Multimodal Feature-based Surface Material Classification

Matti Strese, *Member, IEEE,* Clemens Schuwerk, *Member, IEEE,* Albert Iepure, and Eckehard Steinbach, *Fellow, IEEE*

Abstract—When a tool is tapped on or dragged over an object surface, vibrations are induced in the tool, which can be captured using acceleration sensors. The tool-surface interaction additionally creates audible sound waves, which can be recorded using microphones. Features extracted from camera images provide additional information about the surfaces. We present an approach for tool-mediated surface classification that combines these signals and demonstrate that the proposed method is robust against variable scan-time parameters. We examine freehand recordings of 69 textured surfaces recorded by different users and propose a classification system that uses perception-related features, such as hardness, roughness, and friction; selected features adapted from speech recognition, such as modified cepstral coefficients applied to our acceleration signals; and surface texture-related image features. We focus on mitigating the effect of variable contact force and exploration velocity conditions on these features as a prerequisite for a robust machine-learning-based approach for surface classification. The proposed system works without explicit scan force and velocity measurements. Experimental results show that our proposed approach allows for successful classification of textured surfaces under variable freehand movement conditions, exerted by different human operators. The proposed subset of six features, selected from the described sound, image, friction force and acceleration features, leads to a classification accuracy of 74% in our experiments when combined with a Naive Bayes classifier.

Index Terms—Tool-mediated surface classification, feature-based surface material recognition, multimodal surface classification

1 Introduction

Stroking a rigid tool over the surface of an object or tapping onto the object surface induces accelerations in the tool. These vibrations are measured with an acceleration sensor, and the corresponding surface signals represent the surface characteristics. In current technical systems, these signals can be used to recreate the feel of real surfaces using voice coil actuators [1], [2], [3], to enhance traditional teleoperation systems [4] and to recognize surfaces using robots [5], [6], [7], [8] or from human freehand movements [9].

In this work, we go beyond accelerometer-based signal acquisition and envision a hand-held surface classification system using multiple modalities, such as acceleration, image and sound signals, inspired by human surface classification capabilities. Possible real-life scenarios comprise the usage of smartphones or custom made sensorized and low-cost tools to characterize surfaces and to transmit features to a remote receiver for redisplay and recreation. A human user may wield, e.g., her/his smartphone to scan a surface, extract a set of features and transmit these over a network. At the destination, the features can be used to locally recreate the feel using novel tactile output devices with low bandwidth demands. To make such an application possible, we would like to prove in this work that features from different signal domains can be used to classify among surfaces without explicit scan force or velocity measurements.

Another use case of such a classification system can be found in the remote perception of surfaces using teleoperation setups. In current scenarios, teleoperating systems convey remote tactile impressions in almost real-time, but have no memory about, e.g., the previously touched object surfaces. A surface classification system can be used to build up a haptic scene description of the remote environment. The unknown remote surfaces can be matched to known surfaces from a tactile database, which are linked to tactile models of the surfaces. For realistic recreation, these models have to match the human perception of surfaces.

• *All authors are with the Chair of Media Technology, Technical University of Munich, Munich, Germany, 80333.*
E-mail:matti.strese@tum.de, clemens.schuwerk@tum.de, albert.iepure@tum.de, eckehard.steinbach@tum.de

1.1 Challenges

When a human strokes a rigid tool over an object surface, the applied force and the scan velocity typically vary during the surface exploration and between subsequent exploration sessions. These scan-time parameters strongly affect the acquired acceleration signals [1] [3] and how humans perceive artificially recreated surfaces [10] using voice coil actuators. As a consequence, acceleration-based features for a robust surface classification system have to mitigate the dependency of these two parameters. Related studies in the area of robotic texture recognition, reviewed in Section II, use well-defined exploration trajectories or controlled scan-time parameters to overcome these challenges. With humans performing the surface explorations, such well-defined scan-time parameters, however, are not feasible.

