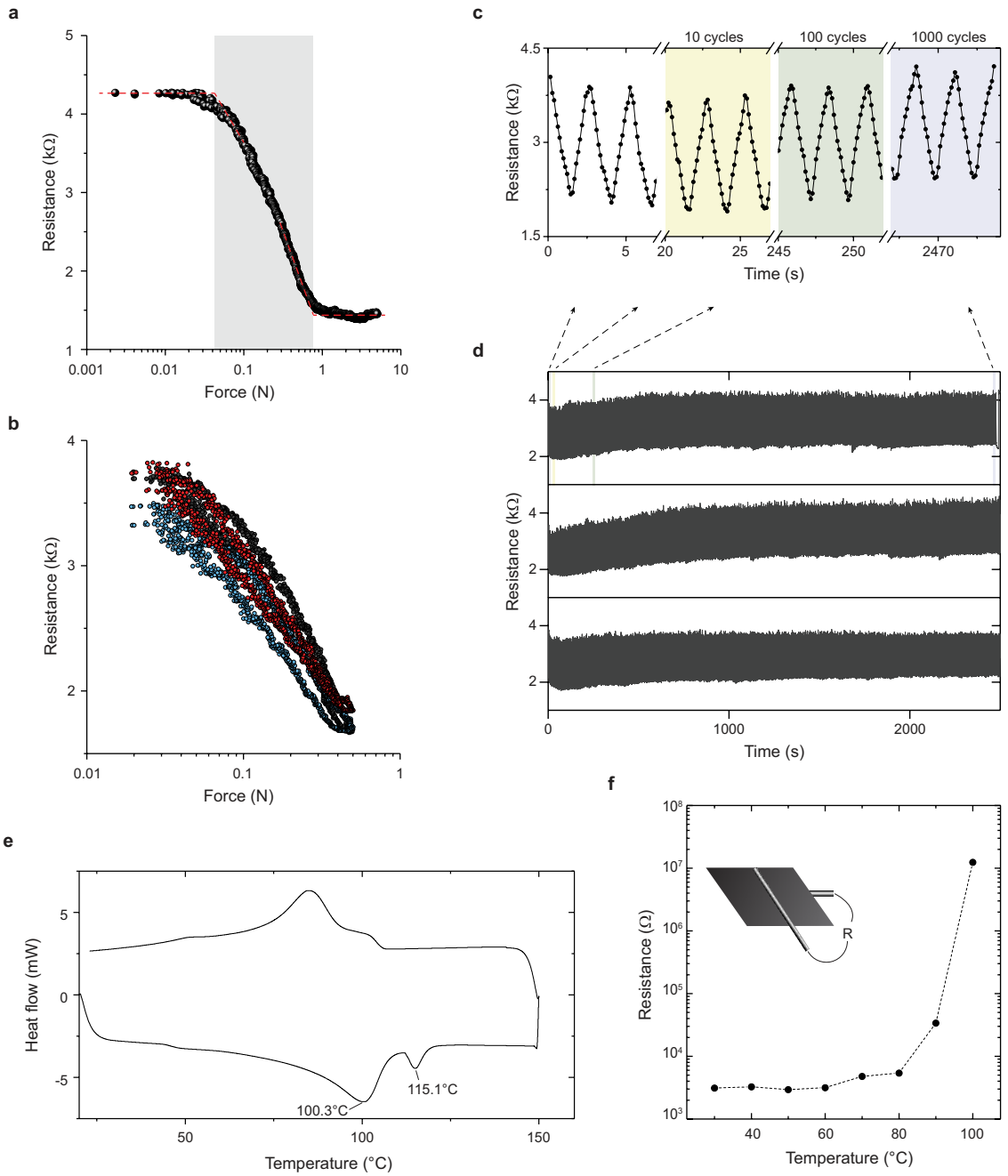
Data availability

Source data for key figures in the manuscript are included as interactive maps in Supplementary Data 1–3. Please load (and refresh) all '∗.html' pages in Firefox or Chrome. The tactile datasets generated and analysed during this study are available from the corresponding author on request.

