Surface Perception through Haptic-Auditory Contact Data

Behnam Khojasteh[1] , Yitian Shao[1,2], and Katherine J. Kuchenbecker[1]

***Abstract*— Sliding a finger or tool along a surface generates rich haptic and auditory contact signals that encode properties crucial for manipulation, such as friction and hardness. To engage in contact-rich manipulation, future robots would benefit from having surface-characterization capabilities similar to humans, but the optimal sensing configuration is not yet known. Thus, we developed a test bed for capturing high-quality measurements as a human touches surfaces with different tools: it includes optical motion capture, a force/torque sensor under the surface sample, high-bandwidth accelerometers on the tool and the fingertip, and a high-fidelity microphone. After recording data from three tool diameters and nine surfaces, we describe a surface-classification pipeline that uses the maximum mean discrepancy (MMD) to compare newly gathered data to each surface in our known library. The results achieved under several pipeline variations are compared, and future investigations are outlined.**

I. Motivation

The biological sensing and transduction processes that occur during finger-surface and tool-surface interactions are remarkably sophisticated, enabling humans to perform ubiquitous tasks such as fine material discrimination and dexterous manipulation. Accurate surface perception is often a necessary step toward targeted and effective object manipulation, as motor commands need to be adjusted to fit the physical interaction taking place. Classifying surfaces can involve the identification and categorization of different types of surfaces based on their physical properties, such as roughness, friction, hardness, and shape.

It is an important challenge for machine perception to try to capture and process the rich contact signals elicited during surface exploration with a level of success similar to humans. Prior research introduced a diverse set of surface-sensing systems [1], [2], [3], but it is not clear *what combination, quality, resolution, and acuity of sensor data are necessary* to reach the efficiency and accuracy of humans. Contact vibrations in particular offer high temporal transient resolution for spatial touch information decoding [4] and for effective multimodal surface classification [5]. Similarly advantageous is the fact that accelerometers are compact, low-cost, energy-efficient sensors with simple mounting, a straightforward electrical interface, and easy calibration.

To increase our understanding about artificial surface perception, we have designed a novel haptic-auditory test bed to capture data while a human explores a variety of surface samples with either a handheld tool or their fingertip. For the goal of effective computational sensing, we leverage ideas from recognizing surfaces with kernel mean embeddings [5] to elucidate the configuration of sensor information that are necessary for highly accurate surface recognition. We further propose methods for out-of-distribution tasks in settings with different sensing tools with the goal to generalize beyond training data. By leveraging modern machine-learning techniques, we believe haptic and auditory time series can be used more efficiently for surface perception in the absence of any visual information, thereby enabling industrial applications in manufacturing, material processing, and inspection. We anticipate that the eventual findings from this ongoing research project can help guide the design of new artificial sensing tools and robotic hands.

II. Haptic-Auditory Sensing

To investigate the mechanical basis of surface encoding, we designed a novel high-fidelity measurement apparatus to capture haptic-auditory data from surface interactions.

A. Test Bed

Our test bed consists of an optical motion-capture system (Vicon, Vantage 5), two miniature high-bandwidth accelerometers (STMicroelectronics, IIS3DWB), a six-axis force/torque sensor (ATI Industrial Automation Inc., Nano43), and a high-fidelity microphone (Brüel and Kjaer, 4955) to capture relevant haptic and auditory data from contact interactions (Fig. 1(a)). The assembly of the selected texture, a rigid platform, and the force-torque sensor underneath are mounted through a plate on an optical table (Thorlabs Inc.). In preparation for our computational analysis, we calculate the tool-tip position, tool-tip speed, and tool orientation from the motion-tracking system (mot); three-axis contact forces (for) and torques (tor); 3D tactile accelerations on the tool (act) and finger (acf); and contact sound from the microphone (mic); sample data are shown in Fig. 1(a).

B. Data Acquisition

For this study, we selected a set of nine surface textures (Fig. 1(b)) inspired by prior surface datasets [1], [2], [3]. In addition, we considered three hemispherical steel tools with thermally hardened tool tips of 4, 6, and 8 mm diameter (Fig. 1(c)). During data acquisition, an experimenter recorded rich intra-class surface data between the selected tool and surface by varying their speed and applied normal force (Fig. 1(d)). They were asked to choose a free but circular motion in order to capture rich signals from different contact conditions and phenomena. From two long data recordings for each surface c, we extracted ten trials that are each five

[1] B. Khojasteh, Y. Shao and K. J. Kuchenbecker are with the Max Planck Institute for Intelligent Systems (MPI-IS), Stuttgart, Germany. e mail: {khojasteh, ytshao, kjk}@is.mpg.de

[2] Y. Shao is also with the Centre for Tactile Internet with Human-in-the-Loop (CeTI), Technische Universität Dresden, Dresden, Germany.

