Learning to Feel Textures: Predicting Perceptual Similarities From Unconstrained Finger-Surface Interactions

Benjamin A. Richardson, Yasemin Vardar, *Member, IEEE*, Christian Wallraven, *Member, IEEE*, and Katherine J. Kuchenbecker, *Fellow, IEEE*

***Abstract*—Whenever we touch a surface with our fingers, we perceive distinct tactile properties that are based on the underlying dynamics of the interaction. However, little is known about how the brain aggregates the sensory information from these dynamics to form abstract representations of textures. Earlier studies in surface perception all used general surface descriptors measured in controlled conditions instead of considering the unique dynamics of specific interactions, reducing the comprehensiveness and interpretability of the results. Here, we present an interpretable modeling method that predicts the perceptual similarity of surfaces by comparing probability distributions of features calculated from short time windows of specific physical signals (finger motion, contact force, fingernail acceleration) elicited during unconstrained finger-surface interactions. The results show that our method can predict the similarity judgments of individual participants with a maximum Spearman's correlation of 0.7. Furthermore, we found evidence that different participants weight interaction features differently when judging surface similarity. Our findings provide new perspectives on human texture perception during active touch, and our approach could benefit haptic surface assessment, robotic tactile perception, and haptic rendering.**

***Index Terms*—Texture perception, machine learning, finger-surface interaction, predicting human tactile perception, probabilistic representation.**

Manuscript received 5 February 2022; revised 10 August 2022; accepted 6 September 2022. Date of publication 10 October 2022; date of current version 19 December 2022. The work of Christian Wallraven was supported by the Institute for Information and Communications Technology Promotion (IITP), in part by Korea Government, under Grants 2019-0-00079 and 2017-0-00451, and in part by the National Research Foundation of Korea under Grant NRF-2017M3C7A1041824. This work was supported by the German Ministry of Education and Research (BMBF) through the Tübingen AI Center under Grant FKZ 01IS18039B. This article was recommended for publication by Associate Editor Prof. Astrid M.L. Kappers and Editor-in-Chief Dr. Seungmoon Choi upon evaluation of the reviewers' comments. *(Benjamin A. Richardson and Yasemin Vardar contributed equally to this work.) (Corresponding author: Benjamin A. Richardson.)*

Benjamin A. Richardson is with the Haptic Intelligence Department at the Max Planck Institute for Intelligent Systems and with the Computer Science Department, University of Stuttgart, 70569 Stuttgart, Germany (e-mail: richardson@is.mpg.de).

Yasemin Vardar is with the Department of Cognitive Robotics, Delft University of Technology, 2628 Delft, The Netherlands (e-mail: y.vardar@tudelft.nl).

Christian Wallraven is with the Departments of Artificial Intelligence and Brain and Cognitive Engineering, Korea University, 02841 Seoul, South Korea (e-mail: wallraven@korea.ac.kr).

Katherine J. Kuchenbecker is with the Haptic Intelligence Department, Max Planck Institute for Intelligent Systems, 70569 Stuttgart, Germany (e-mail: kjk@is.mpg.de).

This article has supplementary downloadable material available at https://doi.org/10.1109/TOH.2022.3212701, provided by the authors.

Digital Object Identifier 10.1109/TOH.2022.3212701

I. Introduction

When humans touch a surface with their fingers, they feel a rich array of tactile cues revealing the physical properties of the surface, such as friction, roughness, and deformability. The spatio-temporal fingertip deformations activate several types of mechanoreceptors, which send signals to connected tactile afferents, transmitting information to the central nervous system [1]. The skin deformations that occur depend on the material properties and geometry of the finger and the surface [2], [3], normal force [4], and speed [5], and they can vary substantially even for the same person exploring the same texture [6]. Little is known about how the brain distills the information needed to evaluate textures from the combination of skin deformation and exploratory motion.

A common approach to determining the fundamental factors underpinning texture perception is conducting psychophysical experiments in which participants rank the similarity of surfaces or give ratings for their specific features (e.g., hardness, roughness). The results are typically analyzed by a dimensional reduction technique such as multidimensional scaling (MDS) or principal component analysis (PCA), which reveals a compact representation of a resultant perceptual space. In this perceptual space, similarly rated stimuli cluster and dissimilar stimuli separate. The reader can refer to [7], [8] for more details. The current consensus in the literature [7], [9], [10] is that tactile perception of surfaces can be compressed down to three to five perceptual dimensions, with axes roughly aligned with the rating dimensions of micro and macro roughness/smoothness, hardness/softness, stickiness/slipperiness, and coldness/warmness. The perceptual dimensions obtained for any particular study, however, depend highly on the selected set of surfaces.

Although the above approach gives a general understanding of how humans make perceptual judgments about surfaces, it is inadequate to explain the fundamental relationship between the tactile information elicited from the finger-surface interaction and the resulting perception. Revealing this relation is also crucial for many applications, such as robot perception [11], [12],

[13], product design [14], and haptic rendering [15], [16], [17]. Despite the rich, complex, and unique information available from finger-surface interaction, the existing literature has generally forgone interaction-specific analysis in favor of general surface descriptors: most studies have sought correlations between the derived perceptual space and each surface's physical features (e.g., power spectral density, friction coefficient, average power, spectral centroid, and compressibility) measured in a controlled condition (fixed speed and force) [18], [19], [20], [21]. This approach oversimplifies the complex finger-surface interaction depending on user exploration, as people modify their exploratory movements depending on both the perceptual task and scanned texture to make better perceptual judgments [22]. More importantly, some studies [18], [20], [23] overlooked the importance of finger properties during interaction and focused on surface properties measured via a tool or specific machinery when correlating with a perceptual space that was obtained via free finger exploration.

Here, we aim to understand the fundamental relationship between the tactile information obtained from unconstrained finger-surface interaction and human perception. Specifically, we are interested in which of and to what extent common signal features (e.g., average power, spectral centroid, friction coefficient) calculated from free finger-surface interactions play a role in human perceptual judgments. Since the values of these features change with normal force and scanning speed [15], relating them to perceptual judgments is not straightforward for free exploration. To address this challenge, we first propose a methodology that enables both the conversion of finger-surface interaction signals into a distribution of features and the calculation of the distances between feature distributions from different surfaces based on perceptual similarities rated by humans. Then, based on this methodology, we present general and participant-specific models that can predict the perceptual similarity of two surfaces from their corresponding finger interaction signals. The model parameters and predictions suggest relevant physical features and their weighted roles in human texture perception.

The results indicate that our model is able to predict the perceptual judgments for surface dissimilarities with moderate accuracy despite the great variety in the measured fingertip-surface interactions for the same surface, person, and interaction. We also found evidence that people weigh features differently, suggesting they employ individual mental models when distinguishing surfaces.

II. Methods

We tested our approach on perceptual and interaction data collected from a previous study by Vardar et al. [21]: human participants explored pairs of textures drawn from a set of ten and rated each pair's similarity while their finger-surface interaction data were recorded (Section II-A). First these signals were segmented into the two key exploratory procedures used by participants, tapping and sliding. Then, we partitioned these segmented physical signals into overlapping windows and extracted simple features from each window, resulting in feature distributions for each surface (Section II-B). Finally, we projected these features into a low-dimensional space such that the distances between pair-wise feature distributions match the perceived surface-pair dissimilarities (Section II-C); the models and optimization procedure were implemented in PyTorch (Section II-D).

Fig. 1. The ten surfaces used for the study.

A. Data Collection

The data were collected via psychophysical experiments whose details were previously described [21]. As the physical data presented in this study were not analyzed before, we summarize the details of the experiments in this section.

Seven women and three men with an average age of 28.5 years (SD: 4.14) participated in the experiments. The experimental procedures were approved by the Ethics Council of the Max Planck Society. All participants gave informed consent. The ones who were not employed by the Max Planck Society were compensated at a rate of 8 EUR per hour.

Ten surfaces from the Penn Haptic Texture Toolkit [24] were used as stimuli; the selected surfaces vary in material properties, resulting in a haptically diverse stimulus set (Fig. 1). During the experiments, the participant sat in front of two surfaces (Fig. 2(a)). A black divider was placed between the participant and the surfaces, and the participant wore noise-canceling headphones to mask auditory cues. These interventions ensured that the participants used only haptic cues during the experiment. Each surface was placed on top of a force sensor (Nano 17 Titanium, ATI Inc.). The contact force vector, contact torque vector, and finger acceleration vector were measured during experiments. The force and torque data were collected by a data acquisition board (PCIe 6323, NI Inc.) with a sampling rate of 10 kHz. Two custom-built digital accelerometer boards (MPU-9250, Invensense Inc.) were placed on the index fingernails of both hands of the participant. The accelerometer data were collected via a micro-controller (ATmega32U4, Atmel Inc.) with a sampling rate of 4 kHz. The scene was recorded from above by a high-resolution camera (C920, Logitech Inc.)

