Learning to be Multimodal : Co-evolving Sensory Modalities and Sensor Properties

Anonymous Author(s)
Affiliation
Address
email

Abstract: Making a single sensory modality precise and robust enough to get human-level performance and autonomy could be very expensive or intractable. Fusing information from multiple sensory modalities is promising – for example, recent works showed benefits from combining vision with haptic sensors or with audio data. Learning-based methods facilitate faster progress in this field by removing the need for manual feature engineering. However, the sensor properties and the choice of sensory modalities is still usually done manually. Our blue-sky view is that we could simulate/emulate sensors with various properties, then infer which properties and combinations of sensors yield the best learning outcomes. This view would incentivize the development of novel, affordable sensors that can make a noticeable impact on the performance, robustness and ease of training classifiers, models and policies for robotics. This would motivate making hardware that provides signals complementary to the existing ones. As a result: we can significantly expand the realm of applicability of the learning-based approaches.

1 Motivation and Overview

Animals have evolved an impressive variety of sensors. A few examples: vision in the visible light spectrum and beyond (snakes seeing in infrared, bumblebees and reindeer in UV); sound emitters/receptors for ranges beyond the audible (elephant low-frequency calls to reach mates miles away; high-frequency echo-location in bats), electroreception for underwater localization (in dolphins, sharks, platypus); sensing Earth's magnetic field for navigation (in pigeons, turtles, salmon). If 'more-is-better' were the main law of sensing, then we would expect to see impressive agglomerations of sensory capabilities in the more advanced animals. Instead, sensors often seem specialized to the class of tasks an animal needs to perform to survive. Thus, sensors likely co-evolved with the organisms and, in part, defined the class of tasks that each organism could accomplish.

To build sensors for robotics we usually do task-agnostic optimization: maximize speed, accuracy and robustness while minimizing size, cost and energy consumption. With the recent success of machine learning, we now have scalable neural networks (NNs) that can learn to construct representations from raw sensor data. This can drastically improve both speed and precision/success rate. For example, NNs can be trained to work on partial and noisy point clouds [1, 2, 3, 4], eliminating the need for shape completion. Shape completion is computationally expensive and can lead to excessive thickening, which in turn limits the potential of active learning approaches (lack of perception speed) and can degrade manipulation with thin objects (lack of point cloud precision).

A simplistic view is that we can include all the available sensory data and rely on end-to-end learning. However, such view could significantly impede the long-term progress. Below we highlight two major issues and propose how to address them by employing a co-evolution approach to sensing.

The need for smart compromises: One major issue is that we cannot literally include in our hardware platforms all possible sensors: this is intractable at least due to cost and size. Hence, we need to make compromises: select a subset of sensors given the budget, size and energy constraints. The choice of which sensors to buy/include is currently done manually. This could have been sufficient in the past: the number of available sensors was small; the scope of the robotics tasks was limited (structured tasks in isolated environments); traditional approaches used sensor data in fixed ways.

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.

For example, for a specific task and known objects one could select a fingertip force sensor that would report normal force that an object exerts on the gripper's finger. This 1D signal could be enough to guide the manipulation if the initial pose of the object was known. In contrast, many robotics researchers now strive to handle objects on unknown shape, material and weight in unstructured environments, striving to achieve this by automatically learning NN-based representations from high-dimensional sensor data, such as RGBD camera images. The downside of this generality is that the learned representations are not comprehensible at a glance. Hence, it is difficult to manually identify sensory modalities and properties that would be necessary and sufficient for a certain class of tasks. As a result, we usually end up including sensors that happen to be available, rather than searching for sensors that would be optimal for the class of tasks we aim to solve.

The need for smart sensing: Suppose we make a bet on sensors becoming smaller and cheaper, hoping to include more and more sensory modalities in the future. Can we then expect that end-to-end learning-from-scratch would suffice? Many academic researchers do not have access to large-scale compute resources, hence their research would be hampered if large-scale training is required in all cases. Generality and flexibility are not always enough for those working with robot hardware and limited computational resources. Hence the need to retain only the sensory data crucial for supporting a given research agenda, and focus compute resources on the relevant signals, without worrying that a key modality is omitted. Smart hardware design could turn even a low-dimensional combination of signals into a highly efficient hybrid sensory modality. For example, in experiments with the Exploratory Hand [5], adding basic tactile sensing greatly improved contact and object identification when combined with contact particle filtering. Roboticists can deduce what kind of sensing could be useful from model-based (physics) considerations. However, they can only test a few target settings/tasks manually. A more streamline way would help to quickly gauge which novel components are effective, which combinations have the potential to produce low-cost and accurate results.