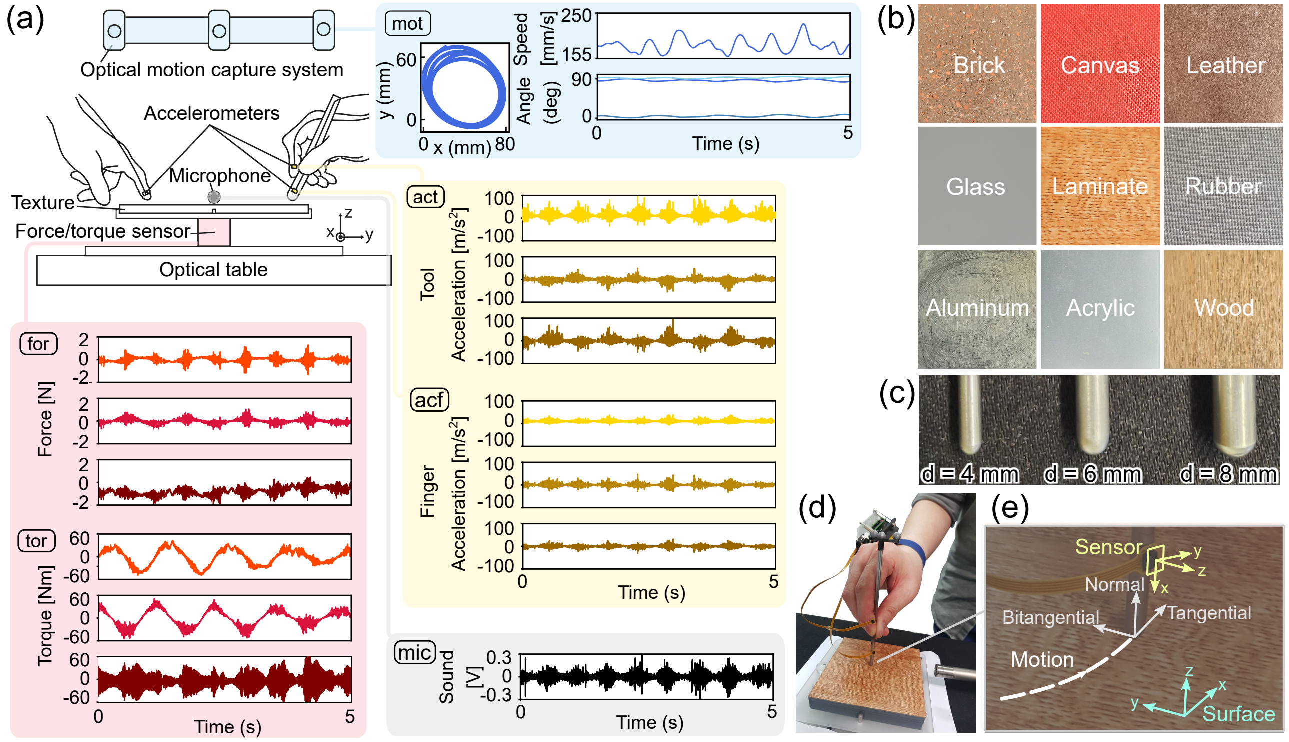

Fig. 1. (a) Test bed and sample recording from each sensor for a steel tool (d = 6 mm) dragging on the wooden surface. (b) Set of nine diverse surfaces. (c) Three steel tools with different diameters. (d) Recording setup. (e) Coordinate system definitions.

seconds long without any transient artifacts. Thus, in total we have 270 trials of multimodal surface data (9 surfaces × 3 steel tools × 10 trials).

III. Machine Surface Perception Studies

To provide insights into both the hardware and the software sides of artificial texture perception, we propose several studies for haptic-auditory surface data collected with our text bed (Fig. 2(a)). Our approach leverages ideas from Khojasteh et al.'s recent work on automatically classifying multimodal surface data [5].