In the experiment, each surface pair was placed on the force sensors by taping them to the holders at the edges. After this preparation, the participant was alerted with a sound. They then freely explored the two surfaces for 5 seconds using only their index fingers. Another sound indicated it was time to remove their fingers from the surfaces. Then, the participant

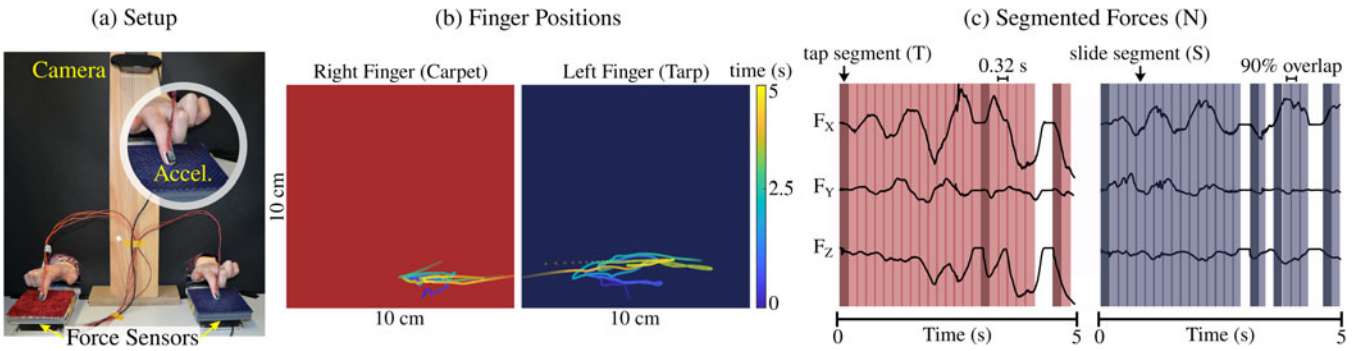

Fig. 2. (a) The experimental setup for data collection. A participant touches a pair of surfaces. The finger-surface interaction data is collected via force sensors placed under each surface and accelerometers (indicated by "Accel.") attached to each index fingernail. A camera records the scene from above. (b) Example of calculated fingertip positions of one participant in one trial. The positions are calculated from the force-torque sensor data assuming each finger makes point contact with its surface. (c) Segmentation process. The force (and simultaneously collected acceleration) data are partitioned into tap and sliding regions based on the velocities of each finger. Each region consists of 320 samples, and each sliding segment overlaps with the former one 90%.

rated the similarity of the pair of surfaces using a nine-point scale. All 45 possible pairs of surfaces were presented twice, with each surface in the pair appearing once on the left and once on the right. Each participant touched the pairs in a different random order. Before each experiment, the participants were given instructions and asked to complete a training session. The training session included one very similar pair (stone tile and leather), one very dissimilar pair (metal foil and carpet), and three random pairs. The very similar and dissimilar pairs were selected based on preliminary experimental results. In total there were 95 trials (5 training + (45 pairs $\times$ 2 locations)). Each participant completed the experiments in two sessions separated by a ten-minute break. The duration of the experiment was about 90 minutes.

B. Fingertip Interaction Features

As opposed to previous studies [18], [19], [20], [21], [25], [26], which represented textures as average features calculated from data collected in controlled conditions, we parse the interaction signals collected in each trial into smaller segments and then calculate features from them. As a result, we obtain a fine-grained distribution of features representing the interactions of each participant with each surface.

1) Segmentation: We compute two types of segments corresponding to the two key exploratory procedures used by participants: tapping and sliding. We define a tap as the moment when contact is initiated between the fingertip and surface, and we define a slide as a period of sustained tangential movement by the fingertip on the surface. To compute the tapping and sliding segments, we first transform the raw force-torque data into position and velocity (Fig. 2(b)) by assuming each fingertip made point contact with the surface. The same technique was used in previous studies [27], [28] to estimate the contact location of a fingertip or a tool on a surface. Before the position was computed, the force and torque signals were down-sampled to 2 kHz using MATLAB's *downsample* function. They were then low-pass filtered using a third-order Butterworth filter with a cut-off frequency of 20 Hz to capture hand motions [15]. The fingertip velocity vectors were calculated by taking the time derivative of the fingertip position vectors. Given the filtered velocity signals, we use MATLAB's *findpeaks* function to select potential taps. Only a peak that immediately follows a region of no contact (exactly zero velocity) is considered a tap peak.

We use the tap peaks to partition 2 kHz down-sampled force and acceleration signals into tap segments and sliding regions (Fig. 2(c)). The tap segment is defined as a 320 sample (0.16 s) window starting from 19 samples before the peak. These values were determined by preliminary screening of the interaction data. Considerably shorter segments would not have captured all the relevant information from a tap interaction, whereas longer ones would have blended tap and sliding interaction data. After computing tap segments, all remaining non-zero velocity regions of the interaction are considered sliding regions. Segments are extracted from slide regions by scanning a 320 sample window (equal size to tap segments) directly after tap segments until the end of the sliding region. Each sliding segment was overlapped 90% with the previous one.

2) Feature Calculation: Select features were calculated from each segment of the 2 kHz signals to represent the three fundamental perceptual dimensions of surfaces: hardness/softness, roughness/smoothness, and friction (sometimes called stickiness/slipperiness) [9], [19]. Features describing surface roughness/smoothness and friction were extracted from slide segments, whereas a feature representing hardness/softness was extracted from tap segments.

Our rationale behind choosing our particular set of features is as follows: previous studies [29], [30] provide evidence that the roughness dimension is composed of both macro and micro roughness, and perceived roughness of the surfaces is related to the intensity and spectral content of the vibrations induced during fingertip sliding [11]. Hence, two metrics were selected to represent the roughness dimension during sliding segments: spectral centroid and vibration power. These two metrics were computed for both the force sensor and the fingernail-mounted accelerometer to enable comparisons between these distinct sources of information. The three-axis force and three-axis acceleration signals were first

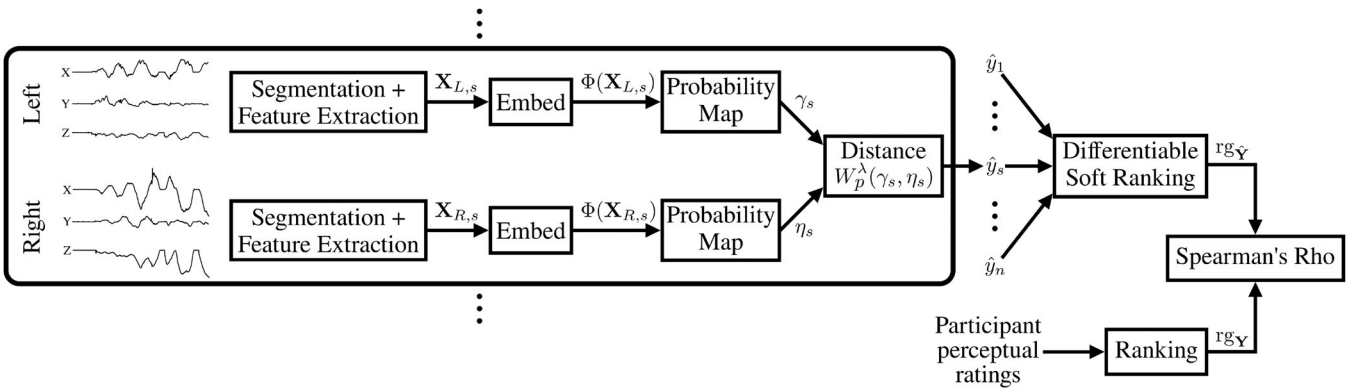

Fig. 3. An arbitrary trial s is comprised of left- and right-handed interactions with different surfaces. The recorded 3D force (shown) and acceleration signals are parsed into many segments over time. Features are then extracted from each of these segment such that each segment is represented as a single point $X_{,s}^{(i)} \in \mathbf{X}_{,s}$ in multidimensional feature space. The two sets of points $\{\mathbf{X}_{L,s}, \mathbf{X}_{R,s}\}$ are then mapped via the function Φ into an embedding space. The point sets are then converted to probability densities γ_s and η_s by assigning probability mass to each embedded point. The optimal transport distance $W_p^\lambda(\gamma_s, \eta_s)$ is computed between the left- and right-handed densities. Finally, the resulting distance $\hat{y}_s$ is ranked relative to the distances of all other trials and compared to rankings of the human similarity ratings. The function Φ is optimized to maximize the Spearman's correlation between distances and rankings.

each combined into one axis using the discrete Fourier transform 3-to-1 (DFT321) method [31]. The spectral centroid was computed by band-pass filtering the compressed signal between 5 Hz and 400 Hz and then taking the fast Fourier transform. For the vibration power, we further filtered the same signals between 20 Hz and 400 Hz and then calculated their average power.