Similar observations were made with regard to tool-surface interaction-based sound signals, which are closely related to the acceleration signals [4]. Changes in the scan velocity and force influence the perceived pitch, timbre and the loudness.

The dynamic friction force between the tool and a surface also depends on the scan-force and scan-velocity and offers an additional dimension, namely friction (see [11]) that cannot be captured by using only acceleration sensors. If force sensors, e.g., force sensing resistors (FSR), are laterally applied to measure such

friction data, the different human operator gripping forces applied while holding the tool constitute another challenge.

In addition to acceleration, friction force and sound data, images of the object surfaces can be captured using a camera. While surface images carry important information about the surface, differences in distance, rotation, lighting, and focus conditions complicate the extraction of robust image-based features for surface classification.

We focus on the design of features extracted from acceleration, sound, friction force and image signals that are robust against these scan-time parameters. The proposed features are used in combination with various machine learning (ML) algorithms to perform surface material classification. Note that, e.g., in a smartphone the scan force or velocity may not be measured at all without mounting external sensors.

1.2 Contributions

In this study, we extend the tool-mediated surface classification system described in [9], which performs the classification using acceleration signals recorded during human freehand movement only. The following are all the novel contributions of this work.

- We created a novel surface classification system consisting of a recording tool, acceleration, sound, friction force and image-based features and a classification pipeline.
- We recorded acceleration, sound and friction force signals and captured images of 69 surfaces as an extension to our work in [9], which are publicly available at our website *http://www.lmt.ei.tum.de/texture/*. The training and testing dataset are used as input for three machine learning approaches and have been recorded by different human operators to provide variant tool-mediated surface exploration data.
- We performed a visual-haptic subjective experiment to evaluate the visual and tactile roughness of our surface material database. The outcome is used to define an image-based roughness feature and to evaluate the visual-haptic mismatch in terms of perceived roughness. The outcome of this experiment demonstrates that human subjects classify surfaces differently, when asked to judge about surface roughness either purely visually or by direct touch.

The rest of the paper has the following structure. In Section 2, we review relevant work in the area of surface texture recognition and classification. Our surface recording setup is introduced in Section 3. The proposed features are described in Section 4 and evaluated in Section 5.

2 RELATED WORK

Acceleration signals have been used in robotic texture recognition. In [12], a biomimetic sensor emulating the human finger, mounted on an industrial robot, was used to scan surface textures and provide acceleration data for the machine classification of materials achieving 95% accuracy. The dominant five peaks of the spectral envelope were used to form a feature vector. It is known that texture surface signals change mainly with the applied force and scan velocity [3], [13]. Therefore, the robot needs to follow well-defined trajectories and uses only the signals acquired at constant velocity for training and classification. Hence, extensive training with various predefined trajectories is necessary. The force applied to the surface during the recordings is explicitly controlled to be constant. Note that industrial robots can repeat different movement trajectories very precisely, normally are equipped with various sensors, and hence are able to measure the scan-time parameters at run times.

In another related study [8], the applied force and scan velocity combinations were explicitly measured and used as features. In addition, the energies of 30 spectral bins served as features in the surface classification. This approach also suffers from the additional sensor requirements and from the shortcomings of explicit training with a range of different scan parameters. For a database comprising 15 different textures, the system achieved a classification accuracy of 72% for a single exploration using a multi-class Support Vector Machine (SVM).

In [6], the fundamental frequency of the acceleration signal was used as a feature representing a texture. In addition to controlled force and velocity conditions during scanning, the surface profiles were constrained to be well-defined (periodic) sinusoidal gratings, for which the fundamental frequency was expected to change almost linearly with the scan velocity, and thereby aid in the surface classification.

