Short Papers

Robust Surface Recognition With the Maximum Mean Discrepancy: Degrading Haptic-Auditory Signals Through Bandwidth and Noise

Behnam Khojasteh, *Graduate Student Member, IEEE*, Yitian Shao, *Member, IEEE*, and Katherine J. Kuchenbecker, *Fellow, IEEE*

Abstract*—Sliding a tool across a surface generates rich sensations that can be analyzed to recognize what is being touched. However, the optimal configuration for capturing these signals is yet unclear. To bridge this gap, we consider haptic-auditory data as a human explores surfaces with different steel tools, including accelerations of the tool and finger, force and torque applied to the surface, and contact sounds. Our classification pipeline uses the maximum mean discrepancy (MMD) to quantify differences in data distributions in a high-dimensional space for inference. With recordings from three hemispherical tool diameters and ten diverse surfaces, we conducted two degradation studies by decreasing sensing bandwidth and increasing added noise. We evaluate the haptic-auditory recognition performance achieved with the MMD to compare newly gathered data to each surface in our known library. The results indicate that acceleration signals alone have great potential for high-accuracy surface recognition and are robust against noise contamination. The optimal accelerometer bandwidth exceeds 1000 Hz, suggesting that useful vibrotactile information extends beyond human perception range. Finally, smaller tool tips generate contact vibrations with better noise robustness. The provided sensing guidelines may enable superhuman performance in portable surface recognition, which could benefit quality control, material documentation, and robotics.

***Index Terms*—Haptic-auditory sensing, haptic surface recognition, kernel methods, machine learning.**

I. Motivation

The biological sensing and transduction processes that occur during 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. Recognizing surfaces is a multidimensional task of sensing and interpreting the complex sensations of contact.

It would be useful if artificial systems could capture, process, and accurately recognize the rich contact signals elicited during surface exploration. Prior research introduced a diverse set of surface-sensing hardware [1], [2], [3], but it is not clear *what combination, quality, bandwidth, and acuity of sensor data are necessary* to match the efficiency and accuracy of humans. Contact vibrations, in particular, present a promising source of information because they exhibit complex patterns [4], [5], propagate widely [6], [7], offer high temporal resolution for spatial touch-information decoding [8], and make multimodal surface classification robust [8], [9]. Similarly advantageous is the fact that the MEMS-based accelerometers typically used to capture these vibrations are compact, low-cost, robust, energy-efficient sensors with simple mounting, a straightforward electrical interface, and easy calibration.

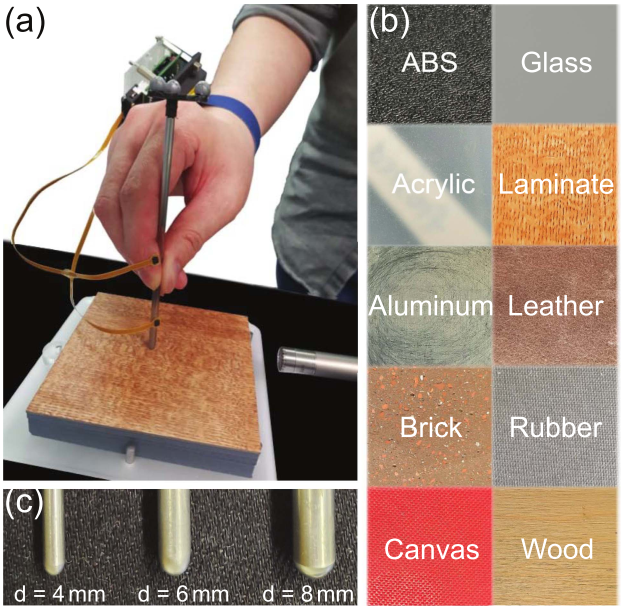

Fig. 1. (a) Recording setup. (b) Set of ten diverse surfaces. (c) Three steel tool tips with different diameters.

Many machine-learning algorithms have been proposed for surface classification in the past decade. The plethora of research includes recognizing surfaces with cues from visual-haptic-auditory [9], [10], [11], [12], [13], [14], haptic-auditory [15], [16], and only haptic [17], [18] data. Generalizing to surface-contact data recorded by a different human is known to be more challenging, and promising cross-user compensation approaches have been developed [9], [10], [11], [16], [19].

Feature-engineered classifiers or deep neural networks may suffer from overfitting due to subjective choices or limited training data [2], [20]. In contrast, mapping distributions of surface time series was

Manuscript received 26 September 2023; revised 6 December 2023; accepted 3 January 2024. Date of publication 22 January 2024; date of current version 21 March 2024. This work was supported in part by the Max Planck Society in Germany and in part by the German Research Foundation (DFG) Project EXC 2050/1, under Grant 390696704. This paper was recommended for publication by Associate Editor M. Ziat and Editor-in-Chief D. Prattichizzo upon evaluation of the reviewers' comments. *(Corresponding authors: Behnam Khojasteh; Katherine J. Kuchenbecker.)*

Behnam Khojasteh and Katherine J. Kuchenbecker are with the Max Planck Institute for Intelligent Systems (MPI-IS), 70569 Stuttgart, Germany, and also with the Faculty of Engineering Design, Production Engineering and Automotive Engineering, University of Stuttgart, 70174 Stuttgart, Germany (e-mail: khojasteh@is.mpg.de; kjk@is.mpg.de).

Yitian Shao is with the Centre for Tactile Internet with Human-in-the-Loop (CeTI), Technische Universität Dresden, 01069 Dresden, Germany, and also with the Max Planck Institute for Intelligent Systems (MPI-IS), 70569 Stuttgart, Germany (e-mail: ytshao@is.mpg.de).