The kinetic friction coefficient was selected as the metric to represent slipperiness. For each slide segment, the kinetic friction coefficient was calculated by fitting a Coulomb friction model to the unfiltered normal and tangential forces.

It has been previously shown that people can discriminate the hardness of a surface from the vibration that occurs after tapping on it with a tool [32]. Because the spectral centroid of this vibration increases with the stiffness of the surface [15], we chose it to represent hardness. Unlike the centroid described above, this spectral centroid was computed during tap segments from the force signal normal to the surface.

In summary, each sliding segment was represented by seven features: finger speed (v), normal force (F_n), kinetic friction coefficient (μ_k), and sliding power ($P_.$) and spectral centroid ($C_.$) calculated from force sensor ($\cdot_f$) and accelerometer ($\cdot_a$) data, whereas each tapping segment was represented by one feature: tap spectral centroid (C_{tap}) obtained from force sensor data. Therefore, the interaction data collected from one finger in each trial was reduced to the collection of seven + one different features calculated from each sliding or tapping segment of the entire interaction.

C. Modeling Framework

Our method aims to learn the relationship between the features extracted from the segments of raw tactile data and the perceptual similarity ratings provided by the participants. We do this by considering the set of segments extracted from the left- and right-handed interactions as two discrete probability distributions. We learn a mapping from feature space into a lower-dimensional embedding space such that the distances between the pairs of embedded distributions agree with the corresponding similarity ratings. We will first introduce the problem definition and give a general overview of the entire modeling pipeline in Section II-C1. We then describe the details of the individual components of the pipeline. Fig. 3 shows a summary of the full pipeline.

1) Learning Problem: Let the set of all trials be denoted $\mathbf{S}$ and the set of corresponding similarity ratings be denoted $\mathbf{Y}$. Given a single trial $s \in \mathbf{S}$ with rating $y_s \in \mathbf{Y}$, the left- and right-handed interactions L_s and R_s with l_s and r_s segments, respectively, can be represented by matrices $\mathbf{X}_{L,s} \in \mathbb{R}^{l_s \times 8}$ and $\mathbf{X}_{R,s} \in \mathbb{R}^{r_s \times 8}$, where 8 is the total number of features. Each row of a matrix $\mathbf{X}$ contains the features calculated from a single segment of that interaction and can be written as

$$\mathbf{X}^{(i)} = \{v,\, F_n, \mu_k, P_f, P_a, C_f, C_a, C_{tap}\}, \tag{1}$$

where i denotes an arbitrary segment. If i is a sliding segment, the feature C_{tap} (last vector element) is assigned zero. Otherwise, the other seven features are assigned zero. Note that interaction matrices $\mathbf{X}$ can have different numbers of rows/ segments.

Additionally, to learn a compact representation of the features that more closely represents the human perceptual space, we define a mapping function $\Phi : \mathbb{R}^m \mapsto \mathbb{R}^n$ from the m-dimensional fingertip interaction feature space to an n-dimensional embedding space. We will describe this mapping function in greater detail in Section II-C3. This mapping function $\Phi(\mathbf{X})$ embeds each row of $\mathbf{X}$ as a unique point in $\mathbb{R}^n$. The projections of the left- and right-handed interactions $(\Phi(\mathbf{X}_{L,s}),\ \Phi(\mathbf{X}_{R,s}))$ can be represented as discrete probability distributions γ_s and η_s, with

$$\gamma_s = \sum_{i=0}^{l_s} \mathbf{g}_i \delta_{\Phi_i(\mathbf{X}_{L,s})} \quad \text{and} \quad \eta_s = \sum_{i=0}^{r_s} \mathbf{h}_i \delta_{\Phi_i(\mathbf{X}_{R,s})}, \tag{2}$$

where $\mathbf{g}$ and $\mathbf{h}$ are non-negative vectors summing to 1 and $\delta_{\Phi_i(\cdot)}$ is the Dirac delta function centered at the point indicated

by the i-th row of $\Phi(\mathbf{X})$. Then, $\hat{y}_s \in \hat{\mathbf{Y}} := \{\hat{y}_s \; \forall \; s \in \mathbf{S}\}$ is defined as the distance between probability distributions γ_s and η_s for the specific trial s. Specifically, we use the Wasserstein distance function, which we describe in Section II-C2.

Given this notation, the learning problem can generally be described as optimizing a parameterized mapping function Φ that maximizes the correlation between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$. Because Likert scales provide qualitative, ordinal data, we are specifically interested in maximizing the *rank correlation* between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$. This is called the Spearman's correlation, and it can be defined specifically for this problem as

$$\rho_{\text{sp}}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\text{cov}(\text{rg}_{\hat{\mathbf{Y}}}, \text{rg}_{\mathbf{Y}})}{\sigma_{\text{rg}_{\hat{\mathbf{Y}}}} \sigma_{\text{rg}_{\mathbf{Y}}}}, \tag{3}$$

where $\text{rg}_{\hat{\mathbf{Y}}}$ and $\text{rg}_{\mathbf{Y}}$ are the rank variables of $\hat{\mathbf{Y}}$ and $\mathbf{Y}$, $\text{cov}(\text{rg}_{\hat{\mathbf{Y}}}, \text{rg}_{\mathbf{Y}})$ is the covariance of the rank variables, and $\sigma_{\text{rg}_{\hat{\mathbf{Y}}}}$ and $\sigma_{\text{rg}_{\mathbf{Y}}}$ are the standard deviations of the rank variables. We implement a differentiable ranking function that is described in Section II-D.

2) Regularized Wasserstein Distance: To compute the distance between probability distributions, we use the p-Wasserstein distance, which is the solution to the traditional optimal transport problem and essentially measures the minimum cost of transporting the mass from one probability distribution to another in a metric space [33]. Although there are other popular methods of measuring the similarity between probability distributions, such as the Kullback–Leibler (KL) and Jensen-Shannon divergence, we chose the Wasserstein metric because it is symmetric (unlike KL-divergence), can be computed on distributions that do not share a support set, and has usable gradients over the entire support set [34]. This distance can be extremely costly to compute for both continuous and discrete distributions. Thus, we use the entropy-regularized p-Wasserstein distance, which approximates the true Wasserstein distance but admits a simpler solution that can be computed orders of magnitude faster using GPUs [35]. Given two discrete measures γ and η with G and H (in our case l_s and r_s) support points, respectively, the discrete, entropy-regularized p-Wasserstein distance with regularization parameter λ is defined as

$$\begin{aligned} W_p^{\lambda}(\gamma, \eta)^p = & \min_{T \geq 0} \; \text{tr}(D^p T^\top) - \frac{1}{\lambda} h(T) \\ \text{s.t.} \quad & T\mathbf{1} = \gamma, \quad T^\top \mathbf{1} = \eta, \\ \text{with} \quad & h(T) = -\sum_{i=1}^{G} \sum_{j=1}^{H} T_{i,j} \log(T_{i,j}). \end{aligned} \tag{4}$$

$D^p \in \mathbb{R}_+^{G \times H}$ is a matrix of distances with $D_{ij}^p = d(x_i, y_j)^p = ||x_i - y_j||_p^p$ and $T \in \mathbb{R}_+^{G \times H}$ is the discrete transport plan with T_{ij} the probability mass transported from γ_i to η_j [35], [36]. $T\mathbf{1} = \gamma$ and $T^\top \mathbf{1} = \eta$ are the marginal constraints on T. The optimal T can be solved for using Sinkhorn's fixed point iteration. The black lines between points in Fig. 4 display the elements T_{ij} of an example transport plan with a single element highlighted in orange. More information about optimal transport and the Wasserstein distance can be found in [33], and specific details about the discrete Wasserstein distance with entropic regularization appear in [35], [36].

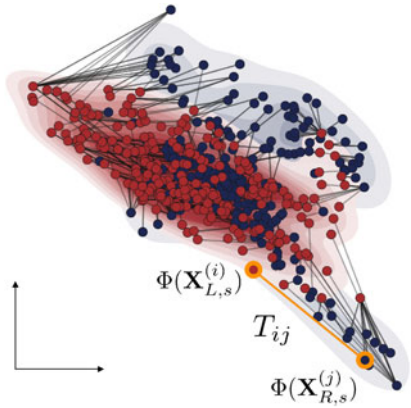

Fig. 4. Example transport plan between two sets of embedded points $(\Phi(\mathbf{X}_{L,s}), \Phi(\mathbf{X}_{R,s}))$ in 2D space. Two example points $\Phi(\mathbf{X}_{L,s}^{(i)})$ and $\Phi(\mathbf{X}_{R,s}^{(j)})$ and their corresponding transport plan T_{ij} are highlighted in orange. The underlying probability mass distributions are shaded by blue and red behind the individual points.