The BioTac sensor, a multimodal tactile sensor that resembles the human fingertip and measures contact forces, vibrations and temperature was mounted on a linear stage to stroke the textures in [7]. A classification rate of 95.4% among 117 textures was reported. Features such as *roughness*, *fineness* and *traction* are identified from the literature on human perception and modeled analytically. Although the proposed features describe well-defined scan parameters, they still show high variance in the applied force and scan velocity [7].

The interaction of a tool tip with a surface does not only generate acceleration signals, but also audible sound waves. This fact is used in [14] to classify textures according to audio features, or objects as in [15] using Hidden Markov Models.

Besides the direct touch impressions, textured surfaces reveal visual clues to humans in the classification of surfaces. Hence, textural features, which approximate the human visual surface perception, are of interest. In [16], a series of psychological measurements were performed and six basic texture features, namely, coarseness, contrast, directionality, line-likeness, regularity, and roughness, were determined. These features have been approximated to obtain a computational form that can be used by a machine to classify and select different textures.

Haralick et al. [17] developed a computational method for a set of 14 features based on image properties such as energy content or local homogeneity. The image-based approaches in [16], [17] have not been used in combination with acceleration and sound data to the best of our knowledge. Using only images for the classification of textured surfaces is not able to robustly capture surface characteristics such as hardness or friction. We believe that image- and acceleration-based features need to be combined to overcome this, just as humans would perform this task.

The works in [18] and [19] use a deep learning network architecture for object recognition and surface classification. Image and acceleration-based signals are used to learn a model, which opens an alternative approach for the task of surface classification compared to hand-crafted feature definition.

In summary, the reviewed approaches [5], [6], [7], [8], [12] in general use expensive sensors and require constant scan parameters (force and velocity) during the acceleration recordings or the scan parameters are explicitly measured and used to match the texture signal to a database that comprises various scan force and

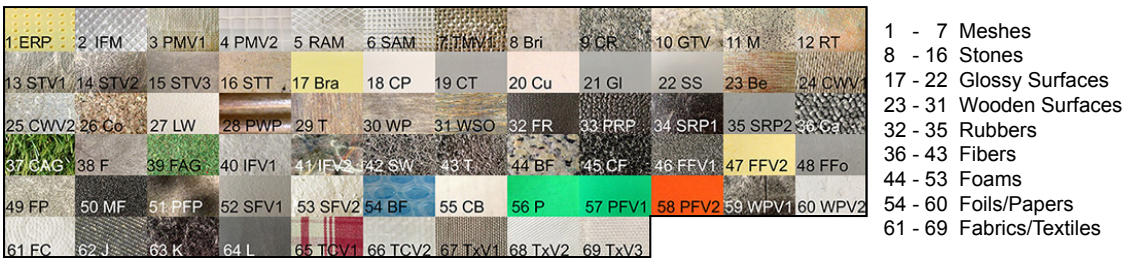

Fig. 1: The materials included in our haptic surface database. The acceleration, sound and friction force signals and magnified images are freely accessible at *http://www.lmt.ei.tum.de/texture/*. We use everyday surface materials and categorize them in common categories such as stones, fabrics and wooden surfaces. The corresponding surface names are listed in Table 2 in Appendix A (see supplementary material).

velocity combinations. The deep learning approaches in [18], [19] propose a promising alternative to the feature definition procedure, despite the higher computational complexity, and may be used in future work alongside with handcrafted features for optimal surface classification, if the device capabilities allow for the higher computational complexity.

In contrast, we use an inexpensive acceleration sensor, a common microphone and a smartphone camera to classify textured surfaces during human freehand recordings. Additionally, two low-cost force-sensing resistors are used with the same electronics as the acceleration sensor for measuring the interaction friction forces. Besides human freehand surface exploration, such robust multimodal features can potentially improve the performance of robotic texture recognition systems because the well-controlled surface exploration and the extensive training of the classifiers can be avoided.