A. Multi-Axis Signal Representation

We plan to represent multi-directional signals such as the three-axis accelerations in three different coordinate systems: sensor, motion, and surface, as defined in Fig. 1(e). The motion-capture system provides the tool-tip position and tool orientation needed to transform the tool accelerometer signals into the motion and surface coordinate systems. For the motion coordinate system, we therefore can represent the 3D accelerations and forces in the tangential, bitangential, and normal directions of the tool tip in motion, which are expected to facilitate physical interpretation.

B. Noise Contamination

We plan to evaluate the robustness of our sensors and processing pipeline by contaminating the sensor readings with noise from a normal distribution $\mathcal{N}(0, \epsilon^2)$. This noise contamination may represent a variety of real-world scenarios, such as a sensor's inherent noise (digital vs. analog) or environmental noise.

C. Time-Series Subsampling

We aim to quantify differences in the distributions of contact data recorded from two different surface interactions, which may come from identical or non-identical surfaces. We extract data points from the time-series data in the time or frequency domain [5]. For temporal subsampling, we equidistantly sample n data points from both time-series sources (Fig. 2(b)). For spectral subsampling, we randomly select the n magnitudes from the discrete Fourier transforms of the signals at the same frequencies (Fig. 2(c)).

D. Out-of-Domain Prediction for Sensing Tools

It is of practical interest to study the generalization capabilities of surface perception to unseen sensing tools so that a given pipeline could be deployed on different instruments and different robot body parts (e.g., from smallest to largest robot finger). Khojasteh et al.'s recent holistic data-driven approach demonstrated the effective mitigation of speed-, force-, and session-dependent effects during tool-surface interaction by manipulating distributions of time-series data [5]. In that surface similarity engine, a simple distribution mean alignment was sufficient to remove these effects and boost recognition performance, presumably because this shift highlighted the distributions' higher-order statistical moments that convey important information about the surface properties. This method may also be relevant for generalization to unseen tools, which we term "out-of-domain prediction".

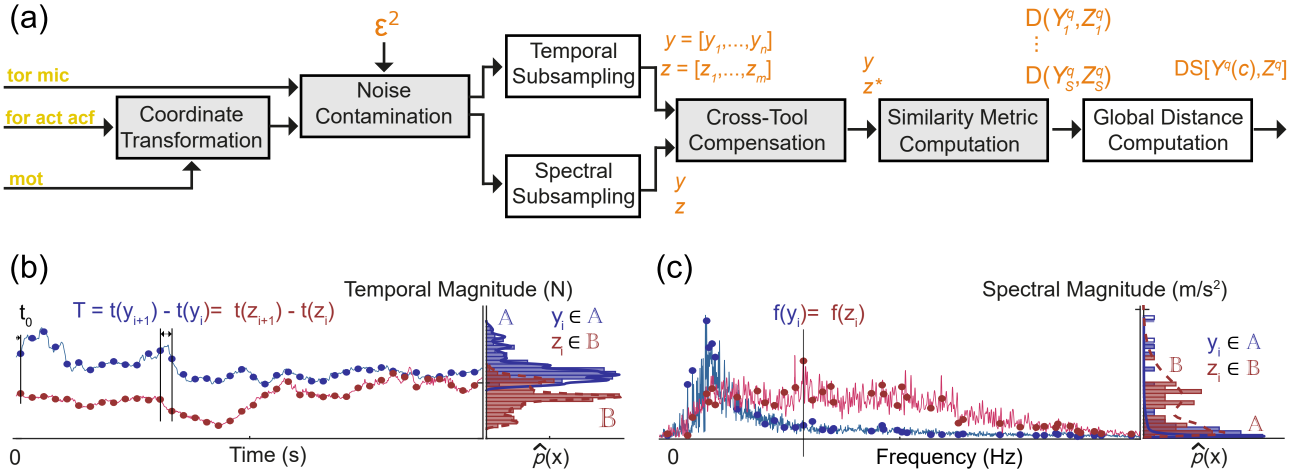

Fig. 2. (a) Proposed pipeline for studying haptic-audio surface encoding. (b) Temporal and (c) spectral subsampling of time-series data from two surfaces.

E. Quantifying Distribution Differences

An efficient way to classify multimodal surface data (e.g., images, haptic signals, and sounds) is to quantify the difference between the distributions of the data sources [5]. For quantifying distribution differences, we focus on kernel-based statistical tests [6], [7], [8] and another notion of divergence with generalized entropy [9].