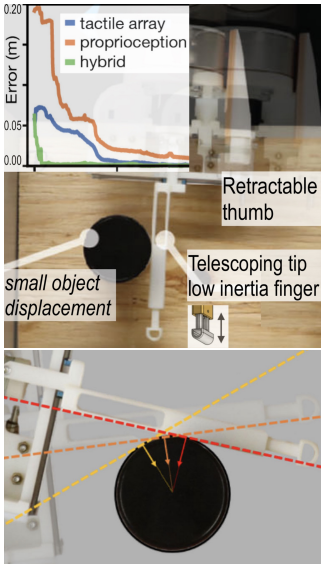

Figure 1: Exploratory hand [5]

1.1 Co-evolving sensors to optimize efficiency of learning algorithms that use sensor data

Our blue-sky paradigm is to integrate development of sensor hardware (and low-level signal processing) with research on using various sensory modalities for learning-based approaches. We can start by creating approximate models of various sensors in simulation, then feed the synthetic sensor data to the target learning methods, and attempt to train on tasks that represent a class of problems of interest. We could vary sensor properties, such as speed and precision, to determine which aspects have significant effects on the training performance. With this, we could infer the minimum requirements for the sensors and find combinations of modalities that enable successful learning. There has been progress in sensor modeling [6, 7, 8], but relying only on simulation is limiting, since it remains challenging to fully close the sim-to-real gap [9]. Hence, after initial simulation-based analysis, it would be crucial to use hardware data. We could collect sensor data from expensive sensors (e.g. high-speed cameras, high-resolution tactile arrays) and identify the maximum level or degradation that still allows learning useful representations. We could also search for optimal combination of sensors: employ sensitivity analysis to determine modalities that contribute the most to the learning success, use explainability methods to show contributions of different sensory modalities.

To motivate the above paradigm, let us consider one concrete challenging example: manipulation of highly deformable objects. More than a decade ago, it has been shown that high-speed cameras and high-speed actuation can help with higly dynamic tasks [10, 11]. However, such systems did not gain widespread use, either due to their high cost or safety concerns of high-speed actuation. We could search for alternative combinations. Perhaps high-speed actuation could be replaced by learning multistep dynamics models: anticipate the motion of the object instead of reacting instantly to its current state. Perhaps such multistep feedforward models would be an inexpensive replacement for the feedback-based approaches that need to rely on high-speed sensing and actuation. We could also test whether feedback-based approaches relying on high-speed cameras can cope with occlusions. If not, we could search for complementary modalities, such as tactile sensing. We could analyze whether adding fingertip sensing is enough. It may turn out that, for some classes of tasks, it is

essential to add sensing to other parts of the robot hands/grippers/arms/body, suggesting the use of sensors closer to an inexpensive 'skin'. If using robotic skin turns out to be highly beneficial, we could identify the minimum requirements for such sensors for a given set of tasks. This paradigm would enable finding an optimal combination of the properties of a vision system (speed, number of cameras) and tactile system (resolution, placement). It would allow to jointly minimize the sensor cost and size requirements, while still ensuring that the learning algorithms get a signal that is sufficiently rich for accomplishing a set of target tasks. We might discover that the well-performing combinations require sensors that have not been built yet. For example, we might find that reporting full contact information with all parts of the grippers in simulation is highly beneficial for successful training, hence demonstrating the potential benefits of developing a sensitive skin that can cover a large part of the grippers. By showing that learning succeeds when using such sensors we could help the teams working on sensor hardware to obtain additional funding and support for the development.

2 Acquiring Novel Sensors and Sensory Modalities

High-speed / high-coverage sensors: [10] showed impressive results on dynamic tasks by combining fast actuation (180 degrees per 0.1 second) and high-speed cameras with 500 frames per second (FPS) rate to solve advanced tasks, such as in-the-air cloth folding and knot tying. High-speed tactile sensors (1kHz) enabled dynamic in-hand manipulation tasks, such as pen spinning [12]. LIDAR sensors have been tested at high speeds for race driving (>150 km/h), showing potential for high-speed localization [13]. High-speed vision sensors is a recurring interest in robotics and computer vision. However, in addition to being expensive, systems beyond 1000 FPS face the challenges of caching and transmitting frames at high rates [14].