Digital Object Identifier 10.1109/TOH.2024.3356609

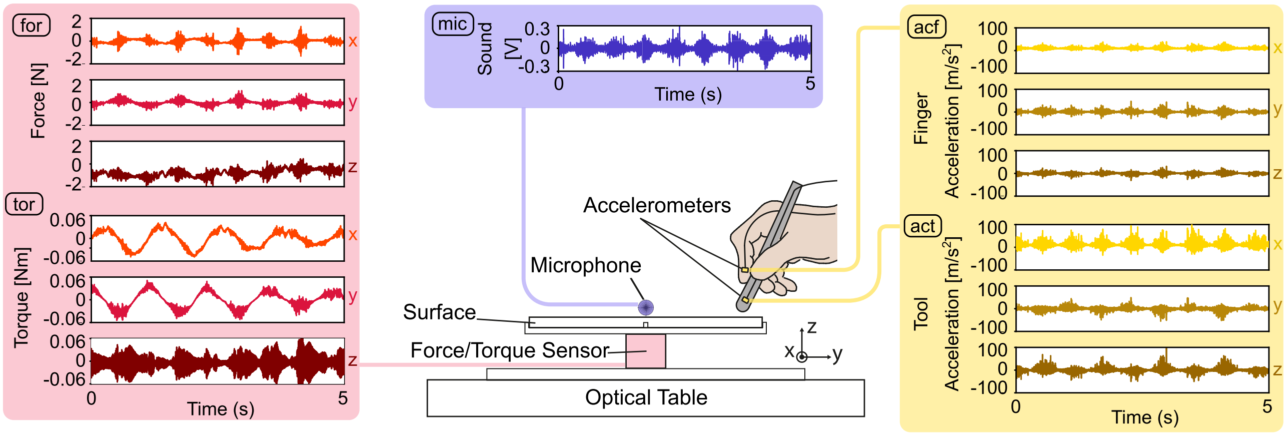

Fig. 2. Design of the test bed and sample recording from each sensor for an experimenter dragging a steel tool ($d = 6$ mm) on the laminate surface.

recently proven to be highly effective for surface recognition [9]. This approach builds on the framework of kernel two-sample testing [21] for data that is not independent and identically distributed (i.i.d.) [22] to quantify differences in distributions of surface data using the maximum mean discrepancy (MMD) – a metric that effectively quantifies the difference between two distributions in a high-dimensional space by considering all their statistical moments. This framework unlocked an automated sample-efficient technique to classify multimodal surface data (e.g., images, haptic signals, and sounds) without the constraints of feature engineering or large training datasets.

The choice of the sensing tool greatly affects the mechanical signals that are produced during surface interactions [23]. However, our understanding of the influence of the sensing tool on the mechanical basis of surface encoding is incomplete. A study by Kirsch et al. showed that vibrotactile signals that are elicited from various hard steel tool tips with surfaces exhibit complex distributions and time-series characteristics [24]. Strese et al. reported better multimodal surface-classification capabilities of their scanning system equipped with a steel tool tip in comparison to a human finger [2]. They pointed out that this trend may originate from the hard-hard contacts that amplify mechanical signals, therefore facilitating surface classification. However, the effects of the sensing tool's dimensions have not been systematically explored.

To guide the choice of sensing hardware and sampling rates, we systematically investigated which *signal bandwidths* are most effective for artificial surface classification with haptic-auditory recordings generated from three steel tools. In addition, due to the ubiquity of mechanical and electrical noise, we evaluated the pipeline's robustness by adding white *noise* to the captured signals. These investigations are all performed on newly captured high-quality surface data using the automated MMD-based recognition framework of Khojasteh et al. [9]. They demonstrated a sample-efficient approach for learning to recognize 108 surface textures from multimodal (visual, auditory, and haptic) sensor readings obtained from a public data set recorded by eleven different people. Their algorithm achieved higher recognition rates than traditional machine-learning models based on expert knowledge, also requiring less training data and optimization compared to standard deep-learning methods. To support research progress in this domain, we also share our haptic-auditory recordings [25].

II. Haptic-Auditory Surface Recordings

To provide insights into both the hardware and software sides of artificial texture perception, we conducted a degradation study involving bandwidth and noise. Haptic-auditory contact data were recorded from human-guided surface exploration (Fig. 1(a)) with an instrumented tool, as such data have previously been shown to reflect the friction, hardness, and texture of the surface being touched [26]. We prefer human-operated rather than automatic haptic data collection because the hardware costs less and can be made portable for field use [3]. We carefully selected a set of $C = 10$ surface textures (Fig. 1(b)) from the Penn Haptic Texture Toolkit [27] drawn from material categories that are also representative of other surface datasets [2], [3]; similar pairs of materials (e.g., wood and laminate) were purposefully included to make recognition more challenging.

We considered three solid steel tools of the same length with thermally hardened hemispherical tool tips of 4, 6, and 8 mm diameter (Fig. 1(c)); their masses are 11.6, 26.1, and 46.4 g, respectively. To capture relevant haptic and auditory data from surface texture interactions, the test bed (Fig. 2) comprises two miniature digital high-bandwidth low-noise accelerometers (STMicroelectronics, IIS3DWB) securely mounted to the tool itself and the experimenter's finger via double-sided tape (tesa SE), a six-axis force/torque sensor (ATI Industrial Automation Inc., Nano43) underneath the surface sample, and a high-fidelity microphone (Brüel and Kjaer, Type 4955) above the rigid surface platform. A motion-capture system (Vicon, Vantage 5) tracked the tool position and orientation, but these data were not analyzed for the reported studies.