3 RECORDING PROCEDURE

3.1 Surfaces

We recorded acceleration, friction force and sound signals and captured surface images from 69 different surfaces, which are shown in Fig. 1. We use everyday materials, such as stones or fabrics, and give a unique label to each surface. We decided not to rely on the publicly available and extensive haptic database in [20], as it contains only acceleration data recorded while a tool moved across different surfaces. Besides the friction and image data, sound signals as well as initial tool-surface impact data is required in our multimodal classification approach.

3.2 Recording tool and camera

The recording tool, denoted as haptic stylus, is a free-to-wield object with a stainless steel tool tip (diameter 5.6 mm), an acceleration sensor, a microphone and two FSRs as shown in Fig. 2. We use the FSRs to record laterally applied forces to the device while dragging it over a surface. In this study, the sampling frequency of the recorded acceleration and friction data is 10 kHz. We use a three-axis LIS344ALH acceleration sensor (ST Electronics) with a range of ± 6 g and combine all three axis to one using the DFT321 (see [21]). We use this approach, which preserves the spectral characteristics of the recorded acceleration signals, to reduce the acceleration data dependency of the device inclination [1] toward the surface and to process only one signal trace. Single-axis acceleration signals would suffer from the fact that the device is not constantly held perpendicular toward the surface during human freehand exploration of surfaces.

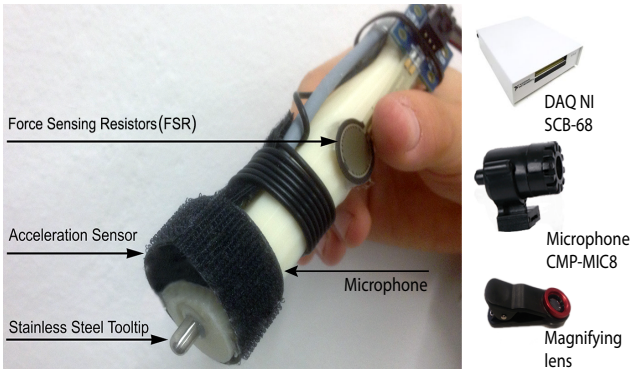

Fig. 2: The haptic stylus (left) used in the textured surface analysis. A microphone (middle right) is placed inside the body of this device to record sound signals during the exploratory movement, while the LIS344ALH acceleration sensor gathers acceleration signals using a DAQ (upper right). Simultaneously, two FSR are used to measure the lateral applied forces to deduce the dynamic friction properties of the material. A magnifying lens (lower right), mounted on an iPhone 6 (not depicted), is used to enhance textural structures on the surfaces.

We place a CMP-MIC8 microphone (Koenig Electronics) inside the body of the custom haptic stylus (see Fig. 2) to capture sound waves generated during the surface-tool interaction at a sampling rate of 44 100 Hz. These sound signals are acquired in parallel to the acceleration signals. Acceleration and sound signals convey information about the surface-tool interaction in different frequency ranges and, hence, we consider both as relevant for surface classification.

The images of the surfaces, which we use as an additional modality for surface classification in this paper, were taken by the rear camera of a common smartphone (Apple iPhone 6) that has a resolution of 8 MP. We applied a smartphone magnifying lens (QUMOX macro lens, 10x magnification, see Fig. 2) to enhance the textural properties of the surfaces. The capturing conditions differ for each image of the same surface. For example, the subset of magnified surface images shown in Fig. 4 were captured under different illumination, focus and camera distance conditions. Note that, e.g., a web cam can be mounted on top of the haptic stylus to capture images, meaning that our image acquisition is not bound to a smartphone camera.