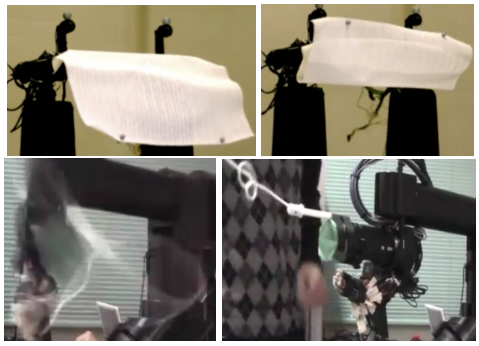

Figure 2: In-the-air cloth folding & knot tying [10]

One way to address this is to compress the signal at the time of sensing [17]. Another solution is event-based methods: report only the salient changes [18, 19, 20]. This could allow high-resolution sensing without overwhelming the learning approaches with redundant data. This could be particularly beneficial for high-coverage skin sensors, since we know that humans rely on such ability (e.g. event-based aspects of fast-adapting mechanoreceptors on human fingertips [21]; sensing clothing only when we initially put it on). If we optimize such functionality separately from understanding how the signals are used by the learning methods – we risk creating input data that is not easy for the learning methods to interpret. Adopting the paradigm of co-evolving the sensor properties together with the learning methods would allow us to ensure the overall training success.

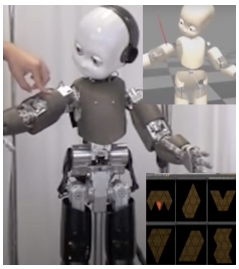

Figure 3: 2000 skin receptors on iCub [15, 16]

More than the sum of its parts: A number of recent works showed that combining multiple sensory modalities is beneficial. Vision (from conventional 30Hz cameras) has been combined with tactile and haptic sensors [22, 23, 24, 25, 26], range sensors [27, 28], auditory data [29, 30, 31, 32, 33]. Withholding certain modalities can be detrimental even for the most powerful and adaptive learning systems – the human brain. Research in prosthetics shows that the absence of tactile information can lower device adoption rates [34]; incorporating tactile sensing is important for making it easy for the brain to adapt to using the hand and arm prosthetics [35, 36]. As the available learning algorithms evolve, it is important to keep track of which modalities are necessary and sufficient for the new state-of-the-art learning algorithms to succeed. Following the co-evolution paradigm would allow us to jointly optimize the individual sensor properties and experiment with various combinations of sensory modalities. As a result, we could discover the need for new hybrid sensory modalities. For example, if high-speed visuo-tactile sensors show promise, but face issues with caching, transmission and synchronization – we could argue for developing an integrated sensing system. Placing vision and tactile sensing on the same platform/chip would reduce cost, communication overhead, improve synchronization, allow integrated compression and event-driven transmission (FingerVision [37] showed some of these benefits). The co-evolution paradigm would let us present a convincing case by clearly showing the expected benefits on the learning system. This could strengthen

the motivation and enthusiasm for funding the hardware development. Once affordable hardware is available, the learning-for-robotics approaches would be able to show the benefits in hardware experiments, thus furthering the adoption of advanced learning-based methods in core robotics.

The 6th sense and beyond: It can be easy to settle into a local optimum of using a common inexpensive set of visual and tactile sensors. To overcome this, the co-evolution view can help perform initial estimation of what the novel sensing modalities could offer. Consider the recently developed techniques of 'seeing through walls' with WiFi [38] and radio signals [39]. We could make a rough model of these by post-processing real fully-observable data to approximate the quality of the data that such sensors would output under occlusions. We can then quickly find out whether such novel sensors can significantly improve the performance of the learning approaches. Similarly, we can explore other unconventional sensors, for example, the miniature versions of spectrometers [40, 41]. Finally, we can compare the benefits of multistep predictive models versus using high-speed hardware sensors. If multistep models can anticipate the near future well enough to replace a high-speed sensor – then we can spend resources on acquiring data and compute power to train such models.