Maximum Mean Discrepancy: The maximum mean discrepancy (MMD) is a metric to quantify the distance between two probability distributions by considering all their statistical moments in a high-dimensional space. The embedding of probability distributions is called kernel mean embeddings and can be approximated by kernel-based estimates (the so-called kernel trick). We use the computationally efficient MMD estimator by Gretton et al. [7],

$$\mathrm{MMD}_b^2[\mathbb{P}_Y, \mathbb{P}_Z] = \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j) \\ + \frac{1}{m^2}\sum_{i,j=1}^{m} k(z_i, z_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(y_i, z_j), \tag{1}$$

where $[y_1, ..., y_n]$ and $[z_1, ..., z_m]$ are i.i.d. random variables. In our case, these are n and m samples from surface data streams Y_s^q and Z_s^q with unknown distributions $\mathbb{P}_{Y_s^q}$ and $\mathbb{P}_{Z_s^q}$. This MMD estimator is well suited for recognition of surface classes from multimodal data [5].

Other Kernel-Mean-Embedding Metrics: Jitkrittum et al. propose the mean embedding (ME) test and the smoothed characteristic function (SCF) test, which we plan to evaluate in the future for higher interpretability [8]. Unlike the squared-time MMD test, the linear-time ME test evaluates the squared-exponential-kernel mean embeddings at optimized test locations chosen to maximize distribution differences, while also taking variance into account. Similarly, smoothed characteristic functions are distances based on optimized frequency locations. Jitkrittum et al. present the theoretical foundation and implementation details for both tests [8].

H-Divergence Discrepancy Measure: Zhao et al. propose a generalized class of divergences for two-sample testing that includes existing measures such as the MMD and H-Jensen Shannon Divergence [9]. The H-divergence measure directly takes the loss function into account based on the decision space. This new measure outperforms several existing test statistics in terms of test power in multiple real-world experiments, so we plan to check its suitability for our application.

IV. Experiments

We adopt the same setting as Khojasteh et al. [5], who achieved competitive classification results on a public database of 108 textures [2]. The recognition settings for the preliminary results (Fig. 3) and future steps are presented in the following.

A. Recognition Settings

We pursue a random spectral subsampling strategy for all information sources (for, tor, act, acf, mic), as they all exhibit salient AC components. Inspired by human sensing capabilities, we chose the frequency range to be 0–1 kHz for tactile and 0–20 kHz for auditory vibrations. The frequency range for subsampling forces and torques is 0–2 kHz, below all resonant frequencies of the sensor. Our subsample size is $n = m = 500$. We repeatedly compute our MMD scores for each trial-to-trial comparison and each information source $R = 10$ times for subsampling random spectral magnitudes The classification accuracy is reported for the mean and standard deviation of the R repetitions in various conditions. We obtain our global distances ("all") by multiplying the MMD scores of all five information sources. Classification decision is made by selecting smallest distance between the test instance and all training instances (1-nearest-neighbor).

In this work, we conduct classification by testing with five of the ten trials collected for each surface-tool pair; when focusing on a single tool diameter, every trial in the testing set is compared to a library of 45 other trials (analogous to the training set in other approaches). In our out-of-domain setting, we test the data from interactions with both the 4 mm

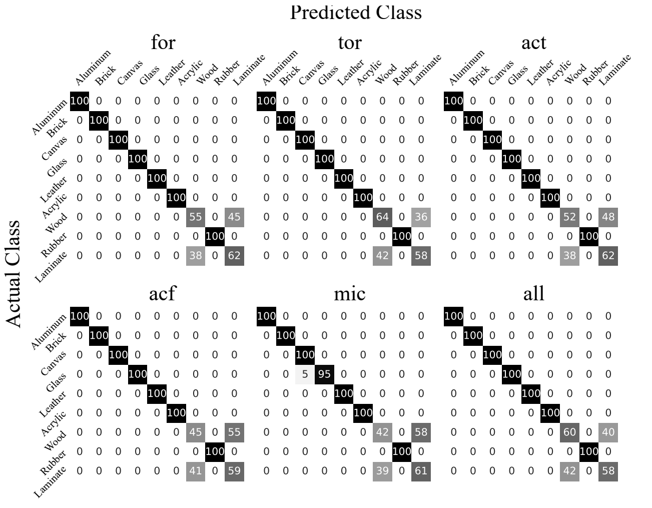

Fig. 3. Confusion matrices for the five individual information sources (force, torque, tool accelerometer, finger accelerometer, and microphone) and their combination (all). Perfect surface recognition would be a solid black diagonal.

and 8 mm tools by relying on the training data from the other tool diameter, i.e., 6 mm. Here, we also evaluate whether aligning the distribution means [5] is a promising approach for out-of-domain recognition.