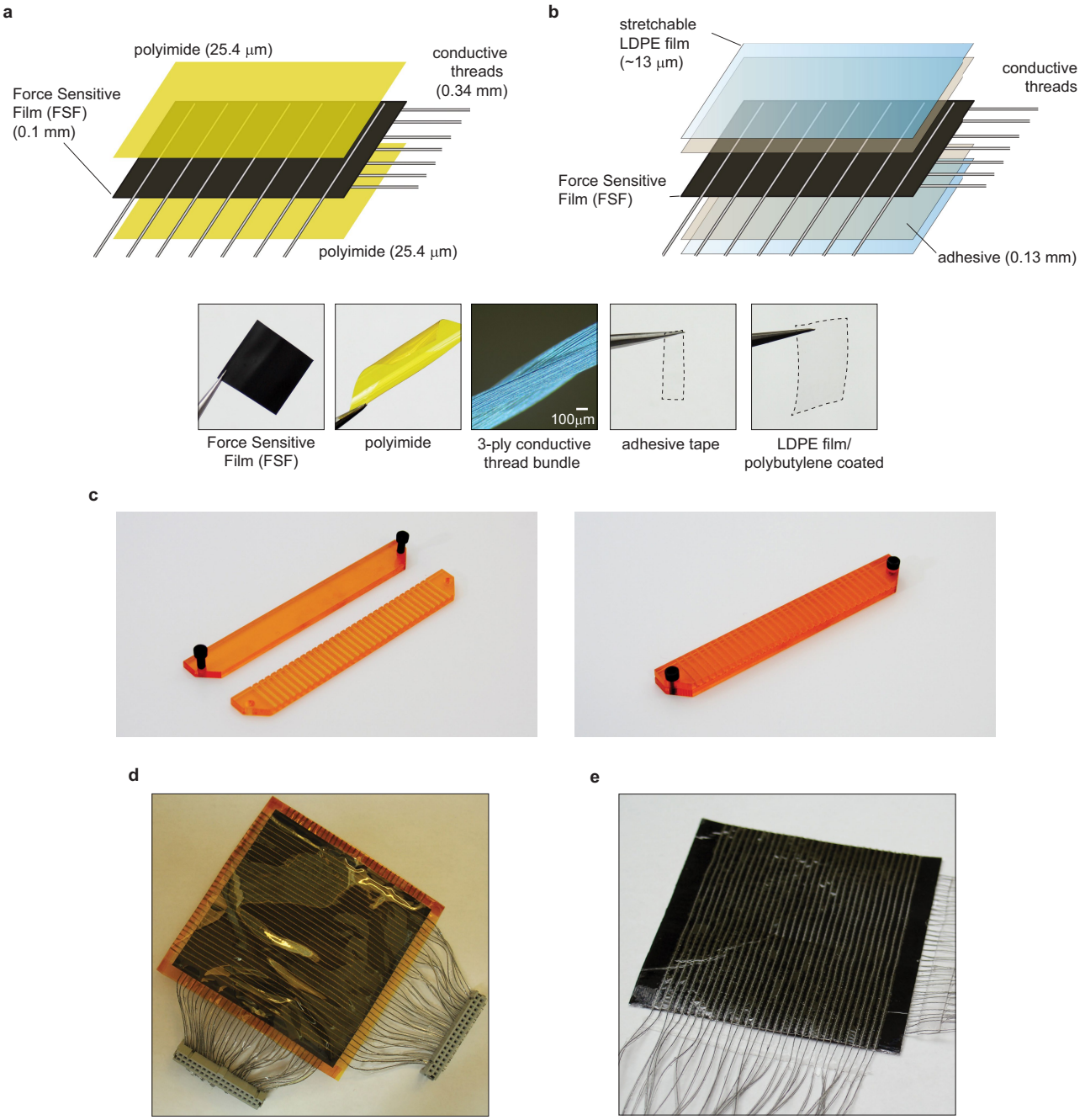
31. Lazzarini, R., Magni, R. & Dario, P. A tactile array sensor layered in an artificial skin. In *Proc. IEEE International Conference on Intelligent Robots and Systems (Human Robot Interaction and Cooperative Robots)* 114–119 https://doi.org/10.1109/IROS.1995.525871 (IEEE/RSJ, 1995).
32. Newell, F., Ernst, M., Tjan, B. & Bülthoff, H. Viewpoint dependence in visual and haptic object recognition. *Psychol. Sci.* **12**, 37–42 (2001).
33. Higy, B., Ciliberto, C., Rosasco, L. & Natale, L. Combining sensory modalities and exploratory procedures to improve haptic object recognition in robotics. In *Proc. 16th International Conference on Humanoid Robots (Humanoids)* 117–124 https://doi.org/10.1109/HUMANOIDS.2016.7803263 (IEEE-RAS, 2016).
34. Klatzky, R., Lederman, S. & Metzger, V. Identifying objects by touch: an "expert system". *Percept. Psychophys.* **37**, 299–302 (1985).
35. Lederman, S. & Klatzky, R. Haptic perception: a tutorial. *Atten. Percept. Psychophys.* **71**, 1439–1459 (2009).
36. Kappassov, Z., Corrales, J. & Perdereau, V. Tactile sensing in dexterous robot hands. *Robot. Auton. Syst.* **74**, 195–220 (2015).
37. Gao, Y., Hendricks, L. A., Kuchenbecker, K. J. & Darrell, T. Deep learning for tactile understanding from visual and haptic data. In *Proc. International Conference on Robotics and Automation (ICRA)* 536–543 https://doi.org/10.1109/ICRA.2016.7487176 (IEEE, 2016).
38. Meier, M., Walck, G., Haschke, R. & Ritter, H. J. Distinguishing sliding from slipping during object pushing. In *Proc. IEEE Intelligent Robots and Systems (IROS)* 5579–5584 https://doi.org/10.1109/IROS.2016.7759820 (2016).
39. Baishya, S. S. & Bäuml, B. Robust material classification with a tactile skin using deep learning. In *Proc. IEEE Intelligent Robots and Systems (IROS)* 8–15 https://doi.org/10.1109/IROS.2016.7758088 (2016).
40. Tompson, J., Goroshin, R., Jain, A., LeCun, Y. & Bregler, C. Efficient object localization using convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 648–656 https://doi.org/10.1109/CVPR.2015.7298664 (IEEE, 2015).
41. Paszke, A. et al. Automatic differentiation in PyTorch. In *Proc. 31st Conference on Neural Information Processing Systems (NIPS)* 1–4 (2017).
42. Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. Network dissection: quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3319–3327 https://doi.org/10.1109/CVPR.2017.354 (IEEE, 2017).
43. Flanagan, J. & Bandomir, C. Coming to grips with weight perception: effects of grasp configuration on perceived heaviness. *Percept. Psychophys.* **62**, 1204–1219 (2000).
44. Maaten, L. V. D. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
45. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).