3.3 Recording Procedure

This work focuses on surface classification of human freehand exploration recordings. We recorded different data for the feature design (Section 3.3.1), training procedure (Section 3.3.2) and

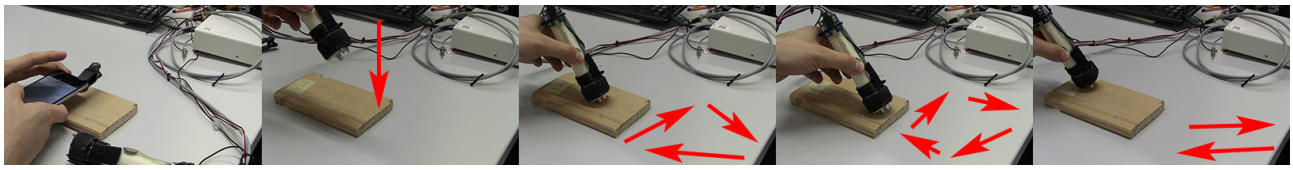

Fig. 3: Recording procedure, exemplary shown for an operator. First, a surface image is captured using the camera preview. Secondly, the recording device impacts the surface. Thereafter, an operator may arbitrarily move the tool over the surface.

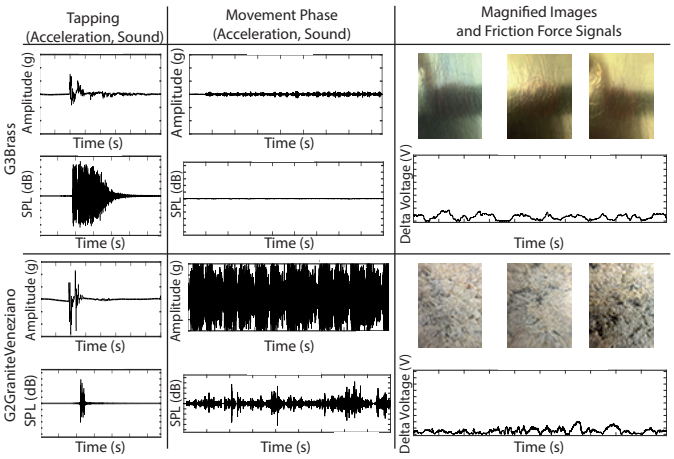

Fig. 4: Examples of the image, friction force, sound and acceleration signals for two surfaces (brass and granite). SPL stands for sound pressure level.

evaluation of the trained model (Section 3.3.3). We evaluate our proposed classification system using recordings from different human operators. All images, regardless of belonging to the training or testing data set, were captured in a similar way. Using the camera preview and auto-focus, all operators have been given the task of taking a full-screen and not blurred picture of the surface. During this procedure, the distance, perspective, rotation and resulting light reflections differ among the different operator recordings.

3.3.1 Robot-Controlled Recordings

We use robot-controlled recordings from our preliminary work (Phantom Omni device with PID controller, see [22]) to assess the feature robustness with respect to different scan forces and scan velocities. In [22], 43 surfaces have been recorded under linearly increasing scan forces and scan velocities. Exemplary acceleration signals are shown in Fig. 26 in the appendix (see supplementary material) to illustrate the impact of both parameters on the acceleration signal over time. However, certain signal properties, e.g., the signal spikiness, show some level of invariance and are hence characteristic for the surface. These controlled conditions were realized using a custom-made rotatory plate and a controllable Phantom Omni device. We use these controlled recordings to tune the parameters during the feature design procedure and to increase the robustness of the proposed and adapted features. Additionally, we recorded tapping data (linearly increasing hand velocity) for the derivation of our hardness feature. The tapping data comprise a series of impact impulses of different strength, which are created by increasing the tool velocity toward the surface. Note that these recordings are not used for the proposed surface classification pipeline in this work, but for the feature definition only.

3.3.2 Training Data

The first author of this work recorded ten 25-s-long freehand signals for each surface as training data by wielding the tool manually. Each dataset contains acceleration, sound and friction force signals as well as images. According to [23], tool-mediated recordings typically cover a scan force range of $0-3$ N and a scan velocity range of $0-400$ mm/s. We do not measure these two parameters for our proposed surface classification approach. We deliberately applied different scan velocities and scan forces and movement patterns during each recording to also ensure intra-class variance for each surface. For completeness, we uploaded all signal plots to our website[1] for further inspection. Also, Fig. 3 illustrates the freehand recording procedure in a sequence of images.