A. Measurement Protocol

Khojasteh et al.'s data-driven method effectively mitigated speed-, force-, and session-dependent effects during tool-surface interaction in order to generalize from one user to the other ten users [9]. In this context, a simple distribution shift of time-series data was sufficient to boost the recognition accuracy by 9%, up to the near-perfect score of 97.2%. As we adopt their effective multimodal multi-user surface-recognition framework, we focus on data recordings by one human. During data acquisition, an experimenter recorded rich surface data between the selected tool and surface by varying their tool speed and applied normal force, which ranged from 10 to 440 mm/s and 0 to 5.1 N, respectively. The experimenter was asked to choose a free circular motion to capture rich signals from different contact conditions and phenomena, as in [1], [27]. From two long data recordings for each surface c, we obtained ten trials that are each five seconds long without any transient artifacts. Thus, in total we have 300 trials of multimodal surface data ($C = 10$ surfaces × 3 steel tools × 10 trials). During data

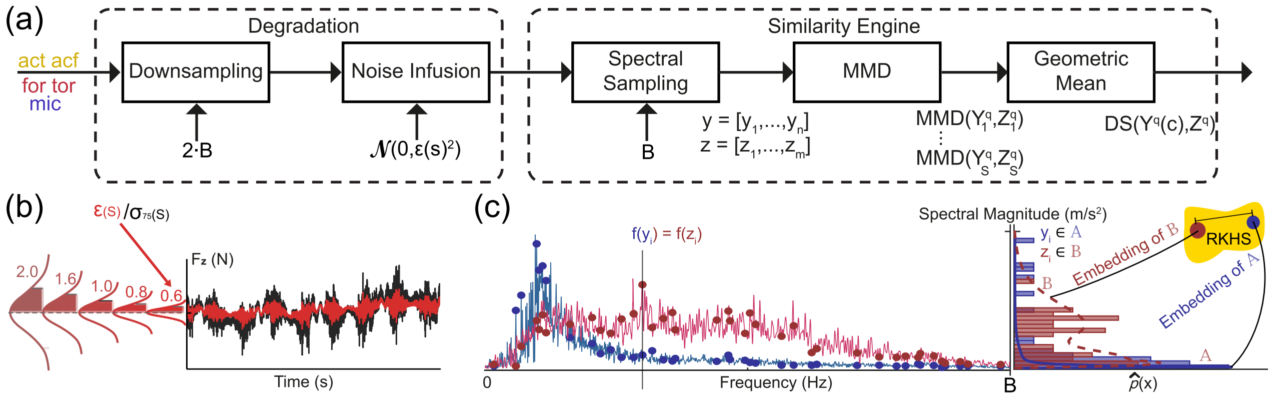

Fig. 3. (a) Pipeline with degradation studies and classification framework. (b) Illustration of white noise addition on a time-domain (force) signal. (c) Adopted spectral sampling strategy of time-series (acceleration) data from two surfaces [9]. The two distributions are mapped into a reproducing kernel Hilbert space (RKHS) in order to estimate the MMD through MMD_b^2.

recording, the sampling rate was 26 667 Hz for the two accelerometers and 44 100 Hz for the force/torque sensor and the microphone.

B. Information Sources

Each surface trial consists of five information sources from the four sensors during tool dragging on the surface. The time-series information sources are: three-axis acceleration of the tool (`act`) and fingertip (`acf`), three-axis contact forces (`for`) and the corresponding three-axis torques (`tor`), and one-dimensional contact sounds (`mic`); Fig. 2 shows sample data.

III. Degradation and Classification

We investigated the role of (1) frequency bandwidth B (i.e., Nyquist frequency) and (2) additive noise on tool-mediated surface recognition by degrading the haptic-auditory signals through these parameters. This paper analyzes 110 conditions of 10 varying signal bandwidths and 11 wide-ranging noise levels. Our approach leverages ideas from recent work on automatic classification of visual-haptic-auditory data of 108 surfaces [9]. The degradation studies and the recognition framework are elaborated in the following.

A. Problem Formulation

Our goal is to classify unseen multimodal sensor recordings from physical surface interactions with the same sensing tool. From a mathematical perspective, we address this surface-recognition task by focusing exclusively on data distribution differences. We model surface interactions as realizations of a dynamical system [22]. We assume that a set of $C \in \mathbb{N}$ unique surfaces will induce different distributions $\mathbb{P}_1, \ldots, \mathbb{P}_C$, respectively. The classifier compares unlabeled surface trials to the known surface distributions in the library to determine the surface class c from which it most likely came. To infer whether an unseen testing trial Z and a training trial $Y(c)$ from a library come from identical or non-identical surfaces, we quantify distribution differences between the two trials.

B. Downsampling

To identify the relevant bandwidth B for surface classification, we systematically downsample all five information sources after applying a low-pass filter and mirror padding to prevent aliasing and edge artifacts, respectively. In all conditions, the sampling rate is twice $(2 \cdot B)$ the bandwidth (Nyquist frequency) of focus (Fig. 3(a)). The sampling rates in all configurations of the degradation studies are lower than the actual data acquisition rate for all sensor readings. These reduced sampling rates entail lower spectral resolution, so this degradation simulates digital sensors with lower sampling rates, as are often used in surface classification. The bandwidth of interest always starts at DC (0 Hz), and we vary the upper end of the considered frequency range. For the upper frequency end, we consider ten logarithmically spaced values between the sensor's bandwidth (given by the manufacturer, reduced for the force/torque sensor due to the attached surface) and a value that is 1000 times smaller. The maximum spectral bandwidth for the analysis is 0–6300 Hz for the accelerometer readings (`act` and `acf`), 0–2100 Hz for the force/torque sensor (`for` and `tor`), and 0–20 000 Hz for the microphone data (`mic`), which includes the full human-audible perception band.