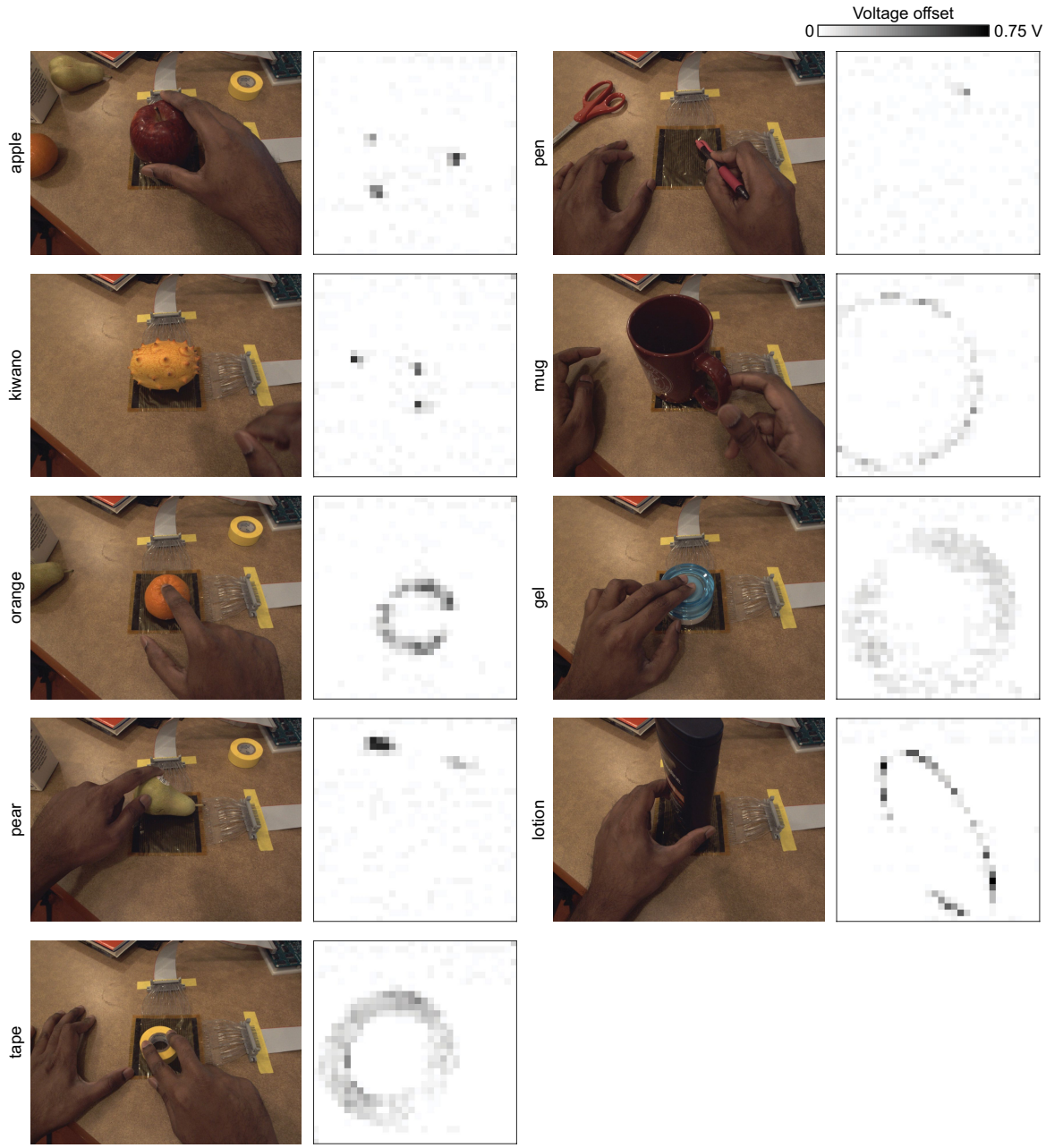
Extended Data Figure 1 | STAG images and readout circuit architecture. **a**, Image of the finished STAG just before the electrodes are insulated. **b**, Scan of the STAG. **c**, Electrical-grounding-based signal isolation circuit (based on ref. [22]). The active row during readout is selected by grounding one of the 32 single-pole double throw (SPDT) switches. A 32:1 analog switch is used to select one of the 32 columns at a time. Here R_c is the charging resistor, V_{ref} is the reference voltage, and R_g sets the amplifier gain. **d**, Fabricated printed circuit board that interfaces with the STAG. The two connectors shown on the top right and bottom are connected to the column and row electrodes of the sensor matrix. The charging resistors (R_c) are on the back of the printed circuit board.