3.3.3 Testing Data

We asked ten subjects, which are not authors of this work, to record 15 seconds of acceleration, sound, friction force signals and images (one per person), which we use as an independent testing data set for the evaluation of our features. The subjects were instructed to start after a visual signal occurred and to freely explore each given surface for ten seconds. The subjects were not instructed to follow any specific pattern. They were also instructed to capture one surface image by choosing an arbitrary recording distance and angle, which should fill the camera preview and should not be blurred.

4 FEATURES FOR SURFACE CLASSIFICATION

This section details the features used for surface classification. Note that our feature definition abbreviations follow a specific spelling convention. The first letter indicates whether the feature originates from acceleration (A), friction (F), image (I) or sound (S) signals. The second letter indicates the phase, Impact (I) or Movement (M), from which the feature is calculated. Since these phases do not exist in the recording procedure of the images, this letter is omitted in the definition of the image features.

4.1 Acceleration Features

We first introduce a modified version of the Mel Frequency Cepstral Coefficients (MFCC) in Section 4.1.1, which is widely used in speech recognition and adapt it to our scenario of surface material classification. Afterwards, we propose a set of acceleration-related features (Sections 4.1.2 - 4.1.4) that describe the hardness and roughness of surfaces.

1. http://www.lmt.ei.tum.de/texture/

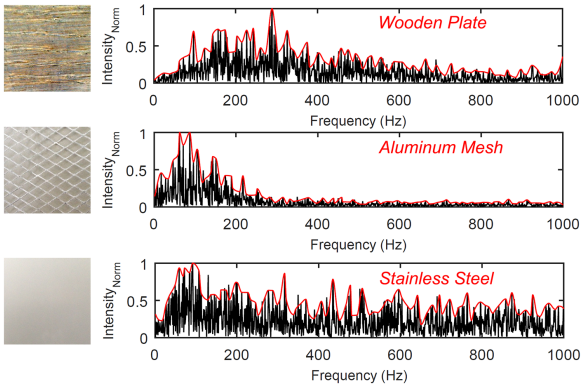

Fig. 5: Spectral plots for different surfaces during the acceleration movement phase. The acceleration spectra of the surfaces reveal characteristic shapes, especially with regard to their spectral envelope.

4.1.1 Modified Frequency Cepstral Coefficients

The field of audio signal processing offers a variety of methods that can be used in the analysis of acceleration signals as stated in [8]. Figure 5 shows that the spectra of acceleration signals and, especially the spectral envelopes, are characteristic for various surfaces. Conceptually, Mel-Frequency Cepstral Coefficients (MFCC) are known to capture the spectral envelope of a given data frame by calculating the discrete cosine transform (DCT) of the sound signal spectra. We adapt the MFCC implementation from [24] and modify it for our acceleration signals. Figure 6 shows the processing pipeline for a 25-ms-long acceleration signal segment, which is extracted from the acceleration movement phase. Most importantly, the applied filter bank has to be adjusted to the acceleration sensor bandwidth. In addition, the audio specific processing steps such as the alpha value (pre-intensification of the sound signal) and the liftering procedure (cepstral filtering) [24] are removed. We use the Fourier Transform to get the spectrum of the acceleration signal, calculate the corresponding filter bank entries, use the logarithm to decrease the dynamic range and calculate the cepstrum using the DCT. The first entry denotes the constant component of the signal, so only the entries 2 to 14 are used. This calculation is repeated and averaged for overlapping data frames (10 ms overlap) to obtain 13 mean cepstral coefficients and to define them as our acceleration movement cepstral coefficients feature vector (AMCC). It has been argued that the exponential decay in the spectral amplitudes is directly related to tactile perception [8]. We tested a filter bank with constant as well as exponentially decaying spectral amplitudes.