C. Noise Infusion

We infuse broad-bandwidth white noise from a zero-mean normal distribution $\mathcal{N}(0, \varepsilon(s)^2)$ into the sensor readings (Fig. 3(b)). Such a signal spread indicates the noise intensity through the variance, $\varepsilon^2 \in \mathbb{R}^{D(s)}$, where $D(s)$ represents the number of directions or axes in the corresponding information source. To ensure uniformity across information sources and suitability of dataset-specific noise, we incorporate noise that is proportional to the standard deviation of the original recorded signals. The standard deviation provides a balanced representation of the signal spread for time series with and without a DC component. The infused noise magnitude $\epsilon(s)$ of information source s is statistically weighted with

$$\sigma_{75}(s) = \mathrm{median}_{75\mathrm{th}}\{\sigma_{\max,1}(s), \ldots, \sigma_{\max,Q}(s)\} \tag{1}$$

that first identifies the maximum standard deviation values over time for each sensor axis from each of the 300 tool-surface trials, and then determines the top 75th percentile of these maximum values. Selecting the median from the top quartile's maximum standard deviations serves as a reliable representation of the variability within the higher range of all tool-surface data. This approach is resistant to the impact of outliers, so that a variety of noise magnitudes can be modelled without distortion. For each information source s, the noise magnitude is

$$\varepsilon(s) = w_\epsilon \cdot \sigma_{75}(s) \tag{2}$$

with weights, $w_\epsilon \in \{0, \frac{1}{5}, \ldots, 2\}$, varying from no noise through ten linearly spaced noise levels with an overall maximum noise magnitude

TABLE I
MAXIMUM NOISE MAGNITUDES FOR EACH INFORMATION SOURCE IN THREE DIRECTIONS

Information	$2 \cdot \sigma_{75}$		
Source s	x	y	z
act (m/s²)	15.52	12.36	10.18
acf (m/s²)	6.89	7.17	4.28
for (N)	0.84	0.88	0.72
tor (Nm)	0.11	0.11	0.02
mic (V)		0.07	

of $2 \cdot \sigma_{75}(s)$. Table I lists these maximum noise magnitudes with units for the five information sources in all directions. Our noise infusion approach serves as a measure for signal-to-noise ratio and therefore may represent a variety of real-world scenarios, such as a sensor's inherent noise (digital vs. analog), environmental mechanical and electrical noise, or other independent noise sources.

D. Spectral Sampling

To quantify surface similarity between a testing trial and a library (training) trial, we consider the spectral magnitude distribution of all Fourier-transformed information sources (Fig. 3(c)). We use the frequency-domain representation due to the distinct high-frequency nature of all sensor readings (Fig. 2). For consistency across information sources, we always compute $N = 6000$ spectral bins for the discrete-time Fourier transform, so that the selected bandwidth B determines the spectral resolution. With the given resolution, we randomly extract $n = m = 300$ unique spectral magnitudes at consistent frequencies from both testing and training trials to quantify surface similarity (Fig. 3(c)). For this approach, we have observed no significant improvement in classification performance when more spectral magnitudes were considered; thus, we maintained 300 data points throughout the study to avoid unnecessary increases in computation time.

E. Similarity Metric

The two sets of extracted spectral magnitudes, y and z, are fed into the core component of our surface similarity engine, i.e., the MMD (Fig. 3(a) and (c)). This metric gauges the distance between two probability distributions by considering their statistical moments in a reproducing kernel Hilbert space (RKHS), which is a space equipped with inner products. Given its efficacy in multimodal multi-user surface recognition [9], we use the squared bias MMD estimator by Gretton et al. [21],

$$\mathrm{MMD}_b^2[\mathbb{P}_Y, \mathbb{P}_Z] = \frac{1}{n^2} \sum_{i,j=1}^{n} k(y_i, y_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(z_i, z_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(y_i, z_j), \tag{3}$$

where $[y_1, \ldots, y_n]$ and $[z_1, \ldots, z_m]$ are i.i.d. random variables. In our case, these are n and m samples from surface data streams Y_s and Z_s with unknown distributions $\mathbb{P}_{Y_s}$ and $\mathbb{P}_{Z_s}$. For our kernel function $k(\cdot, \cdot)$, we use the squared exponential function along with its well-established hyperparameter heuristics for all statistical tests due to its suitability for visual-haptic-auditory surface data [9].

F. Global Similarity Decision

For multi-source classification, we use the geometric mean to unify the MMD scores of multiple information sources to an overall discrepancy score DS (Fig. 3(a)). This framework uses the arithmetic mean of individual logarithm-transformed MMD values and therefore improves MMD scale-invariance across information sources. Our full classifier combines all $S = 5$ information sources, which we term `all`. Based on the MMD estimator in (3), we compute the global discrepancy score

$$\mathrm{DS}[Y, Z] = \sqrt[S]{\prod\nolimits_{s=1}^{S} \left(\mathrm{MMD}_b^2[\mathbb{P}_{Y_s}, \mathbb{P}_{Z_s}]\right)^{w_s}} \tag{4}$$

between two trials, Y and Z; greater values for DS imply that the two trials have a higher discrepancy. While the exponential weights $w_s \in \mathbb{R}^+$ allow one to consider some information sources more strongly than others, we choose unit MMD weights, $w_s = w_{\mathrm{MMD}} = 1$, in all our experiments. To avoid issues with geometric means, we confirmed that all individual MMD scores are positive, i.e., $\mathrm{MMD} \in \mathbb{R}^+$.

Our algorithm leverages the k-nearest neighbors principle to make classification predictions with the global DS scores. An unlabeled surface trial Z will be classified to the class c in the library with C surfaces according to $\min_{c \in C} \mathrm{DS}[Y(c), Z]$; it predicts a surface class through the test-train trial pair with the smallest global discrepancy distance, i.e., the nearest neighbor. The size of the library plays a significant role in determining the classification complexity. When the library contains a limited number of training instances, effectively generalizing to new, unseen data may be challenging. This aspect is particularly crucial when employing few-shot learning approaches, where the goal is to achieve accurate predictions with a minimal amount of training data.

IV. EXPERIMENTS

This section begins by outlining the two recognition settings of our experiments with different training sets: the (1) five-shot and (2) one-shot learning experiments. Subsequently, we describe the two performance metrics employed to evaluate the recognition results. These two performance metrics are the basis for computing the 30 optimal bandwidths (two performance metrics × $S = 5$ information sources × three tools).

A. Few-Shot Learning Experiments

In our experiments, we independently test each of our 100 surface trials, grouped by tool-specific datasets. For testing, an unlabeled trial is compared to a class-balanced random subset (training) of other trials from the library for that tool diameter. The two degradation studies are our main experiments; they consider five library trials per class for prediction, i.e., five-shot learning. After determining the optimal sensor bandwidths, we perform one-shot learning experiments, in which we consider only one five-second library trial per surface class. This more difficult scenario represents a setting that is constrained by limited training data. For this experiment, we also perform dual-source classification to investigate which paired combinations of tactile, kinesthetic, and auditory cues are complementary for multimodal surface recognition.