3 Optimization for the Co-evolution of Sensors

Global search for optimal sensing: To avoid the difficulties of modeling the physics of sensors, we can instead create simulated data streams that match the precision, speed and basic noise patterns of existing or hypothetical/new sensors. To find combinations of sensory modalities and sensor properties that yield the best performance, we can use data-efficient global search methods, e.g. a recent variant of Bayesian optimization (BO) for mixed categorical-continuous spaces [42]. We can start by sampling a batch of N sensor combinations $\{s_1^{(n)}, ..., s_k^{(n)}\}_{n=1}^N$ along with their properties $\{x_1^{(n)}, ..., x_k^{(n)}\}_{n=1}^N$, where each x is a vector of sensor characteristics (sampling rate, noise level, resolution, etc). We can launch N training runs in parallel on a set of target tasks, then add the results to the BO posterior. This posterior would model the objective function $f(combo)$, expressing the expected performance of the learning algorithms that use a given combination of sensors. With BO we also get uncertainty estimates. If the posterior indicates benefits from a novel combination of sensors, we can run the optimization further until the uncertainty shrinks enough to present a clear case for purchasing the new sensors or developing the new hardware. More generally, we can view the search for optimal combination of sensors as hierarchical hyperparameter optimization and try a variety of relevant methods [43, 44]. Multi-objective optimization [45] could find a Pareto frontier for optimizing performance versus cost to accommodate various budget levels.

Incorporating real data: To avoid over-reliance on simulation, we need to use real sensor data when possible. For example, we could construct a dataset that pertains to the aspects of interest/tasks from a high-resolution/speed sensor. Data collection could be done in a lab that develops the high-resolution/speed sensors. Other labs can use the dataset to mimic lower-resolution/speed (by degrading the sensor readings) to find the sensor resolution/speed that is sufficient for learning.

4 Relevant Work in Morphological Computation and Explainability

Co-evolving morphology: Recent works in searching for optimal morphology provide a level of assurance that our proposed co-evolution approach would be tractable in practice. [46] provide a way to evolve morphology of a rigid hand; [47, 48] develop automated co-design of soft hand morphology and grasping controllers; [49] propose a way to specialize sensor readings for a soft hand for new tasks. While these works do not suggest a longer-term strategy for sensor evolution, they do show that co-evolving hardware and algorithms (e.g. controllers) is beneficial and tractable. The classical sensor selection problem [50, 51] is related, though it views sensor properties as fixed.

Explainability: To analyze the contributions of various sensor modalities we can leverage recent work in explainability for deep learning [53, 54, 55, 56, 57]. This would be especially useful for the versions of sensors that appear promising but have not been widely adopted yet. Visualizing which aspects of the sensor readings cause a large impact on the learned models would allow sensor developers to focus on enhancing these aspects.

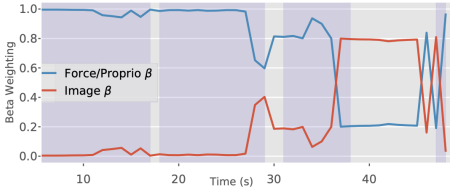

Figure 4: Learning when to use force readings (blue) on contact (shaded) vs images (red) [52]

References

[1] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[2] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, 2017.

[3] P. Schmidt, N. Vahrenkamp, M. Wächter, and T. Asfour. Grasping of unknown objects using deep convolutional neural networks based on depth images. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[4] R. Antonova, M. Kokic, J. A. Stork, and D. Kragic. Global search with bernoulli alternation kernel for task-oriented grasping informed by simulation. In *Proceedings of The 2nd Conference on Robot Learning (CoRL)*. PMLR, 2018.

[5] M. A. Lin, R. Thomasson, G. Uribe, H. Choi, and M. R. Cutkosky. Exploratory hand: Leveraging safe contact to facilitate manipulation in cluttered spaces. *IEEE Robotics and Automation Letters*, 6(3):5159–5166, 2021.

[6] W. Guerra, E. Tal, V. Murali, G. Ryou, and S. Karaman. Flightgoggles: Photorealistic sensor simulation for perception-driven robotics using photogrammetry and virtual reality. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6941–6948. IEEE, 2019.

[7] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra. Tacto: A fast, flexible and open-source simulator for high-resolution vision-based tactile sensors. *arXiv preprint arXiv:2012.08456*, 2020.