Given this probability metric, $\hat{y}_s = W_p^{\lambda}(\gamma_s, \eta_s)$, where γ_s and η_s from Eq. (2) are the discrete probability distributions defined over the embedding space for trial s.

3) Mapping Functions: We use two different types of mapping functions Φ in our experiments to embed the features extracted from the tactile data: affine maps and fully connected neural networks. These two choices represent two different levels of embedding complexity, with the affine maps having the simpler, more constrained embedding resulting from fewer degrees of freedom compared to the neural networks. For the affine maps,

$$\Phi_{\text{af}}(\mathbf{X}) = \theta \mathbf{X}^\top + \beta, \tag{5}$$

where $\theta \in \mathbb{R}^{m \times n}$ are the linear mapping parameters and $\beta \in \mathbb{R}^n$ are the biases.

For the neural network, we employ a single hidden-layer architecture with rectified linear unit (ReLU) activation functions. The general structure is then

$$\Phi_{\text{nn}}(\mathbf{X}) = \theta_{(o)} \cdot \text{ReLU}(\theta_{(h)} \mathbf{X}^\top + \beta_{(h)}) + \beta_{(o)}, \tag{6}$$

where $\theta_{(h)} \in \mathbb{R}^{m \times k}$ and $\beta_{(h)} \in \mathbb{R}^k$ are the weights and biases of the hidden layer with output dimension k and $\theta_{(o)} \in \mathbb{R}^{k \times n}$ and $\beta_{(o)} \in \mathbb{R}^n$ are the weights and biases of the output layer with dimension n.

D. Implementation

All optimization of the parameters θ of Φ was performed using stochastic gradient descent and back-propagation with a loss function of

$$\mathcal{L}(\theta) = 1 - \rho_{\text{sp}}, \tag{7}$$

where ρ_{sp} is the Spearman's correlation from Eq. (3).

One difficulty of implementing this loss function is that computing rank variables (e.g., $\mathrm{rg}_{\hat{\mathbf{Y}}}$ and $\mathrm{rg}_{\mathbf{Y}}$) is typically non-differentiable. To address this issue, we use a regularized, differentiable soft-rank function that approximates exact rankings [37]. The soft-rank function uses regularization to trade off between a more accurate ranking (smaller regularization) and a more strongly convex (and continuously differentiable) optimization (larger regularization).

The full optimization procedure was implemented in Python and PyTorch. The built-in Adam optimizer was used with a learning rate of 0.01 and default values for the remaining parameters. The ranking of the Wasserstein distances was performed using the soft-rank PyTorch implementation from Blondel et al. [37] with a regularization of 0.1, a value which provided a reasonable trade off between accuracy and convexity in preliminary experiments.

We used the regularized 1-Wasserstein distance (with the distance function $d(x_i, y_j)$ the L_1 norm) and computed it using the auto-differentiable Sinkhorn implementation by Gabriel Peyré[1] with a regularization of 0.1 chosen from preliminary experiments. Additionally, the two weight vectors $\mathbf{g}$ and $\mathbf{h}$ from Eq. (2) were defined such that probability mass was distributed uniformly across all points in an interaction. That is, for trial s with l_s and r_s segments, $\mathbf{g}_s = \mathbf{1}/l_s$ and $\mathbf{h}_s = \mathbf{1}/r_s$.

III. Modeling Procedure and Computational Experiments

Computational experiments were conducted to both evaluate the performance of the method and to learn more about the perceptual models of individual participants. As such, it was important to balance model interpretability with performance.

With this goal in mind, we first compared the performances of more complex, non-linear models with simpler affine models across a variety of embedding dimensions, demonstrating that simpler, more interpretable models were sufficient.

We then trained simple models to test the generalizability of the method to unseen participants and fine-tuned those general representations to individual participants. We analyze and compare the model structures to try to understand differences between the general, "average" representations and the representational perceptual structures of the individual participants. Additionally, this analysis allows us to look at differences between individual participants.

A. Constructing General Models

General models were trained in two distinct ways. First, we ran a preliminary experiment to compare the performance of neural networks and affine maps as a function of the embedding dimension. We trained both types of models on data from all participants. We used a small neural network architecture of one hidden layer with eight nodes. We found that larger networks with a variety of regularization schemes and nonlinear activation functions did not outperform the smaller models (details shown in Section S1).

[1] https://github.com/gpeyre/SinkhornAutoDiff

Second, we trained general affine map models on data from a subset of participants and evaluated those models on unseen participants. We did not perform this second training procedure with neural networks because the neural networks' slight edge in performance in the first experiment did not outweigh the greater interpretability of the affine maps. This finding is explained in greater detail in Section IV-A.

1) Model Comparison: For the first case, five-fold nested cross-validation was used to train preliminary comparison models. To form the folds, the samples from each participant were partitioned into five equally-sized, stratified groups, with each group having a roughly equal distribution over the ratings. Then, each of the five groups was added to a separate fold. A single fold was held out of the training process for testing, and a model was trained and evaluated on every possible three-one split of the remaining four folds. Thus, there were four models trained for each hold-out. Each fold was held out as a test set, yielding a total of 20 trained models (4 per fold $\times$ 5 folds). For each training run the features were mean-centered for each participant independently using the data in the three training folds.

2) General Affine Models: For the training procedure of the general affine models, there were ten folds with each fold containing all the data from a different participant. The same process described above was performed, yielding a total of 90 trained models (9 per fold $\times$ 10 folds). In this case, the features for each participant were independently mean-centered using all their data.

In all cases, models were trained with a batch size of 180 for 200 epochs. The model state with the best validation performance over the 200 epochs was kept. Additionally, the loss was calculated on a per-participant basis and then averaged over participants. The participant-wise loss differs slightly from Eq. (7) and can be formulated as

$$\mathcal{L}(\theta) = 1 - \frac{1}{|J|}\sum_{j \in J} \rho_{\mathrm{sp}}(\hat{\mathbf{Y}}_j, \mathbf{Y}_j), \tag{8}$$

where J is the set of participants and $\hat{\mathbf{Y}}_j \subset \hat{\mathbf{Y}}$ and $\mathbf{Y}_j \subset \mathbf{Y}$ are the subsets of distances and ratings, respectively, for participant j. That is, the Spearman's correlation ρ_{sp} was calculated independently for each participant.

B. Participant-Specific Modeling

To measure how the perceptual representations of individual participants differed from the generalized representations trained on other participants, we tuned general models to specific participants instead of training participant models from random initial conditions. Specifically, the participant-specific models for a particular participant were initialized using the best-performing (on the validation set) of the nine general models that were trained with that participant held out. To train the participant-specific models, a participant's data were split into the same five folds used in the comparison model training. The models were trained for 100 epochs instead of 200 while the rest of the training, validation, and testing

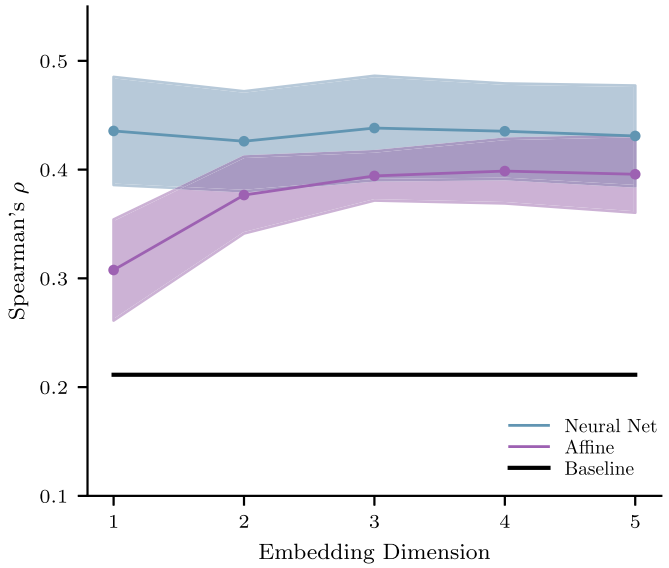

Fig. 5. Mean and standard deviation of model performance vs. embedding dimension by model type. The neural networks all have one hidden layer with eight nodes. The baseline loss on the dataset with no feature mapping is indicated by the solid black line.