B. Classification Performance Metrics

To reduce the influence of different data distributions on recognition performance, we run our classification pipeline in $R = 5$ repeated iterations for each surface trial of the tool-specific testing sets. In every iteration, we consider spectral magnitudes from a distinct set of frequencies for the MMD tests. Our primary performance metric is the classification accuracy (ACC) of the tool-specific testing set. From

TABLE II
OPTIMAL BANDWIDTHS IN HZ FOR
FIVE INFORMATION SOURCES AND THREE SENSING TOOLS

Information Sources	$\mathrm{B}_{\mathrm{opt}}$(Accuracy) (Hz) 4 mm	6 mm	8 mm	$\mathrm{B}_{\mathrm{opt}}$(Margin) (Hz) 4 mm	6 mm	8 mm
act	1801	2048	1768	744	388	392
acf	2162	1681	875	949	352	321
for	1188	1000	896	496	429	338
tor	884	666	464	948	195	195
mic	1784	2599	1957	1729	2510	1923

the R iterations, we calculate the overall classification accuracy for both individual and combined information sources. We thus obtain both the mean accuracy and its standard deviation.

The MMD surface similarity metric, inherent in our recognition framework, affords a novel performance metric, the average safety margin, which serves as a measure for confidence and reliability in the classification decision. Unlike the classification accuracy that is bound to the range of 0–100%, the average safety margin can potentially provide deeper insights, allowing configurations to be more precisely assessed and altered to increase their robustness in unseen scenarios. We define the safety margin

$$\mathrm{SM}_s = \mathrm{MMD}_b^2[Y_s'', Z_s] - \mathrm{MMD}_b^2[Y_s', Z_s] \quad (5)$$

as the MMD difference in the s-th information source between a testing trial Z_s and its closest false Y_s'' and true Y_s' library trials. Then, for each tool-specific testing set with the $R = 5$ repetitions, we compute the overall average safety margin.

C. Optimal Bandwidth Definitions

The initial results of our degradation studies highlighted that the optimal region for the bandwidth is below 4000 Hz for all haptic and auditory information sources. This observation aligns with expectations from tool-surface interactions. To precisely identify this optimal region for the bandwidth, we performed additional bandwidth computations without added noise to have a frequency spacing of $\Delta\mathrm{B}_{i,i-1} = 25$ Hz spanning the range of 0–4000 Hz. This finer resolution led to a total of $H = 160$ operating points. When determining the optimal bandwidth, we exclusively incorporated those bandwidths whose performance metrics (accuracy or average safety margin) are within a value of five (percent or unitless, respectively) of the maximum value for that specific information source and sensing tool; in the case of classification accuracy, this would be: $\mathrm{ACC} \geq \mathrm{ACC}_{\mathrm{max}}$–5%. From this selected set of high-accuracy bandwidths, we then compute the optimal bandwidth

$$\mathrm{B}_{\mathrm{opt}}(\mathrm{Accuracy}) = \frac{\sum_{i=1}^{H} \mathrm{B_i} \cdot \mathrm{ACC_i}}{\sum_{i=1}^{H} \mathrm{ACC_i}} \quad (6)$$

where the ith bandwidth B_i is weighted with the corresponding accuracy value. The optimal bandwidth based on the average safety margin, $\mathrm{B}_{\mathrm{opt}}(\mathrm{Margin})$, is computed analogously. This approach can mitigate the effect of outliers and therefore ensures a congruent representation of the optimal bandwidth for each information source and sensing tool (Table II). As mentioned above, these calculations are done for conditions without added noise ($w = 0$) for all five information sources, after confirming that the maximum performance metric value was in these noise-free configurations and not in the ones from the first noise level.

V. RESULTS AND DISCUSSION

A. Optimal Bandwidth of Haptic-Auditory Signals

As seen in the lower noise-free performance-versus-bandwidth plots in Fig. 4, for the majority of conditions we found a plateau for the accuracy curve and a peak for the margin curve. The bounded and unbounded nature of the two evaluation metrics trigger this behavior: every parameter variation (e.g., bandwidth, noise) perturbs the MMD metric, and therefore we usually get a peak value for the safety margin, whereas larger parameter variations are required for the correct and wrong surface classifications to flip. In terms of accuracy-based optimal bandwidths, we observe the best-performing bandwidth region to end between 200–2500 Hz for the tool accelerations (`act`), 200–3000 Hz for the finger accelerations (`acf`), 200–1600 Hz for the contact forces (`for`), 200–1000 Hz for the torques (`tor`), and 600–4000 Hz for contact sounds `mic`. The optimal accuracy-based bandwidths of all information sources and tools lie in these regions (Table II), validating the approach we chose to compute these values.

Compared to the accuracy-inferred bandwidths, the optimal bandwidths from the margins are smaller in all cases. This consistent trend happens because the margin peaks are closer in frequency to the beginning of the high-accuracy plateaus, suggesting that the overall best configuration might tend to occur with smaller bandwidths. While this finding makes sense, as a smaller sensor bandwidth improves the noise characteristics, more research is needed to verify how optimization choices can best be inferred from the safety margins.

In the majority of configurations, tools with larger diameter have lower optimal bandwidth. This pattern is valid for both the accuracy- and margin-based optimal bandwidths, implying that larger tools generate more surface-relevant low-frequency signals. A smaller tool tip will penetrate more between asperities on the surface than a larger tool [23], thereby giving the haptic-auditory signals higher-frequency content.