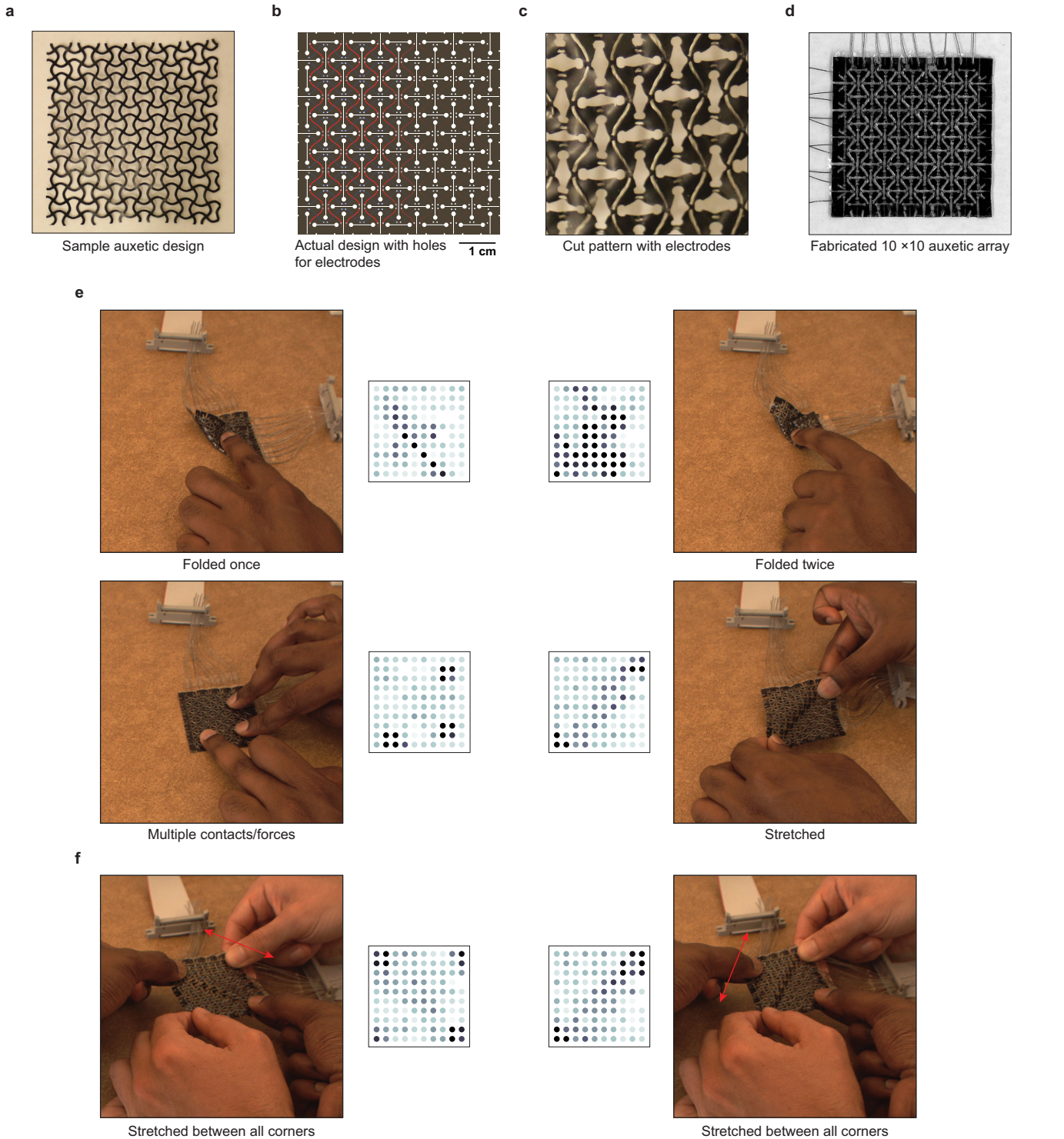
Extended Data Figure 2 | Characteristics of the STAG sensing elements. **a**, The resistance of a single sensing element shows the linear working range (in logarithmic force units). The sensor is not sensitive below about 20 mN of force and saturates in response when a load exceeding 0.8 N is applied. **b**, Response of three separate sensors in the force range 20 mN to 0.5 N. The sensors show minimal hysteresis (17.5 ± 2.8%; see Supplementary Fig. 2). **c**, The sensor response after 10, 100 and 1,000 cycles of linear force ramps up to 0.5 N for three separate devices. The resistance measurements are shown in **d** over the entire set of cycles. **e**, Differential scanning calorimetry measurements of the FSF material shows a two-polymer blend response with softening/melting temperatures of around 100 °C and 115.1 °C. **f**, Through-film resistance of an unloaded sensor after treating at different temperatures in a convection oven for 10 min. The film becomes insulating above about 80 °C.