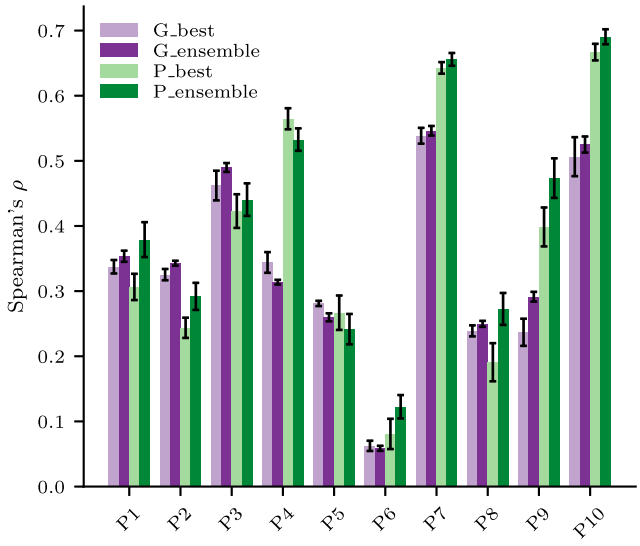

Fig. 6. Means and standard errors of the best general (G_best), general ensemble (G_ensemble), best participant-specific (P_best), and participant-specific ensemble (P_ensemble) models.

procedure remained the same. Features were mean-centered using the data in the training folds.

IV. Results

A. Model Type and Embedding Dimension

To measure the modeling performance as a function of model type and embedding dimension, we trained and evaluated neural networks and affine map models with outputs from one to five dimensions using five-fold nested cross-validation, as described in Section III-A1. Four models were trained for each testing fold, and of those four models, the one that performed the best on the validation set was then evaluated on the test set. Thus, for every full training procedure, five models were evaluated, one for each fold. To account for the random initialization of model parameters, the full modeling procedure described above was performed ten times for each embedding dimension and each model type. Thus, there are a total of 50 (5 folds $\times$ 10 random model seeds) evaluated models of each type (neural net and affine map) for each embedding dimension. The means and standard deviations of these test set evaluations are shown in Fig. 5. To make the results clearer, we show $1 - \mathcal{L}(\theta)$ instead of $\mathcal{L}(\theta)$, which represents the Spearman correlation ρ between the predictions and psychophysical ratings. The baseline represents the loss on the original features with no mapping, i.e., $\Phi = \mathbf{1}$.

As can be seen in Fig. 5, the ability to train an additional, low-dimensional embedding represents a considerable increase in performance for all values of embedding dimensionality. Furthermore, the neural network models marginally outperform the affine models, especially for a low embedding dimensionality of 1. However, there seems to be no additional benefit of adding further dimensions for neural network mappings. Given that affine maps in general are more interpretable compared to neural networks and that their performance saturates at an embedding dimension of three, we exclusively learned affine map models into three dimensions for our remaining experiments.

B. Generalizability

To test the generalizability of the modeling method to unseen participants, we trained affine maps into three dimensions on a subset of participants and evaluated them on unseen participants, as described in Section III-A2. Again, we analyze the performance of the best models by evaluating only the best validation model on the associated test fold (remember, each fold is a single participant). However, evaluating only the top-performing models could introduce bias if particular validation sets were always modeled more accurately than others. Thus, we also measure the ensemble performance of all the models trained for each test fold. Specifically, we compute $\hat{\mathbf{Y}}$ for each of the nine models, normalize each $\hat{\mathbf{Y}}$ so that all distances are between 0 and 1, take the average across all $\hat{\mathbf{Y}}$, and then compute the Spearman's correlation between the averaged distances and the corresponding similarity ratings.

As above, we repeat the full modeling process ten times to account for randomness in the initial model parameters. The mean performances of the best validation models (G_best) and the ensembles (G_ensemble) are shown in Fig. 6, with error bars indicating the standard error of the mean.

Although the generalization performance differs substantially by participant, the average performance across all participants is very similar to the performance of the 3D affine model. Additionally, there is little change in performance between the best and ensemble predictions. Participants 3, 7, and 10 are modeled fairly well, whereas participant 6 is almost completely unpredictable. This finding suggests that much of the information about how most participants rated similarity is either not captured by the model or not contained in the data at all.

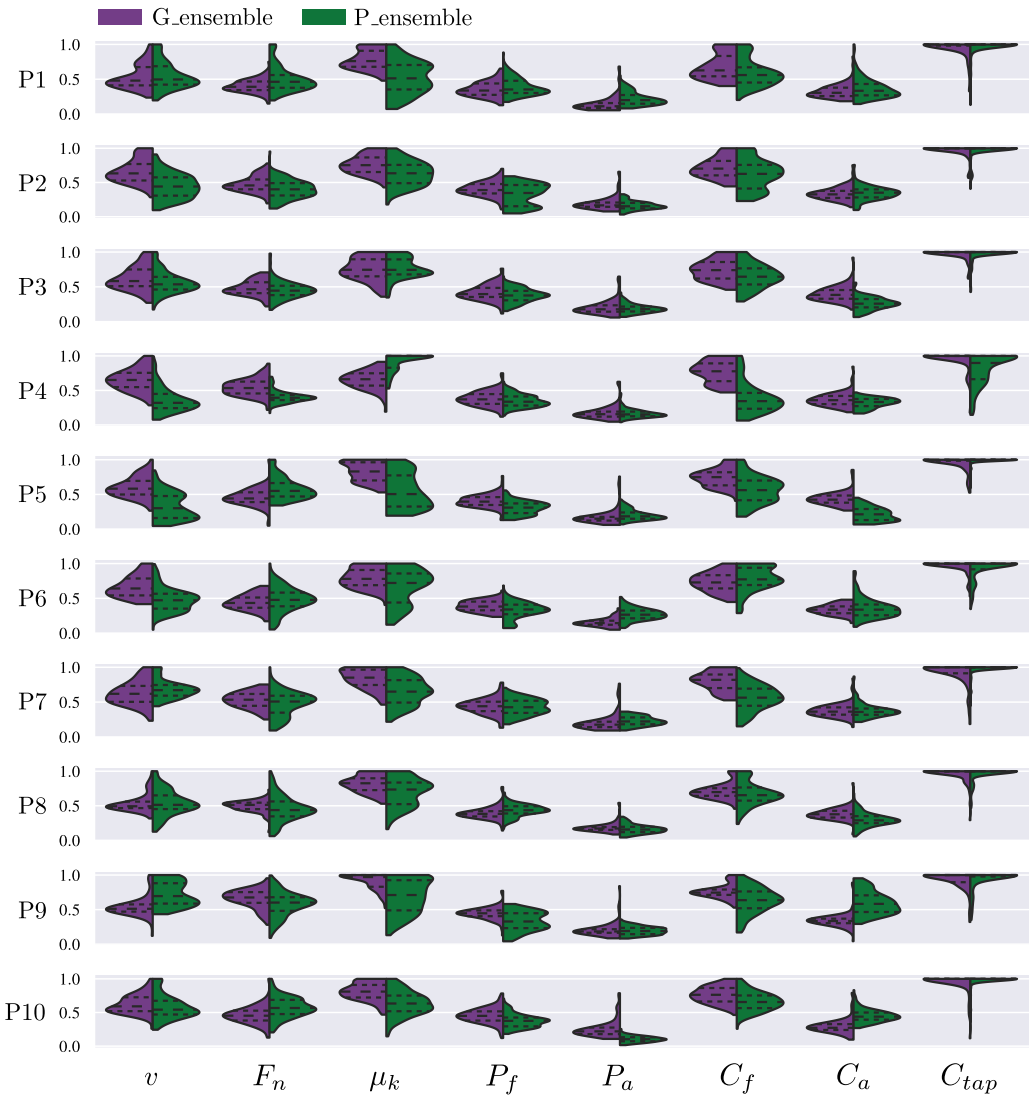

Fig. 7. Distributions of normalized feature axis lengths by participant. Purple regions show distributions of normalized axis lengths for the trained general models by participant holdout. Dark green shows the distributions of normalized axis lengths for the fine-tuned participant models.

C. Participant-Specific Model Tuning

From each of the ten randomized runs, the best general affine model for each participant hold-out was used as the starting configuration to train participant-specific models. Using this method, we can make direct comparisons between the tuned models and the original general models. We measure performance in the same way as above, evaluating both the best validation model for each test fold and the ensemble predictions. The mean performances of the best tuned models (P_best) and the ensembles (P_ensemble) are shown side by side with the general model performance in Fig. 6.

In general, there is an improvement in performance when the models are individually tuned to individual subjects, particularly for participants 4, 7, 9, and 10. The performance for participant 3 is still relatively good, although there is no increase in accuracy. While there is a minor improvement for participant 6, the performance is still particularly poor. Again, there is little difference between the accuracy of the best and the ensemble models in most cases.

D. Model Analysis for Perceptual Characterization

One method of analyzing a simple affine model is to project the original feature axes into the embedding space and measure the relative scales of the axes. Because the Wasserstein distance depends on the distances between points in the metric space, a feature axis with a larger scale contributes more to the overall Wasserstein distance than a feature axis with a smaller scale.