B. Preliminary Results

From the five information-specific confusion matrices (Fig. 3), we find that almost all confusions occur between the two wooden surfaces, wood and laminate. Human perception of these surfaces matches the MMD-based inference for their similarity. The torque readings (tor: $91.3 \pm 1.2\%$) achieve the best recognition accuracy, followed by the contact forces (for: $90.8 \pm 1.4\%$). The better performance of torque sensing in this recognition setting may result from the different thickness of the two wooden surfaces, as the same tangential contact forces cause different torques through their respective lever arms. When we tried using temporal instead of spectral subsampling for force-torque data, we observed similar performance (for: $91.4 \pm 1.2\%$, tor: $91.3 \pm 1.0\%$), validating the utility of both the DC and AC components of these data streams. The tool acceleration data results in a slightly higher recognition rate (act: $90.4 \pm 1.3\%$) compared to the vibrations sensed at the finger (acf: $89.3 \pm 1.9\%$), likely due to vibration dissipation through the tool and the tissue. Auditory vibrations also perform slightly less effectively (mic: $88.7 \pm 1.8\%$) than their tactile counterparts, potentially due to noise contamination and/or their lack of directionality.

Combining the five information sources (all: $90.9 \pm 1.2\%$) performs slightly worse than the torque readings alone, suggesting that additional information does not always improve the current classifier. We also observed no salient change in performance from varying the number of subsamples taken ($100 \leq n \leq 1000$, results not shown). Interestingly, data from the other two sensing tools yielded similar (4 mm: $90.9 \pm 1.7\%$) or slightly higher (8 mm: $91.2 \pm 1.1\%$) combined recognition performance (confusion matrices not shown). However, more investigation is needed to understand how tool geometry and mass affect performance.

First experiments for the out-of-domain task (comparing data collected with a tool that is different from that used to create the surface library) indicate that aligning the mean is not fruitful (results not shown). This finding suggests that the influence of the mass and tool diameter are also considerably represented in the higher-order statistical moments [5]. Successful out-of-domain generalization across sensing tools may be possible through other manipulations of the data distributions, which we will investigate in the future. We may also find that the identity of the tool cannot easily be separated from that of the surface with which it is interacting.

C. Ongoing Analysis

We are currently conducting and evaluating computational studies to assess the influences of 1) coordinate-system transformations for the multi-axial data streams, 2) noise added to all data streams, frequency- versus time-domain subsampling, 3) alternative ways of compensating for tool identity, and 4) other kernel-mean-embedding and discrepancy measures. We envision that the outcome of the results will be a set of guidelines for the design of sensor configurations for surface perception through hand-held tools and diverse robot body parts.

Acknowledgments

The authors thank Jonathan Fiene and Julian Martus for developing the accelerometer boards; Farimah Fazlollahi and Bernard Javot for helpful feedback on the test bed; Dominika Lisy for helping with data acquisition; the German Research Foundation (DFG) Project EXC 2050/1, 390696704, CeTI Cluster of Excellence at TU Dresden for supporting Y. Shao; and the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting B. Khojasteh.

References

[1] H. Culbertson, J. Unwin, and K. J. Kuchenbecker, "Modeling and rendering realistic textures from unconstrained tool-surface interactions," *IEEE Transactions on Haptics*, vol. 7, no. 3, pp. 381–393, 2014.

[2] M. Strese, "Haptic material acquisition, modeling, and display," Ph.D. dissertation, Technische Universität München, 2021.

[3] A. Burka, A. Rajvanshi, S. Allen, and K. J. Kuchenbecker, "Proton 2: Increasing the sensitivity and portability of a visuo-haptic surface interaction recorder," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 439–445.

[4] Y. Shao, V. Hayward, and Y. Visell, "Compression of dynamic tactile information in the human hand," *Science Advances*, vol. 6, no. 16, p. eaaz1158, 2020.

[5] B. Khojasteh, F. Solowjow, S. Trimpe, and K. J. Kuchenbecker, "Multimodal multi-user surface recognition with the kernel two-sample test," *arXiv preprint arXiv:2303.04930*, 2023.

[6] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf *et al.*, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[7] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[8] W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton, "Interpretable distribution features with maximum testing power," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[9] S. Zhao, A. Sinha, Y. He, A. Perreault, J. Song, and S. Ermon, "Comparing distributions by measuring differences that affect decision making," in *Int. Conference on Learning Representations*, 2022.