4.1.2 Hardness Features

According to [11], hardness is considered as a fundamental dimension for the task of surface classification. The acceleration signals recorded during the initial contact of the tool with the surface contain information about the surface hardness and stiffness properties. Motivated by [25], we use the initial high-frequency transient response that results from the tapping on a surface for our hardness definition. We observed that the acceleration values of these initial impacts generally are large and reveal a strongly increasing slope if the surface is not deformable, whereas softer materials deform during an impact and show a weaker inclination from zero to their maximum.

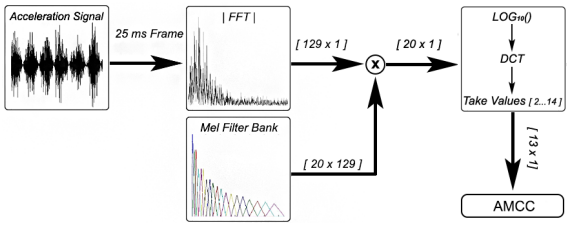

Fig. 6: AMCC feature vector calculation of a 25-ms-long surface signal. The Mel filterbank uses 20 exponentially decaying filters for the cepstral analysis. We extract 13 coefficients for a frame and average these over the entire acceleration signal and use these values as AMCC feature.

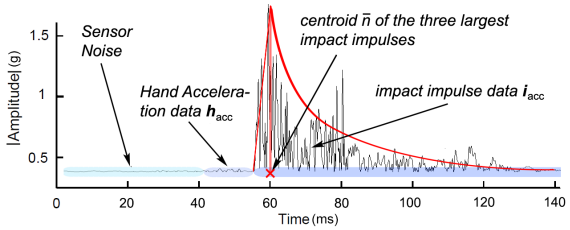

Fig. 7: Absolute value impact impulse data with visible sections of sensor noise, hand acceleration, impact and decay. We use the sensor noise to define a threshold for the impact detection, e.g., 20 times the maximum noise amplitude. If an impulse is detected, a fixed amount of data before and after this impulse is extracted as impact data.

4.1.2.1 Hardness: We extract the hand acceleration data $\mathbf{h}_{Acc} = [|h_{Acc_1}|, \ldots, |h_{Acc_n}|]^T$ and impact impulse data $\mathbf{i}_{Acc} = [|i_{Acc_1}|, \ldots, |i_{Acc_m}|]^T$ shown in Fig. 7 from each segmented acceleration tapping signal. We only use the low-frequency components (1 Hz - 12 Hz) of $\mathbf{h}_{Acc}$, denoted as $\hat{\mathbf{h}}_{Acc}$, since the human hand movement capability is limited to this range [26].

Then, we extract the indexes of the three largest impact impulses $n^{\max}_{1\ldots3}$ in $\mathbf{i}_{Acc}$ and calculate their temporal centroid $\bar{n}$ relative to the start of the impact. Subsequently, we define our acceleration impact hardness feature (AIH) as

$$AIH = \frac{\max(\mathbf{i}_{Acc})}{\bar{n}} \cdot \frac{1}{\sum_{i=1}^{n} \hat{h}_{Acc_i}} \tag{1}$$

The rising slope (first term in (1)) is calculated by dividing the absolute maximum in the tapping data by the index centroid of the three largest tapping impulses. The hand velocity is approximated as the sum of all absolute data values in the hand acceleration data vector $\mathbf{h}_{Acc}$ (denominator of the second term in (1)). Figure 8 shows the proposed hardness feature applied for controlled tapping signals (from Section 3.3.1) of three different materials with increasing impact velocity. It can be seen from Fig. 8 that the proposed hardness feature shows limited intra-surface, but large inter-surface variation for increasing impact velocity.