To compute the relative axis scales for a single model, the unit vector along each feature axis can be projected into the embedding space. The projected vector lengths can all be divided by the magnitude of the longest vector to scale them between zero and one. Different models can be compared by normalizing the projected vector lengths for all models. This process was performed for the general models trained with participant holdouts and for the models that were tuned to specific participants. Fig. 7 shows the density estimates of the relative axis lengths by participant for the general (purple) and participant-specific (green) models. We show the results for the ensemble of models as

opposed to only the best performing models. These are the same models whose performance is plotted in dark purple and dark green in Fig. 6.

There are some observable patterns across subjects and different modeling scales. Clearly, the tap spectral centroid (C_{tap}) is consistently one of the largest embedded feature dimensions, meaning it contributes more to the overall Wasserstein distance than other dimensions. Conversely, the average vibration powers measured from both the force sensor (P_f) and accelerometer (P_a) are the smallest feature dimensions. Thus, the average vibration power does not greatly contribute to the Wasserstein distance. Additionally, for both pairs of features that were computed from both sensors, the feature computed from the accelerometer is always smaller than the corresponding feature computed from force data.

Interestingly, the models seem to get less consistent when they are tuned. For many features, the spread (height of the densities) actually increases from the general to the tuned models. This trend is most clearly demonstrated by the friction coefficient (μ_k) and force sensor slide spectral centroid (C_f).

V. Discussion

In this paper, we tried to solve the unique problem of predicting human perception from individual haptic experiences by aiming to understand the physical factors governing these perceptual judgments. Specifically, we proposed a method that predicts the perceived similarity of two surfaces from the features extracted from the physical signals elicited during the interaction. The results demonstrate that this method somewhat works on both general and participant-specific levels. General representations learned on a subset of participants can partially predict the perceptual similarities of unseen participants with accuracies ranging from low to high depending on the participant. Analysis of the model structures provides a method to interpret the weights of different haptic properties in the perceptual similarity judgments of different people, albeit with limited confidence due to the model performance.

A. Complex Versus Simple Models

A key question about this method is whether a simple model is sufficient to capture the relationships between the tactile features and similarity ratings. The results shown in Fig. 5 answer this question, demonstrating that simple affine models are comparable in performance to more complex neural networks despite having fewer than half the parameters; the additional experiments in Section S1 confirm this finding. The neural network models do perform marginally better, but the small improvement demonstrates that the method is not primarily limited by the model type, at least for this particular dataset and choice of features.

The consistent performance as a function of the number of embedding dimensions, particularly for neural networks, provides additional evidence that the performance limitations are not due to the model architectures and that the Wasserstein metric has large representational capacity across a number of embedding dimensions.

Overall, the average performance reaches "only" levels of $\rho = 0.4$. One reason behind this moderate performance could be the significant noise in the participant ratings. The participant agreement can be measured by computing the Spearman's correlation for each pair of participants over all 90 trials and averaging, yielding an inter-rater agreement of 0.707 [21]. Thus, the consistency of ratings across participants likely provides an approximate upper bound on the modeling performance. It is possible but highly unlikely that all the rating noise can be explained by the data contained in each interaction, as humans are imperfect perceptual machines subject to inconsistency, distraction, and fatigue. Additionally, finding strong correlations between surface properties and human perception has been proven to be difficult. For example, Bergmann Tiest and Kappers [38] had subjects order a set of surfaces by roughness and found Spearman's correlations from 0.4 to 0.8 (depending on the subject) between the perceptual orderings and the physical roughness measures of the surfaces.

Another underlying reason for the moderate prediction performance of our model could be its use of selected features. Although we included the most common physical factors mentioned in the literature, the ones that we did not consider (e.g., thermal conductance, spatial finger deformation, or skewness and kurtosis of the segments) may have significant effects on similarity judgments. It is also possible that human tactile processes do not estimate physical quantities but seek to estimate statistical variations in the tactile signals. This hypothesis has also been proposed for visual [39], [40] and audio [41] senses. In a recent study [42], Metzger and Toscani trained a deep neural network with unsupervised learning to reconstruct vibratory signals elicited by human exploration of surfaces using a tool. They found that the learned latent space could classify different material categories similar to perceptual distances rated by human participants. If this is the case, it would be advantageous to construct a mapping from this latent space to the perceptual space without segmenting and calculating physical features from the original tactile signals. This study omitted that option as we wanted to find relations between physical factors and perception.

B. Generalization and Specialization

By training models on subsets of participants and testing the performance on unseen participants, we demonstrate that our method can find an average perceptual representation across multiple people that can reasonably predict the perceptual similarity judgments of unseen participants. Tailoring these general representations to individual participants suggests that the perceptions of each participant differ uniquely from the average but mostly can be captured by the tuned models.

Figure 6 demonstrates that the general models perform quite differently depending on the participant. They perform exceptionally well for participants 3, 7, and 10, but perform terribly for participant 6. This difference in performance likely indicates that there is some consistency across participants in how they judge similarity, but there are many differences that cannot be explained in an average model. However, it is possible

that participants 3, 7, and 10 all employ a more similar rating strategy than the rest of the participants.

When the models are tuned, the accuracy improves most significantly for participants 4, 7, 9, and 10. Participant 3 still performs well even though the tuned models are not more accurate. This result provides evidence that, at least for these participants, a large part of their perceptual similarity judgments can be explained by the simple models and features that we used. It is particularly interesting that participants 4 and 9 improve quite clearly. It is possible that each of them relies primarily on the features that we included, but they treat them differently from all the other participants.

There are a variety of possible explanations for the comparatively worse performance on the other participants. For example, they may have relied more heavily on tactile signals that were not captured in our small feature set. As mentioned before, one feature in particular that was not included was the thermal conductivity of the surfaces. Temperature perception could have been a dominant cue in many cases, particularly for surface pairs that included aluminium [43]. Other explanations could be that these participants used unique strategies to determine similarity or were inconsistent in applying their strategy. An example strategy could be to consider a surface pair very dissimilar if it differs dramatically in only a single dimension. An alternative strategy could be to consider a surface pair as similar unless it dramatically differs across multiple dimensions. Our method does not currently account for the use of different strategies, although we will discuss how this might be addressed in Section V-D.

C. Inferring Perceptual Structure

The main benefit of using affine maps instead of neural networks is that their simplicity allows us to interpret the learned models and draw inferences about the participants' tactile perceptual representations. We focus on comparing the relative scales of the original feature axes projected into the learned embedding spaces. Despite the large amount of variance in perception that is not captured by our models, we propose that the larger features can be interpreted as more perceptually relevant. Given this assumption, it is immediately clear that, overall, the tap spectral centroid (C_{tap}) is a relevant feature. There are typically many fewer tap segments than slide segments, which means that much less probability mass is assigned to the tap segments overall. The large relative scale of C_{tap} demonstrates that despite the low mass, the tap segments provide unique information and are very important in modeling similarity. This holds true across all participants in both the general and tuned models. Considering the large variety in hardness of the selected surfaces (Fig. 1) and that every trial started with a tap, it is indeed reasonable that hardness-relevant cues played an important role in similarity judgments.

Friction (μ_k) and the slide spectral centroid (C_f) are also relatively important compared to other features. Interestingly, a recent study [17] also found these features correlated with the two main axes in the perceptual space of fine textures created on friction modulation displays. Hence, the results suggest that friction and the slide spectral centroid could be relevant physical parameters for surface perception via direct fingertip touch.

On the other hand, both average vibration power features (P_f and P_a) are consistently the smallest of the features, with P_a being especially small. This means that these features did not contribute substantially to the distance between surface pairs. Thus, it is unlikely that the participants considered vibration power a relevant cue when measuring the similarity of the selected surfaces. Nonetheless, earlier studies [18], [19] found that vibration power correlated with one of the main perceptual dimensions. A likely reason for this discrepancy is the difference in data collection. In both of these earlier studies, the physical interaction data was collected via a tool, whereas we analyzed data that occurred during finger-surface interactions. The variety of selected surfaces and the range of motions used could also contribute to this discrepancy.

Interestingly, both features computed from the accelerometer (P_a and C_a) are typically smaller than their counterparts computed from the force sensor (P_f and C_f). This likely means that the force sensor mounted rigidly to the surface more accurately captured the fingertip-surface interaction than the accelerometer mounted to the fingernail; it is possible that the accelerometer data is even confounding. Due to the complex mechanical properties of the human finger and the fact that vibrations do not travel well from the fingerpad to the fingernail [44], [45], the vibrations transmitted to the accelerometers likely differed substantially from those measured at the force sensors. Additionally, the limited sensitivity and noise susceptibility of the fingernail-mounted accelerometers compared to the force sensor could cause this discrepancy in sensor relevance.