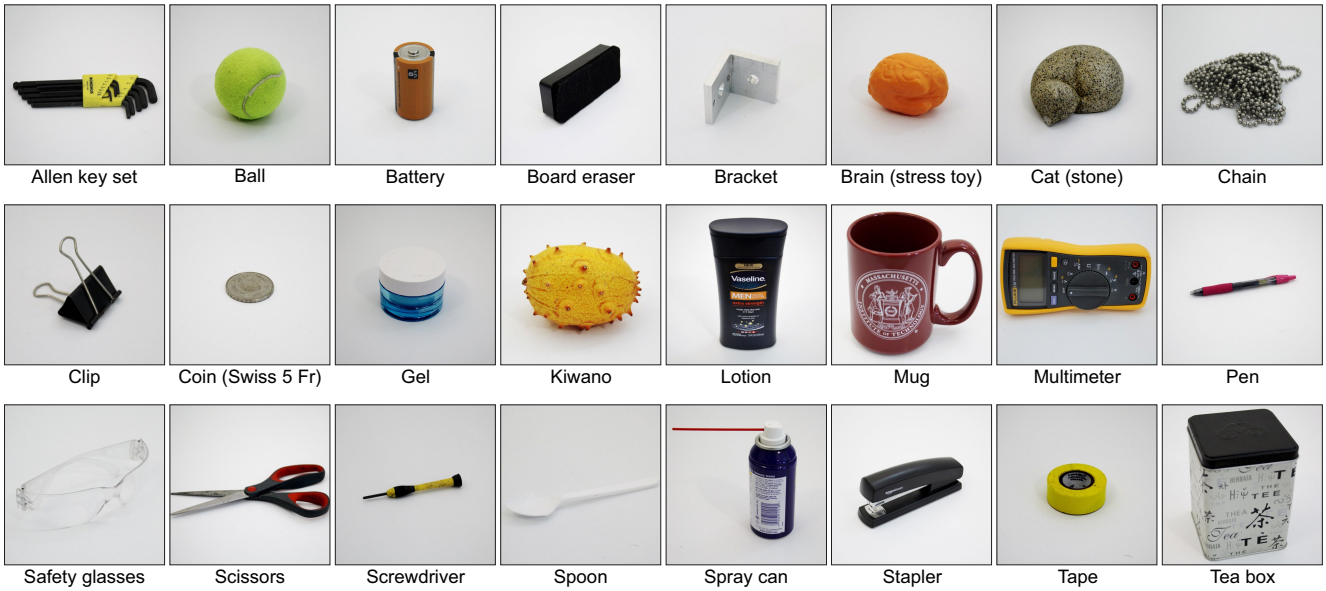
Extended Data Figure 3 | Sensor architectures and regular 32 × 32 arrays. a, A simplified version of the sensor laminate architecture. **b**, The sensor is assembled by laminating a FSF along with orthogonal electrodes on each side, that are held in place and insulated by a layer of two-sided adhesive and a stretchable LDPE film (see Methods). **c**, Fixture used to assemble parallel electrodes. The individual electrodes can be threaded into the structure (like a needle) for assembling parallel electrodes with a spacing of 2.5 mm. **d**, Assembled version of the architecture shown in **a**. **e**, A regular 32 × 32 array version of the STAG based on the design in **b**.



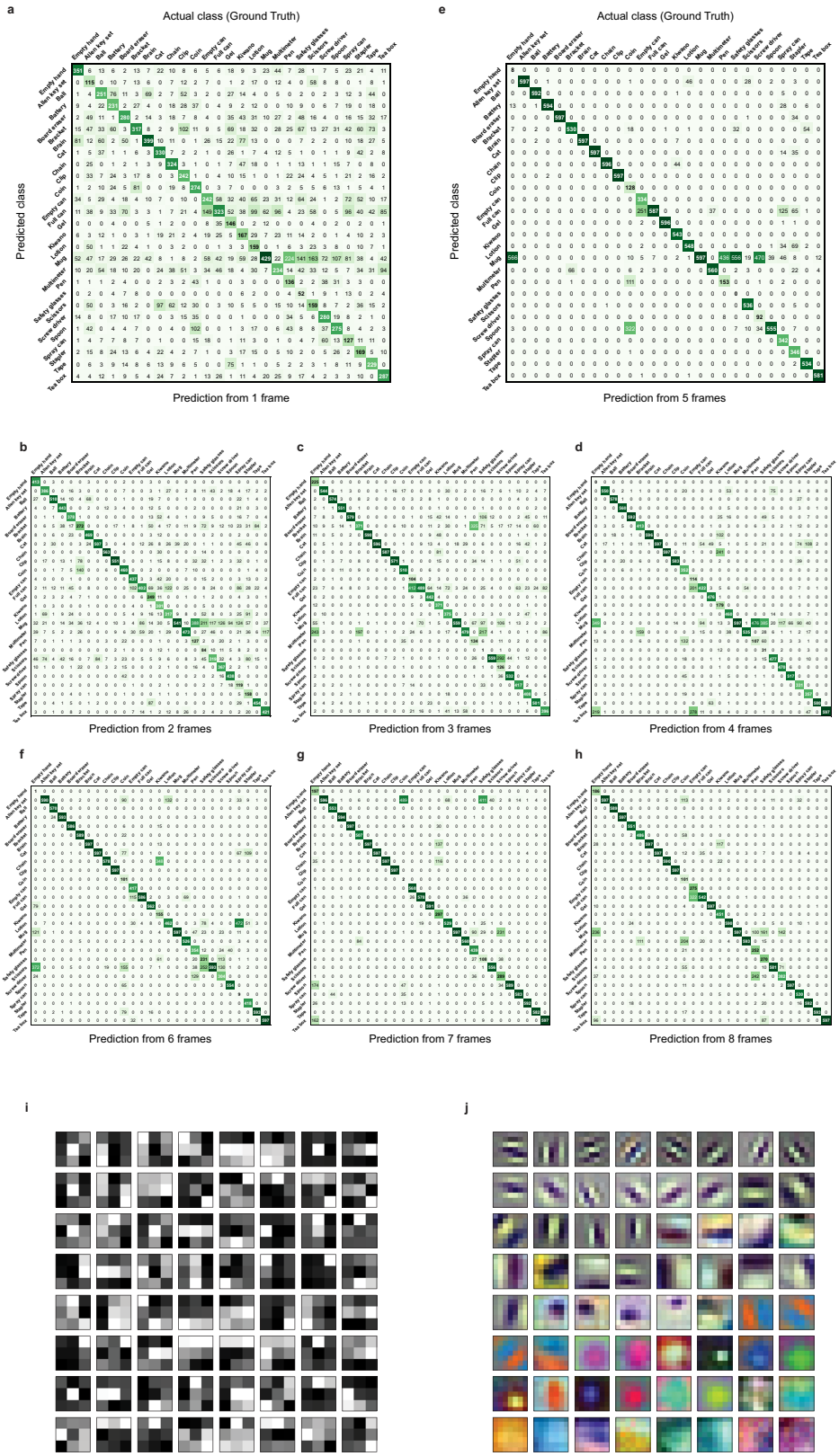
Extended Data Figure 4 | Sample recordings of nine objects on regular 32 × 32 arrays on a flat surface. Nine different objects are manipulated on a regular sensor array (Extended Data Fig. 3d) placed on a flat surface. The resting patterns of these objects can be seen easily. Pressing the tactile array with sharp objects like a pen or the needles of a kiwano yields signals with a single sensor resolution.