Three haptic information sources (`act, acf, for`) achieve 100% perfect recognition rates for several bandwidths, while torques and contact sounds reach their best accuracies (98% and 88%) at 625 Hz and 2075 Hz, respectively. The tool and finger accelerations both enable reliable high-accuracy classification in broad-bandwidth configurations; the flat high-bandwidth frequency response of the accelerometers is also highlighted by the manufacturer. Furthermore, mounting the accelerometer on the rigid tool yields better classification on average than mounting it on the experimenter's soft skin. We believe this trend occurs because the mechanical waves from the hard-hard surface contacts reduce in amplitude and frequency content as they travel through the tool and tissue.

Compared to the accelerations, the top-performing forces have a smaller upper frequency end for the bandwidth. In contrast to the tactile and auditory vibrations, the forces achieve decent recognition accuracy (above 60%) for bandwidths below 10 Hz due to their expressive DC component. The 3D forces contain important information about the frictional surface contact and how the user adjusts their grip force. Compared to the forces, the torques perform less well in classification. While the definition of the torques inherently includes the 3D forces, the lever arm to the contact point is not related the properties of the surface; thus, a reliable surface classifier should consider forces instead of torques. The contact sounds are the worst-performing information source, potentially due to their lack of directionality and their sensitivity to ambient sounds. The same trend was observed in haptic-auditory surface classification of 108 surfaces [9], thereby suggesting that a robust surface classifier should focus on tactile rather than auditory vibrations.

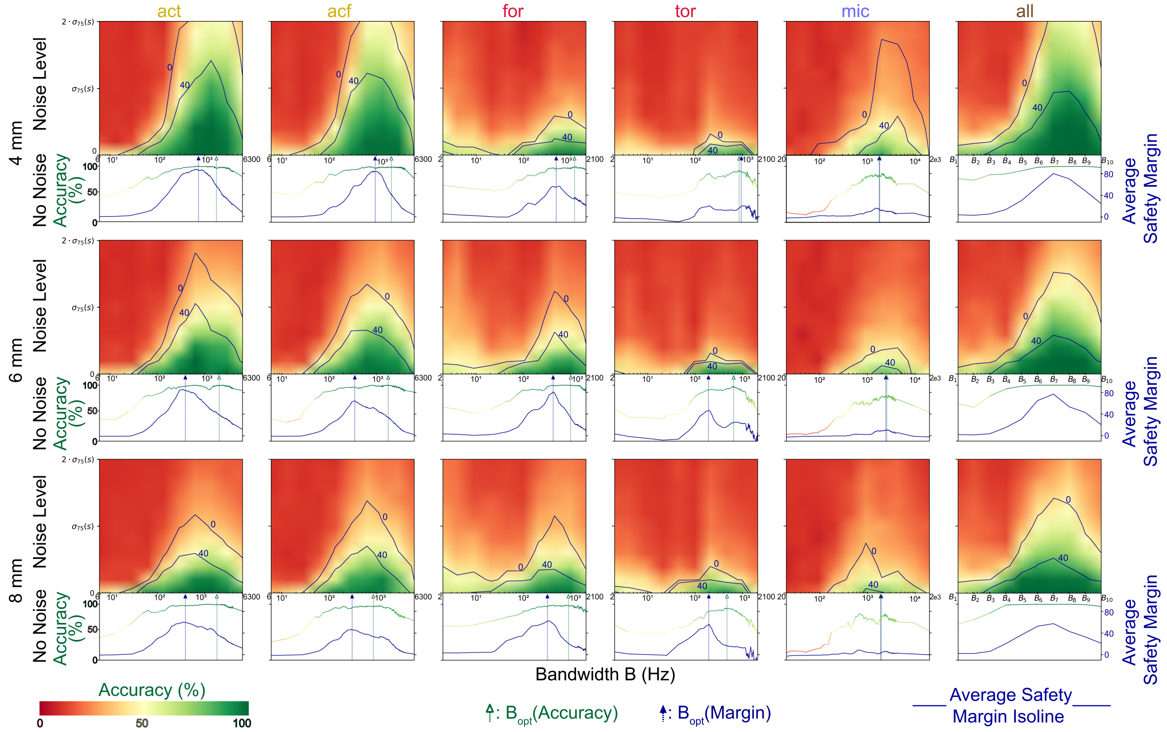

Fig. 4. Main results of the degradation studies for the five individual and all combined information sources and all three sensing tools: the red-yellow-green heatmaps show how classification accuracy varies with bandwidth and noise, while the dark blue curves show average safety margin isolines at 0 and 40. Below each heatmap, more finely calculated results without noise contamination detail how classification accuracy and average safety margin depend on the bandwidth B. Triangles depict the optimal bandwidths for accuracy (green arrow) and average safety margin (blue arrow) from Table II.

B. Noise Robustness of Haptic-Auditory Signals

With regard to the noise resistance between sensing tools (Fig. 4), both performance metrics show that the 4 mm tool has the best noise robustness for the contact vibrations and the 6 mm tool the best robustness for the forces and torques. This behavior originates from the signal energies of the corresponding information sources that determined the maximum noise levels (Table I). The maximum standard deviations σ_{75} for the accelerations and sounds were chosen from the trials generated with the 4 mm tool, and for the forces and torques they were selected from the 6 mm tool's recordings. We believe this trend of higher vibration or force amplitudes is mostly caused by the differing masses and flexural rigidities of the tools. In addition, the aforementioned trend of smaller tool tip diameter engaging with more surface asperities also contributes to the higher vibration magnitudes of the 4 mm tool.

The bandwidth-versus-noise heatmaps (Fig. 4) show that the tool and finger accelerations have the highest noise robustness, followed by the auditory vibrations. The chosen digital accelerometers have very low noise density of 75 $\mu\text{g}/\sqrt{\text{Hz}}$. Starting with almost perfect recognition accuracy ($>99\%$) for the 4 mm tool's surface recognition without noise, adding white noise with magnitudes of 6.2 m/s^2 (σ_{75}) and 12.4 m/s^2 ($2 \cdot \sigma_{75}$) reduces accuracy to 85% and 50%, respectively. The finger accelerations exhibit a similar robust behavior. In the case of the microphone signals for the same setting, the noise-free condition (87%) significantly deteriorates to 30% and 22%. Forces and torques are also more susceptible to spurious noise, potentially because both of these information sources rely on their DC components for the surface-recognition task. Comparing the noise conditions analogously, the 8 mm recognition accuracy for the forces drops from 98% to 32% and 22%. For the torques and the 8 mm tool, the accuracy drops from 92% to 19% and 15%. To conclude, high-frequency tactile vibrations are more robust to noise than the other information sources, thereby highlighting their suitability for classifying surfaces in noisy settings.