For many features, the height of the densities (i.e., the spread of relative feature scales) actually increases from the general to the tuned models. However, we believe that the increase in spread is caused by the much smaller amount of data on which the tuned models are trained and the high variance in the data across folds. With more training examples for individual participants, the models would likely become more uniform and the feature densities narrower.

There is visible variability in the features that different participants relied on when making similarity judgments (Fig. 7). For example, participant 4 seems to consider friction (μ_k) as highly relevant compared to the other participants. Additionally, the narrower densities of many features in the tuned models could explain why the performance increases dramatically from the general to those tuned models; participant 4 models similarity in a predictable way, but somewhat differently from all the other participants. Participant 9 also has tuned model distributions that differ substantially from the general models, particularly with regard to the velocity (v) and slide spectral centroid (C_f and C_a). On the other hand, participants 3, 7, and 10 have tuned model distributions more similar to the corresponding general model distributions, meaning that the general model was able to explain these participants' perceptual similarity judgments as well as possible with the given data.

Nonetheless, it is difficult to conclude much about the participants for which the modeling does not perform well. The

predicted models of these participants could be accurate representations of their perceptual structure within the limitations of the used dataset. The poor prediction performance of their models could be explained by their inconsistent rating strategies among different surfaces. It is also possible that they relied on other tactile cues not presented in the data (e.g., thermal conductivity, stickiness, absorbency).

D. Limitations, Future Directions, and Potential Applications

Our method performed moderately for predicting general perceptual representations and better for some individual participants. This work has several limitations and sources of variability that we believe limited the potential performance; many of these factors could be individually addressed in future experiments.

The dataset has a limited number of participants who each made a limited number of surface comparisons. Likely, with more participants, more surfaces, and more surface comparisons, there would be less noise in the similarity ratings, and it would be possible to learn more predictive models. Additionally, the participants never compared two of the same surface. Comparing identical surfaces could provide valuable information about the consistency of user ratings as well as a powerful comparison that the model might have been able to use to more strongly cluster similar surfaces.

We used a limited set of haptic features to represent the finger-surface interactions. While these features do correspond to primary tactile perceptual dimensions, it might be that secondary properties also contribute to similarity perception. As mentioned earlier, surface thermal conductivity was not included. There are additional vibration-related features, such as the spread or skewness of the frequency spectrum [12], that we did not include, and that could be included in future studies. Additionally, there is some evidence that not only temporal but also spatial features of surfaces play a role for perception during both static and dynamic exploration [5], [46]. As explained earlier, it is also possible that human similarity judgments do not rely on estimation of physical quantities but rather solely on statistical variations in the tactile signals [39], [42]. In the future, this hypothesis can be tested by implementing unsupervised learning methodologies on unsegmented tactile signals elicited from finger-surface interactions.

Our method did not account for the possibility that people can use varying strategies to judge surface similarity. However, we believe that with minor changes this method could be extended to account for at least some strategic variance. The opportunity to provide strategic diversity lies in how probability mass is assigned to individual interaction segments, specifically how the vectors $\mathbf{g}$ and $\mathbf{h}$ are defined in Eq. (2). As described in Section II-C2, we assigned mass uniformly across all segments. Because segments are sampled using discrete time windows, this means that low-velocity regions of the interactions automatically have a higher concentration of probability mass than high-velocity regions and thus contribute more to the Wasserstein distance. As a strategy, this could be described as participants weighing regions of low-velocity more heavily than others. However, normalizing the probability mass assignment by velocity (low-velocity segments have lower mass and high-velocity segments have higher mass) represents a different strategy where unique regions of the feature space are weighed independently of the velocity. These are just two examples, but there are many more strategies that can be captured by modifying the probability mass assignment.

Overall, our method was able to model similarity judgments of many participants with moderate accuracy. The general model performances demonstrate that similarity judgments are extremely complex, and more information and method flexibility are necessary to capture judgments more accurately. However, even with our limited number of features, small model size, and simple mass-assignment strategy, we did find some consistent patterns explaining similarity judgments. By tuning models to specific participants, we found that the judgments can be explained more accurately in many cases. We believe these initial results are promising for the utility of this method to explain complex perceptual processes and how different people weigh various tactile features; future experiments could more precisely test how individual participants use different features. Moreover, given surface-finger interaction data or computed features from two different surfaces, our model can give a good approximation of the perceived similarity of these two surfaces without the need for time-intensive perception experiments.

In general, we believe our approach can help derive a deeper understanding of human tactile perception that can be applied across multiple domains. For example, by considering which tactile properties are relevant in an individual's texture preferences, recommender systems could suggest particular clothing or other textured objects. These properties could be captured by a haptic robot that learns what exploratory procedures most efficiently elicit the relevant data. Alternatively, haptic rendering systems could generate more realistic virtual textures by altering specific characteristics of the haptic output to better match the patterns seen in real textures over short time windows.

ACKNOWLEDGMENT

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Benjamin Richardson. B.A.R. thanks Tom Richardson for the helpful conversations about optimal transport. Y.V. thanks David Abbink for their fruitful discussion about the study.

Data and Code Availability

The code and experimental data are available through https://github.com/MPI-IS/Learn2Feel.

Author Contributions

B.A.R., Y.V., C.W., and K.J.K. conceptualized the study; B.A.R. developed and implemented the modeling methodology and analyzed experimental data; Y.V. acquired and analyzed experimental data and contributed to model development; C.W. and K.J.K. provided guidance for data acquisition and analysis, model development and implementation; K.J.K. supervised the study; B.A.R., Y.V., C.W., and K.J.K. wrote the paper.

REFERENCES

[1] R. Johansson and J. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Rev. Neurosci.*, vol. 10, no. 5, pp. 345–359, 2009.

[2] L. R. Manfredi et al., "Natural scenes in tactile texture," *J. Neuriophysiol.*, vol. 111, no. 9, pp. 1792–1802, 2014.

[3] A. Abdouni, M. Dhaghloul, C. Thieulin, R. Vargiolu, C. Pailler-Mattei, and H. Zhaouani, "Biophysical properties of the human finger for touch comprehension: Influences of ageing and gender," *Roy. Soc. Open Sci.*, vol. 4, no. 8, 2017, Art. no. 170321.

[4] B. Delhaye, A. Barrae, B. B. Edin, P. Lefèvre, and J. L. Thonnard, "Surface strain measurements of fingertip skin under shearing," *J. Roy. Soc. Interface*, vol. 13, no. 115, 2016, Art. no. 20150874.

[5] A. I. Weber et al., "Spatial and temporal codes mediate the tactile perception of natural surfaces," *Proc. Nat. Acad. Sci.*, vol. 110, no. 42, pp. 17 107–17 112, 2013.

[6] D. J. Meyer, M. A. Peshkin, and J. E. Colgate, "Tactile paintbrush: A procedural method for generating spatial haptic texture," in *Proc. IEEE Haptics Symp.*, 2016, pp. 259–264.

[7] S. Okamoto, H. Nagano, and Y. Yamada, "Psychophysical dimensions of tactile perception of textures," *IEEE Trans. Haptics*, vol. 6, no. 1, pp. 81–93, Jan.–Mar. 2013.

[8] K. Drewing, C. Weyel, H. Celebi, and D. Kaya, "Systematic relations between affective and sensory material dimensions in touch," *IEEE Trans. Haptics*, vol. 11, no. 4, pp. 611–622, Oct–Dec. 2018.

[9] M. Hollins, R. Faldowski, S. Rao, and F. Young, "Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis," *Percep. Psyhophysics*, vol. 54, pp. 697–705, 1993.

[10] M. Hollins, S. Bensmaia, K. Karlof, and F. Young, "Individual differences in perceptual space for tactile textures: Evidence from multidimensional scaling," *Percep. Psyhophysics*, vol. 62, pp. 1534–1544, 2000.

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

[12] 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.

[13] B. A. Richardson and K. J. Kuchenbecker, "Learning to predict perceptual distributions of haptic adjecives," *Front. Neurorobot.*, vol. 13, pp. 1–16, 2020.

[14] G. Elkharraz, S. Thumfart, D. Akay, C. Eitzinger, and B. Henson, "Making tactile textures with predefined affective properties," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 57–70, Jan.–Mar. 2014.

[15] 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.

[16] A. Isleyen, Y. Vardar, and C. Basdogan, "Tactile roughness perception of virtual gratings by electrovibration," *IEEE Trans. Haptics*, vol. 13, no. 3, pp. 562–570, Jul.–Sep. 2020.

[17] R. F. Friesen, R. L. Klatzky, M. A. Peshkin, and J. E. Colgate, "Building a navigable fine texture design space," *IEEE Trans. Haptics*, vol. 14, no. 4, pp. 897–906, Oct.–Dec. 2021.