Extended Data Figure 5 | Auxetic designs for stretchable sensor arrays. **a**, Standard auxetic design laser cut from the FSF. **b**, The actual design of the auxetic includes holes to route the electrodes (shown in red and blue), and slots allow the square, sensing island to rotate, enhancing the stretchability of the sensor array. **c**, Close-up of the fabricated array showing the conductive thread electrodes before insulation. **d**, A fully fabricated 10 × 10 array with an auxetic design. **e**, Auxetic patterning allows the sensor array to be folded, crushed and stretched easily with no damage. **f**, The array can also be stretched in multiple directions (see Supplementary Video 2).

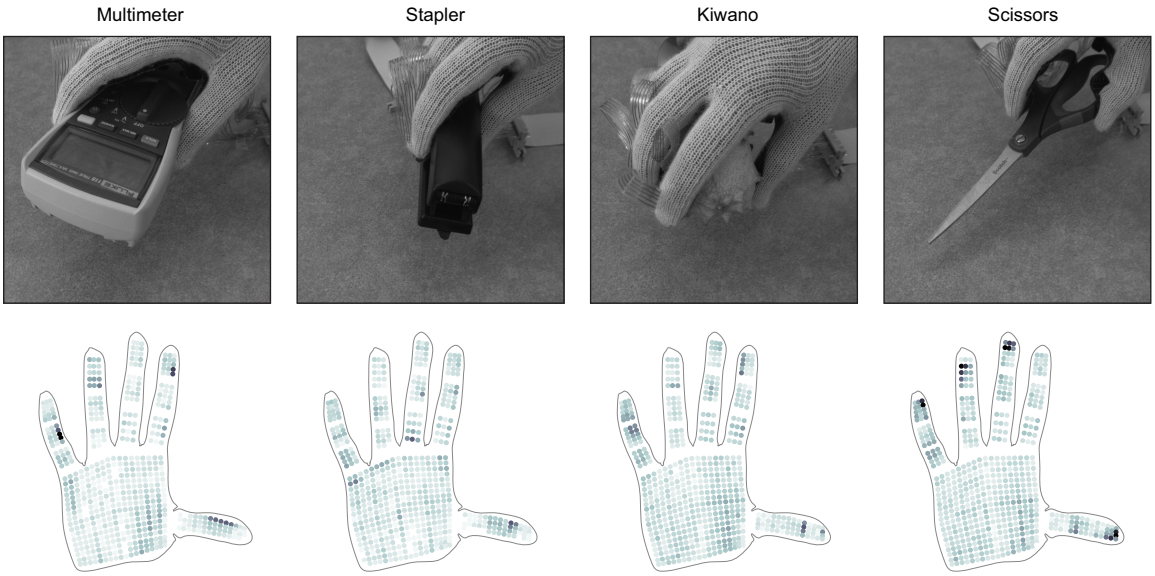


Extended Data Figure 6 | Dataset objects. In total, 26 objects are used in our dataset; images of 24 objects are shown here. In addition to these objects, our dataset includes two cola cans (one empty can and one full can).



Extended Data Figure 7 | Confusion maps and learned convolution filters. a–h, The actual object and predicted object labels are shown in these confusion matrices for different networks, each taking 1 to 8 (or N) inputs where each input is obtained from a distinct cluster for $N > 1$ (approach shown in Fig. 2e; see Methods). These matrices correspond to the 'clustering' curve in Fig. 2b. Objects with similar shapes, sizes or weights are more likely to be confused with each other. For example, the empty can and full can are easily mistaken for each other when they are resting on the table. Likewise, lighter objects such as the safety glasses, plastic spoon, or the coin are more likely to be confused with each other or other objects. Large, heavy objects with distinct signatures such as the tea box have high detection accuracy across different numbers of inputs (N). **i**, Original first-layer convolution filters (3×3) learned by the network shown in Fig. 2a for $N = 1$ inputs. **j**, Visualization of the first-layer convolution filters of ResNet-18 trained on ImageNet.