[8] C. Sweeney, G. Izatt, and R. Tedrake. A supervised approach to predicting noise in depth images. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 796–802. IEEE, 2019.

[9] S. Höfer, K. Bekris, A. Handa, J. C. Gamboa, F. Golemo, M. Mozifian, C. Atkeson, D. Fox, K. Goldberg, J. Leonard, et al. Perspectives on sim2real transfer for robotics: A summary of the r: Ss 2020 workshop. *arXiv preprint arXiv:2012.03806*, 2020.

[10] Y. Yamakawa, A. Namiki, and M. Ishikawa. Motion planning for dynamic folding of a cloth with two high-speed robot hands and two high-speed sliders. In *2011 IEEE International Conference on Robotics and Automation*, pages 5486–5491. IEEE, 2011.

[11] Y. Yamakawa, A. Namiki, and M. Ishikawa. Motion planning for dynamic knotting of a flexible rope with a high-speed robot arm. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 49–54. IEEE, 2010.

[12] T. Ishihara, A. Namiki, M. Ishikawa, and M. Shimojo. Dynamic pen spinning using a high-speed multifingered hand with high-speed tactile sensor. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*, pages 258–263. IEEE, 2006.

[13] T. Stahl, A. Wischnewski, J. Betz, and M. Lienkamp. Ros-based localization of a race vehicle at high-speed using lidar. In *E3S Web of Conferences*, volume 95, page 04002, 2019.

[14] Q.-Y. Gu and I. Ishii. Review of some advances and applications in real-time high-speed vision: Our views and experiences. *International Journal of Automation and Computing*, 13 (4):305–318, 2016.

[15] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. Von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, et al. The icub humanoid robot: An open-systems platform for research in cognitive development. *Neural networks*, 23(8-9):1125–1134, 2010.

[16] M. Hoffmann, Z. Straka, I. Farkaš, M. Vavrečka, and G. Metta. Robotic homunculus: Learning of artificial skin representation in a humanoid robot motivated by primary somatosensory cortex. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2):163–176, 2017.

[17] M. Rani, S. B. Dhok, and R. B. Deshmukh. A systematic review of compressive sensing: Concepts, implementations and applications. *IEEE Access*, 6:4875–4894, 2018.

[18] E. Müggler. *Event-based vision for high-speed robotics*. PhD thesis, University of Zurich, 2017.

[19] G. Gallego, T. Delbruck, G. M. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[20] T. Taunyazov, W. Sng, H. H. See, B. Lim, J. Kuan, A. F. Ansari, B. C. Tee, and H. Soh. Event-driven visual-tactile sensing and learning for robots. 2020.

[21] R. S. Johansson. Tactile sensibility in the human hand: receptive field characteristics of mechanoreceptive units in the glabrous skin area. *The Journal of physiology*, 281(1):101–125, 1978.

[22] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36(3):582–596, 2020.

[23] Y. Bekiroglu, R. Detry, and D. Kragic. Learning tactile characterizations of object-and pose-specific grasps. In *2011 IEEE/RSJ international conference on Intelligent Robots and Systems*, pages 1554–1560. IEEE, 2011.

[24] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[25] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell. Deep learning for tactile understanding from visual and haptic data. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 536–543. IEEE, 2016.

[26] J. Sinapov, C. Schenck, and A. Stoytchev. Learning relational object categories using behavioral exploration and multimodal perception. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5691–5698. IEEE, 2014.

[27] G.-H. Liu, A. Siravuru, S. Prabhakar, M. Veloso, and G. Kantor. Learning end-to-end multimodal sensor policies for autonomous navigation. In *Conference on Robot Learning*, pages 249–261. PMLR, 2017.

[28] J. Sung, J. K. Salisbury, and A. Saxena. Learning to represent haptic feedback for partially-observable tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2802–2809. IEEE, 2017.

[29] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[30] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[31] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016.

[32] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo. Deep multimodal representation learning from temporal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5455, 2017.

[33] C. Matl, Y. Narang, D. Fox, R. Bajcsy, and F. Ramos. Stressd: Sim-to-real from sound for stochastic dynamics. 2020.

[34] B. T. Nghiem, I. C. Sando, R. B. Gillespie, B. L. McLaughlin, G. J. Gerling, N. B. Langhals, M. G. Urbanchek, and P. S. Cederna. Providing a sense of touch to prosthetic hands. *Plastic and reconstructive surgery*, 135(6):1652–1663, 2015.