C. Combined Information Sources

Combining all five information sources enables the full classifier to achieve perfect recognition rates for a range of bandwidths, even at higher noise levels. This robust classification is enabled by considering the advantages of the different information sources (e.g., expressive DC and AC components) The MMD is very effective at detecting salient distribution differences in data, thereby making our classifier very robust.

In dual-source one-shot classification (Fig. 5), combining tactile (`act`, `acf`) and kinesthetic (`for`) cues boosts recognition performance, potentially due to complementary information in the high- and low-frequency ranges. In particular, we report superior surface recognition when combining three-axis tool accelerations and contact forces in this sparse data setting. This observation in artificial surface perception closely resembles the mechanisms of human touch, where slow- and fast-adapting mechanoreceptors provide complementary cues for steady pressure and rapidly varying stimuli, respectively. High-frequency contact sounds (`mic`) also perform best with contact forces, gaining more new surface-relevant information in this setting than when combined with tactile vibrations. Adding another accelerometer at a farther location does not help in surface recognition, suggesting that an accelerometer close to the surface contact is sufficient. Improvements in combined force-torque classification may arise from the better noise suppression of the strain-gauge-based F/T sensor. The torques may have stronger frictional (planar) signals due to the lever arm, compared to the forces. Choosing the bandwidths from the safety margins (Fig. 5(b)) is less successful than the accuracy-based approach (Fig. 5(a)) in this

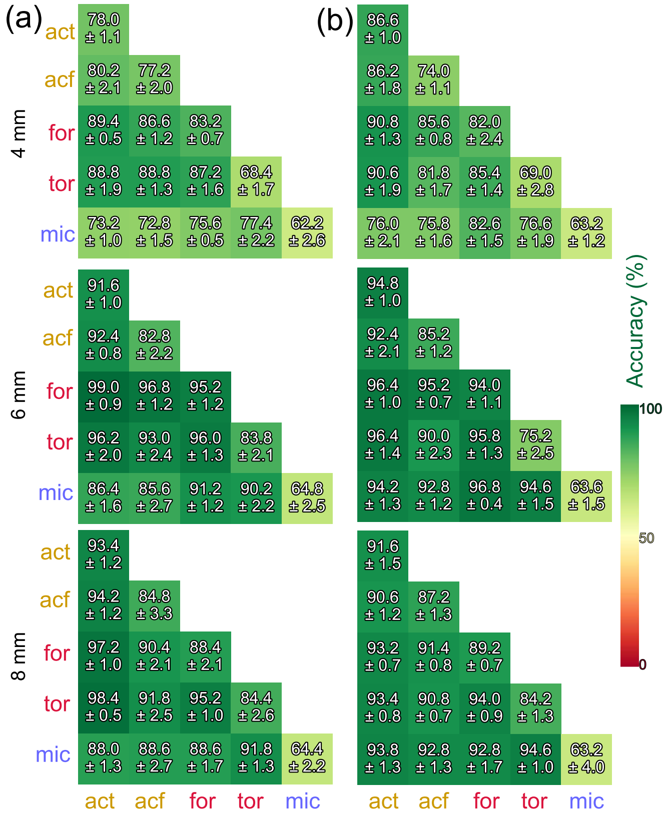

Fig. 5. Individual and dual-source classification results from one-shot learning with optimal bandwidths from (a) accuracy and (b) average safety margin. Each box shows the mean and standard deviation across trials.

one-shot-learning approach with the same data. Therefore, primarily considering classification accuracy for unseen configurations seems more promising than safety margin, but this finding needs to be investigated more in future work.

In dual-source classification, the 4 mm results show that recognition is more difficult with sparse training data. The variability between the 4 mm trials may be higher, potentially confusing our classifier when only one trial per surface is considered. At the same time, smaller-diameter sensing tools have better noise robustness in terms of accuracy and margin.

VI. Conclusion

Artificial surface perception is highly relevant in manufacturing, quality control, and robotics. To identify the optimal sensing configuration, we captured high-quality haptic-auditory data of ten surfaces with three tools for open-source use. With these recordings, we conducted two degradation studies involving signal bandwidth and noise for robust surface recognition with our automated MMD-based classification pipeline. We identified the optimal bandwidths and noise resistance for the tactile, auditory, and force-torque surface signals. In particular, high-frequency tactile contact vibrations enable the highest-accuracy robust surface classification even with noisier accelerometers. Unlike the common choice of 1000 Hz, we found that the optimal bandwidth is higher for vibrotactile information in artificial surface recognition. High-frequency transient contact forces with an expressive DC component are also successful for surface recognition, but several commercial haptic sensors have only low-bandwidth force sensing. In contrast to higher default choices, our findings further suggest that tool-surface sounds up to only 4000 Hz are helpful in classification. Sensing tools with smaller tip diameters amplify contact vibrations, thereby enabling better and more noise-robust recognition rates. These results provide a set of guidelines for the design of sensor configurations for surface perception through rigid hand-held tools.

For the future, non-Gaussian noise models could be explored to try to represent additional complex contact conditions and external factors such as surface contaminants and sensor malfunctions. Investigating the effect of sensor resolution and sensor sensitivity on surface-recognition performance could also be an area of focus. Other tools and end-effectors should also be explored to identify the optimal design. For applications involving compliant end-effectors, it would be important to contextualize the performance and optimal bandwidth of a soft sensing tool alongside our results for hard tools. Conducting a comparative, systematic analysis of soft and hard tools with regard to geometrical, material and mechanical parameters could provide insightful information about both types of contacts. Validating our results for acceleration and force sensors with different frequency responses and noise characteristics could also offer additional insights. Finally, our approach could be compared to and potentially combined with traditional metrology methods like profilometry for characterizing surface topography.