[18] W. M. Bergmann Tiest and A. M. L. Kappers, "Analysis of haptic perception of materials by multidimensional scaling and physical measurements of roughness and compressibility," *Acta Psychologica*, vol. 121, pp. 1–20, 2006.

[19] T. Yoshioka, S. J. Bensmaia, J. C. Craig, and S. S. Hsiao, "Texture perception through direct and indirect touch: An analysis of perceptual space for tactile textures in two modes of exploration," *Somatosensory Motor Res.*, vol. 24, no. 1-2, pp. 53–70, 2007.

[20] L. Skedung, K. L. Harris, E. S. Collier, and M. W. Rutland, "The finishing touches: The role of friction and roughness in haptic perception of surface coatings," *Exp. Brain Res.*, vol. 238, no. 238, pp. 1511–1524, 2020.

[21] Y. Vardar, C. Wallraven, and K. J. Kuchenbecker, "Fingertip interaction metrics correlate with visual and haptic perception of real surfaces," in *Proc. IEEE World Haptics Conf. (WHC)*, 2019, pp. 395–400.

[22] T. Callier, H. P. Saal, E. C. Davis-Berg, and S. J. Bensmaia, "Kinematics of unconstrained tactile texture exploration," *J. Neuriophysiol.*, vol. 113, no. 7, pp. 3013–3020, 2015.

[23] K. Priyadarshini, S. Chaudhuri, and S. Chaudhuri, "PerceptNet: Learning perceptual similarity of haptic textures in presence of unorderable triplets," in *Proc. IEEE World Haptics Conf. (WHC)*, 2019, pp. 163–168.

[24] H. Culbertson, J. J. López 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.*, Houston, Texas, USA, 2014, pp. 319–325.

[25] A. Klocker, M. Wiertlewski, V. Theate, V. Hayward, and J. L. Thonnard, "Physical factors influencing pleasant touch during tactile exploration," *PLoS One*, vol. 8, no. 11, pp. 1–8, 2013.

[26] M. Arvidsson, L. Ringstad, L. Skedung, and K. Duvefelt, "Feeling fine - the effect of topography and friction on perceived roughness and slipperiness," *Biotribol.*, vol. 11, pp. 92–101, 2017.

[27] H. Culbertson and K. J. Kuchenbecker, "Ungrounded haptic augmented reality system for displaying roughness and friction," *IEEE/ASME Trans. Mechatron.*, vol. 22, no. 4, pp. 1839–1849, Aug. 2017.

[28] A. Bicchi, K. Salisbury, and L. Brock, "Contact sensing from force measurements," *Int. J. Robot. Res.*, vol. 12, no. 3, pp. 249–262, 1993.

[29] D. Picard, C. Dacremont, D. Valentin, and A. Giboreau, "Perceptual dimensions of tactile textures," *Acta Psychologica*, vol. 114, no. 2, pp. 165–184, 2003.

[30] G. A. Gescheider, S. J. Bolanowski, T. G. Greenfield, and K. E. Brunette , "Perception of the tactile texture of raised-dot patterns: A multidimensional analysis," *Somatosensory Motor Res.*, vol. 22, no. 3, pp. 127–140, 2005.

[31] N. Landin, J. M. Romano, W. McMahan, and K. J. Kuchenbecker, "Dimensional reduction of high-frequency accelerations for haptic rendering," in Haptics: Generating and Perceiving Tangible Sensations: Part II (Proceedings of EuroHaptics), *(Lecture Notes in Computer Science Series)*, Berlin, Germany,2010, vol. 6192, pp. 79–86.

[32] R. H. LaMotte, "Softness discrimination with a tool," *J. Neuriophysiol.*, vol. 83, no. 4, pp. 1777–1786, 2000.

[33] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin, Germany: Springer2008.

[34] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde, "Sliced wasserstein auto-encoders," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019, pp. 1–11.

[35] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, vol. 26, pp. 2292–2300.

[36] C. Frogner, F. Mirzazadeh, and J. Solomon, "Learning embeddings into entropic wasserstein spaces," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019, pp. 1–12.

[37] M. Blondel, O. Teboul, Q. Berthet, and J. Djolonga, "Fast differentiable sorting and ranking," in *Proc. Int. Conf. Mach.*2020, pp. 950–959.

[38] W. M. Bergmann Tiest and A. M. L. Kappers, "Haptic and visual perception of roughness," *Acta Psychologica*, vol. 124, no. 2, pp. 177–189, 2007.

[39] R. W. Fleming and K. R. Storrs, "Learning to see stuff," *Curr. Opin. Behav. Sci.*, vol. 30, pp. 100–108, 2019.

[40] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vis.*, vol. 40, pp. 49–70, 2000.

[41] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[42] A. Metzger and M. Toscani, "Unsupervised learning of haptic material properties," *eLife*, vol. 11, 2022, Art. no. e64876.

[43] H.-N. Ho and L. Jones, "Contribution of thermal cues to material discrimination and localization," *Percep. Psychophys.*, vol. 68, no. 1, pp. 118–128, 2006.

[44] G. Serhat and K. J. Kuchenbecker, "Free and forced vibration modes of the human fingertip," *Appl. Sci.*, vol. 11, no. 12, 2021, Art. no. 5709.

[45] J. Z. Wu, K. Krajnak, D. E. Welcome, and R. G. Dong, "Analysis of the dynamic strains in a fingertip exposed to vibrations: Correlation to the mechanical stimuli on mechanoreceptors," *J. Biomech.*, vol. 39, no. 13, pp. 2445–2456, 2006.

[46] M. Wiertlewski, J. Lozada, and V. Hayward, "The spatial spectrum of tangential skin displacement can encode tactual texture," *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 461–472, Jun. 2011.

Benjamin A. Richardson received the B.S. degree in materials science and engineering and the M.S. degree in mechanical engineering from Northwestern University, Evanston, IL, USA, in 2013 and 2015, respectively. He is currently working toward the Ph.D. degree in computer science with the University of Stuttgart, Stuttgart, Germany. He is with the Haptic Intelligence Department, Max Planck Institute for Intelligent Systems, Stuttgart, Germany. His research interests include human and robot perception of haptic object properties.

Yasemin Vardar (Member, IEEE) received the Ph.D. degree in mechanical engineering from Koç University, Istanbul, Turkey, in 2018. She is currently an Assistant Professor with the Delft University of Technology, Delft, The Netherlands. She completed Postdoctoral Research with the Max Planck Institute for Intelligent Systems, Stuttgart, Germany, till 2020. Her research interests include human tactile perception and haptic interfaces. She was the recipient of the 2021 NWO VENI Grant, 2018 Eurohaptics Best Ph.D. Thesis Award, IEEE WHC 2017 Best Poster Presentation Award, and TUBITAK Ph.D. Fellowship, she was selected for 2019 Sign Up! Career-building Program. She is currently the Co-Chair of the Technical Committee on Haptics.

Christian Wallraven (Member, IEEE) received the Ph.D. degree in physics for work on a perceptually motivated computer vision system from the Max Planck Institute for Biological Cybernetics, Tübingen, Germany. He then moved to Korea University, Seoul, South Korea, where he is currently a Full Professor and the Head of the Cognitive Systems Laboratory with research focusing on multisensory information integration in the brain with a special focus on vision and touch, social face processing in humans and machines, understanding decision-making processes, and interfacing artificial with human intelligence. He has coauthored more than 200 publications with an interdisciplinary approach integrating artificial intelligence, neuroscience, and immersive computer graphics. In 2021, he co-chaired the 6th Asian Conference on Pattern Recognition.

Katherine J. Kuchenbecker (Fellow, IEEE) received the Ph.D. degree in mechanical engineering from Stanford University, Stanford, CA, USA, in 2006. She is currently the Director with the Max Planck Institute for Intelligent Systems, Stuttgart, Germany. She completed Postdoctoral Research with the Johns Hopkins University, Baltimore, MD, USA, and was a Faculty Member with the GRASP Laboratory, University of Pennsylvania, Philadelphia, PA, USA, from 2007 to 2016. Her research interests include robotics and human-computer interaction with focuses on haptics, teleoperation, physical human-robot interaction, tactile sensing, and medical applications. She delivered a TEDYouth talk on haptics in 2012 and was honored with a 2009 NSF CAREER Award, the 2012 IEEE RAS Academic Early Career Award, a 2014 Penn Lindback Award for Distinguished Teaching, elevation to IEEE Fellow in 2022, and various best paper, poster, demonstration, and reviewer awards. She co-chaired the IEEE RAS Technical Committee on Haptics from 2015 to 2017 and the IEEE Haptics Symposium in 2016 and 2018.