[35] R. Darie, M. Powell, and D. Borton. Delivering the sense of touch to the human brain. *Neuron*, 93(4):728–730, 2017.

[36] S. N. Flesher, J. E. Downey, J. M. Weiss, C. L. Hughes, A. J. Herrera, E. C. Tyler-Kabara, M. L. Boninger, J. L. Collinger, and R. A. Gaunt. A brain-computer interface that evokes tactile sensations improves robotic arm control. *Science*, 372(6544):831–836, 2021.

[37] A. Yamaguchi and C. G. Atkeson. Tactile behaviors with the vision-based tactile sensor fingervision. *International Journal of Humanoid Robotics*, 16(03):1940002, 2019.

[38] F. Adib and D. Katabi. See through walls with wifi! In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, pages 75–86, 2013.

[39] N. Savage. Seeing through walls. *Communications of the ACM*, 63(6):15–16, 2020.

[40] M. Li, Z. Qian, B. Shi, J. Medlicott, and A. East. Evaluating the performance of a consumer scale scio molecular sensor to predict quality of horticultural products. *Postharvest Biology and Technology*, 145:183–192, 2018.

[41] V. Wiedemair and C. W. Huck. Evaluation of the performance of three hand-held near-infrared spectrometer through investigation of total antioxidant capacity in gluten-free grains. *Talanta*, 189:233–240, 2018.

[42] B. Ru, A. Alvi, V. Nguyen, M. A. Osborne, and S. Roberts. Bayesian optimisation over multiple continuous and categorical inputs. In *International Conference on Machine Learning*, pages 8276–8285. PMLR, 2020.

[43] J.-C. Lévesque, A. Durand, C. Gagné, and R. Sabourin. Bayesian optimization for conditional hyperparameter spaces. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 286–293. IEEE, 2017.

[44] L. Yang and A. Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[45] L. Brevault, J. Pelamatti, A. Hebbal, M. Balesdent, E.-G. Talbi, and N. Melab. Mdo related issues: Multi-objective and mixed continuous/discrete optimization. In *Aerospace System Analysis and Optimization in Uncertainty*, pages 321–358. Springer, 2020.

[46] X. Pan, A. Garg, A. Anandkumar, and Y. Zhu. Emergent hand morphology and control from optimizing robust grasps of diverse objects. *arXiv preprint arXiv:2012.12209*, 2020.

[47] K. Ghazi-Zahedi, R. Deimel, G. Montúfar, V. Wall, and O. Brock. Morphological computation: the good, the bad, and the ugly. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 464–469. IEEE, 2017.

[48] R. Deimel, P. Irmisch, V. Wall, and O. Brock. Automated co-design of soft hand morphology and control strategy for grasping. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1213–1218. IEEE, 2017.

[49] V. Wall and O. Brock. Multi-task sensorization of soft actuators using prior knowledge. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9416–9421. IEEE, 2019.

[50] H. Rowaihy, S. Eswaran, M. Johnson, D. Verma, A. Bar-Noy, T. Brown, and T. La Porta. A survey of sensor selection schemes in wireless sensor networks. In *Unattended Ground, Sea, and Air Sensor Technologies and Applications IX*, volume 6562, page 65621A. International Society for Optics and Photonics, 2007.

[51] A. M. Elbir and K. V. Mishra. Sparse array selection across arbitrary sensor geometries with deep transfer learning. *IEEE Transactions on Cognitive Communications and Networking*, 2020.

[52] M. A. Lee, B. Yi, R. Martín-Martín, S. Savarese, and J. Bohg. Multimodal sensor fusion with differentiable filters. *arXiv preprint arXiv:2010.13021*, 2020.

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[54] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[55] NeurIPS 2020 Tutorial . Explaining ML Predictions: State-of-the-art, Challenges, and Opportunities. https://explainml-tutorial.github.io/neurips20, 2020. Accessed: 2021-06-06.

[56] N. Burkart and M. F. Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

[57] M. H. Danesh, A. Koul, A. Fern, and S. Khorram. Re-understanding finite-state representations of recurrent policy networks. In *International Conference on Machine Learning*, pages 2388–2397. PMLR, 2021.