Data Availability Statement

The haptic-auditory dataset [25] of this article is available on the open research data repository of the Max Planck Society, Edmond: https://doi.org/10.17617/3.PM8R94

Acknowledgment

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 Trans. Haptics*, vol. 7, no. 3, pp. 381–393, Jul.-Sep. 2014.
[2] M. Strese, "Haptic material acquisition, modeling, and display," Ph.D. dissertation, Technische Universität München, Munich, Germany, 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 *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 439–445.
[4] M. Janko, R. Primerano, and Y. Visell, "On frictional forces between the finger and a textured surface during active touch," *IEEE Trans. Haptics*, vol. 9, no. 2, pp. 221–232, Apr.-Jun., 2016.
[5] B. Khojasteh, M. Janko, and Y. Visell, "Complexity, rate, and scale in sliding friction dynamics between a finger and textured surface," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 13710.
[6] Y. Shao, V. Hayward, and Y. Visell, "Spatial patterns of cutaneous vibration during whole-hand haptic interactions," *Proc. Nat. Acad. Sci.*, vol. 113, no. 15, pp. 4188–4193, 2016.
[7] Y. Shao, H. Hu, and Y. Visell, "A wearable tactile sensor array for large area remote vibration sensing in the hand," *IEEE Sensors J.*, vol. 20, no. 12, pp. 6612–6623, Jun. 2020.
[8] Y. Shao, V. Hayward, and Y. Visell, "Compression of dynamic tactile information in the human hand," *Sci. Adv.*, vol. 6, no. 16, 2020, Art. no. eaaz1158.

[9] B. Khojasteh, F. Solowjow, S. Trimpe, and K. J. Kuchenbecker, "Multimodal multi-user surface recognition with the Kernel two-sample test," *IEEE Trans. Automat. Sci. Eng.*, 2023,pp. 1–16,10.1109/TASE.2023.3296569.

[10] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, "Multimodal feature-based surface material classification," *IEEE Trans. Haptics*, vol. 10, no. 2, pp. 226–239, Apr.-Jun. 2017.

[11] M. Strese, Y. Boeck, and E. Steinbach, "Content-based surface material retrieval," in *Proc. IEEE World Haptics Conf.*, 2017, pp. 352–357.

[12] A. Burka and K. J. Kuchenbecker, "Handling scan-time parameters in haptic surface classification," in *Proc. IEEE World Haptics Conf.*, 2017, pp. 424–429.

[13] M. Strese, L. Brudermueller, J. Kirsch, and E. Steinbach, "Haptic material analysis and classification inspired by human exploratory procedures," *IEEE Trans. Haptics*, vol. 13, no. 2, pp. 404–424, Apr.-Jun. 2020.

[14] A. Devillard, A. Ramasamy, D. Faux, V. Hayward, and E. Burdet, "Concurrent haptic, audio, and visual data set during bare finger interaction with textured surfaces," in *Proc. IEEE World Haptics Conf.*, 2023, pp. 101–106.

[15] J. Wei, C. Shaowei, P. Hao, J. Hu, S. Wang, and Z. Lou, "Multimodal unknown surface material classification and its application to physical reasoning," *IEEE Trans. Ind. Inform.*, vol. 18, no. 7, pp. 4406–4416, Jul. 2022.

[16] Y. Liu, S. Lu, and H. Culbertson, "Texture classification by audio-tactile crossmodal congruence," in *Proc. IEEE Haptics Symp.*, 2022, pp. 1–7.

[17] J. B. Joolee, M. A. Uddin, and S. Jeon, "Deep multi-model fusion network based real object tactile understanding from haptic data," *Appl. Intell.*, vol. 52, pp. 1–16, 2022.

[18] J. A. Fishel and G. E. Loeb, "Bayesian exploration for intelligent identification of textures," *Front. Neurorobot.*, vol. 6, 2012, Art. no. 4.

[19] B. A. Richardson, Y. Vardar, C. Wallraven, and K. J. Kuchenbecker, "Learning to feel textures: Predicting perceptual similarities from unconstrained finger-surface interactions," *IEEE Trans. Haptics*, vol. 15, no. 4, pp. 705–717, Oct.-Dec. 2022.

[20] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.

[21] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A Kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.

[22] F. Solowjow, D. Baumann, C. Fiedler, A. Jocham, T. Seel, and S. Trimpe, "A kernel two-sample test for dynamical systems," 2020, *arXiv:2004.11098*.

[23] C. G. McDonald and K. J. Kuchenbecker, "Dynamic simulation of tool-mediated texture interaction," in *Proc. IEEE World Haptics Conf.*, 2013, pp. 307–312.

[24] J. Kirsch, A. Noll, M. Strese, Q. Liu, and E. Steinbach, "A low-cost acquisition, display, and evaluation setup for tactile codec development," in *Proc. IEEE Int. Symp. Haptic Audio Vis. Environ. Games*, 2018, pp. 1–6.

[25] B. Khojasteh, Y. Shao, and K. J. Kuchenbecker, "MPI-10: Haptic-auditory measurements from tool-surface interactions," 2024. [Online]. Available: https://doi.org/10.17617/3.PM8R94

[26] H. Culbertson and K. J. Kuchenbecker, "Importance of matching physical friction, hardness, and texture in creating realistic haptic virtual surfaces," *IEEE Trans. Haptics*, vol. 10, no. 1, pp. 63–74, Jan.-Mar. 2017.

[27] H. Culbertson, J. J. L. Delgado, and K. J. Kuchenbecker, "One hundred data-driven haptic texture models and open-source methods for rendering on 3D objects," in *Proc. IEEE Haptics Symp.*, 2014, pp. 319–325.