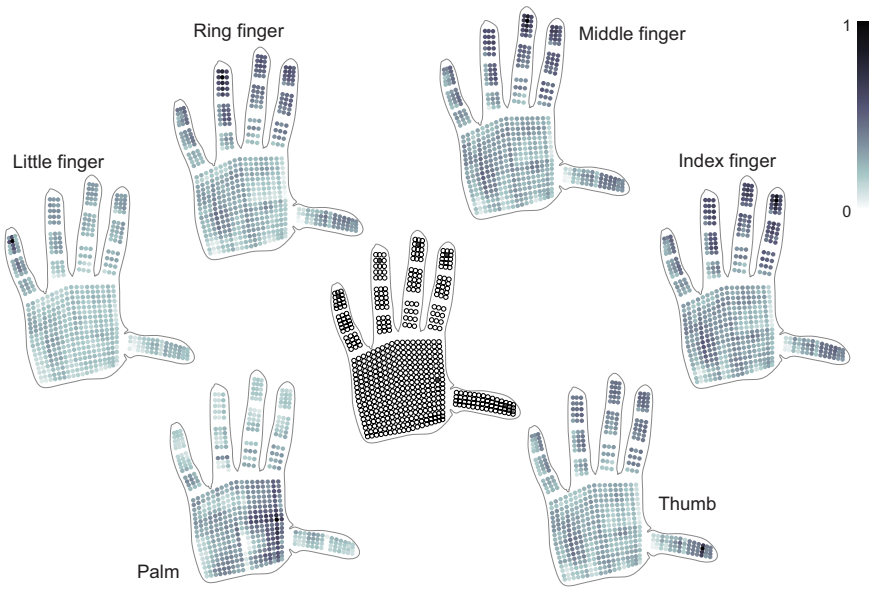
a



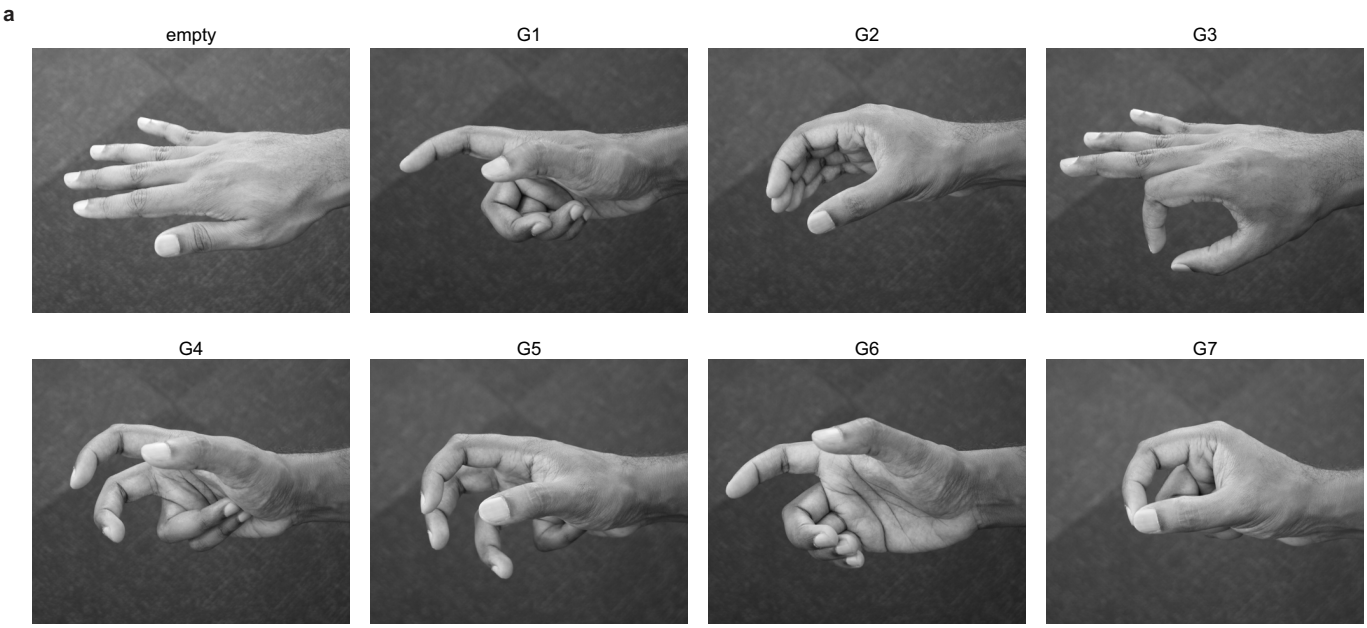
b

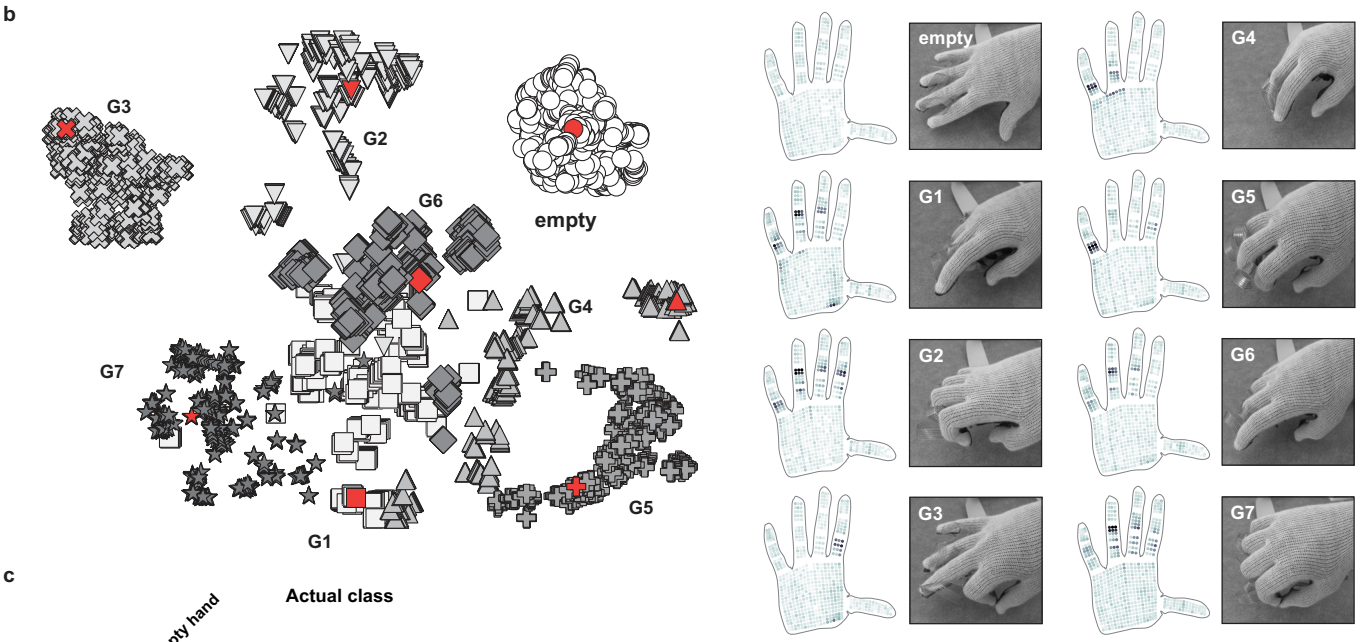
Weight range (g)	Errors					
	Linear baseline		Linear baseline (Hand pose removed)		CNN	
	Abs. (g)	Rel. (Weber)	Abs. (g)	Rel. (Weber)	Abs. (g)	Rel. (Weber)
< 30	57.09	4.63	65.73	5.33	16.49	1.34
31 - 150	83.36	1.07	79.82	1.02	53.30	0.68
151 - 700	136.75	0.45	144.30	0.47	110.97	0.36
Overall error	89.68	2.37	94.24	2.52	56.88	0.69

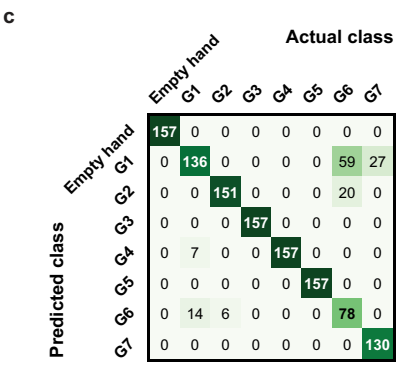
Extended Data Figure 8 | Weight estimation examples and performance. a, Four representative examples from the weight estimation dataset, in which the objects are lifted using multi-finger grasps from the top (see Supplementary Video 6 for an example recording). **b**, The weight estimation performance is shown in terms of the mean absolute and relative errors (normalized to the weight of each object) in each weight interval. The relative error is analogous to the Weber fraction. We observe that the CNN outperforms the linear baseline with or without the hand pose signal removed. The overall errors of the two linear baselines are comparable.



Extended Data Figure 9 | Correspondence maps for six individual sensors using the decomposed hand pose signal. The hand pose signal decomposed from object interactions is used to collectively extract correlations between the sensors and the full hand (analogous to Fig. 3b where the decomposed object-related signal is used). The pixels at the fingertips show less structured correlations with the remaining fingers, unlike in Fig. 3b.







Extended Data Figure 10 | Hand pose signals from articulated hands. **a**, Images of the hand poses used in the hand pose dataset. The poses G1 to G7 are extracted from a recent grasp taxonomy. In the recordings, each pose is continuously articulated from the neutral empty hand pose. **b**, When the tactile data from this dataset is clustered using t-SNE, each distinct group represents a hand pose. Sample tactile maps are shown on the right. The corresponding samples are marked in red (see Supplementary Data 3). **c**, The hand pose signals can be classified with 89.4% accuracy (average of ten runs with 3,080 training frames and 1,256 distinct test frames) using the same CNN architecture shown in Fig. 2a. The confusion matrix elements denote how often each hand pose (column) is classified as one of the possible hand poses (rows). It shows that hand poses G1 and G6 are sometimes misidentified but the other hand poses are identified nearly perfectly.