4.1.2.2 Impact spectral centroid: We identified the spectral centroid (SC) (see [7]) of the impact impulse data as inversely related to the stiffness of a material. If $\mathbf{I}(f_k)$ denotes the discrete cosine transform (DCT, length $m = 4096$ in this work) value of the tapping data $\mathbf{i}$ at frequency f_k, we calculate the stiffness-related acceleration impact SC feature (AISC) as

$$AISC = \frac{\sum_{k=1}^{\frac{m}{2}} \mathbf{I}(f_k)^2 \cdot f_k}{\sum_{k=1}^{\frac{m}{2}} \mathbf{I}(f_k)^2} \tag{2}$$

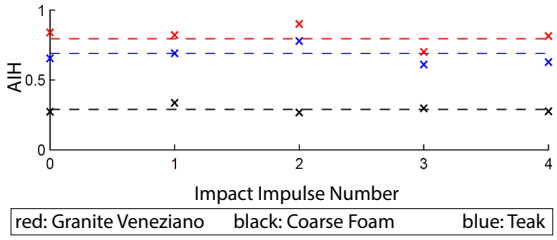

Fig. 8: Impact hardness (AIH) for different materials. The impact velocity increases from left to right.

Softer materials apparently behave as low-pass filters and show lower stiffness (smaller SC values). In contrast, stiff surfaces like stones and metals have high clamping spectral components regardless of the impact velocity and the resulting variable spectral energy of the impact impulse.

4.1.3 Microscopic Roughness Features

The microscopic texture roughness feature proposed in [7] depends on the variations of the scan-time parameters. To improve its robustness against these variations, we first apply an FIR high-pass filter to each signal with a cutoff frequency of 100 Hz. We then compute the wavelet transformation (Coiflet 3) and extract the detail levels $\mathbf{d}_1 = [|d_{1_1}|,\ldots,|d_{1_n}|]^T$ and $\mathbf{d}_5 = [|d_{5_1}|,\ldots,|d_{5_n}|]^T$ for further processing. We assume that rougher surfaces have stronger differences in these detail levels and calculate the acceleration movement temporal roughness (AMTR) as

$$AMTR = \log_{10}(\overline{\mathbf{d}}) \quad with \quad \mathbf{d} = \mathbf{d}_1 - \frac{\overline{\mathbf{d}}_1}{\overline{\mathbf{d}}_5} \cdot \mathbf{d}_5 \tag{3}$$

with $\overline{\mathbf{d}}$ being the arithmetic mean of vector $\mathbf{d}$. Both detail levels are affected by changes in the scan-time parameters; thus, this ratio mitigates the effect of variable velocity and force conditions. Figure 9 shows the result for three different surfaces. In Appendix A, Fig. 24 and Fig. 25 (see supplementary material), we compare the roughness definition from [7] to our AMTR definition on our controlled database recordings. Figure 10 shows the boxplot of our AMTR feature for all 69 surfaces in our database.

Figure 11 compares the definition in [7] and the proposed roughness feature from equation (3). Note that this is only an example for one surface for visualization. Figure 12 shows that this new roughness definition can be used in the classification because different surfaces generate roughness values which are approximately scan velocity robust, but show large differences for different materials.

We also consider a spectral interpretation of surface roughness, where we assume that rapid changes in the spectral domain indicate larger roughness of a textured surface. For the acceleration movement data $\mathbf{x}$, we define a data frame with a fixed size of 5000 data points as $\mathbf{x}_n = [x_n,\ldots,x_{n+5000}]^T$ and its DCT as $\mathbf{X}_n$, where we take the absolute value of each DCT coefficient. We calculate the difference spectrum $\mathbf{D}_k = \mathbf{X}_n - \mathbf{X}_{n+100}$ of each frame and advance in steps of 100 data points until the end of $\mathbf{x}$ is reached. This leads to K difference spectra $\mathbf{D}_k$. Our acceleration movement spectral roughness feature (AMSR) is then defined as

$$AMSR = \log_{10}\left(\sum_{k=1}^{K} \mathbf{D}_k^2\right) \tag{4}$$

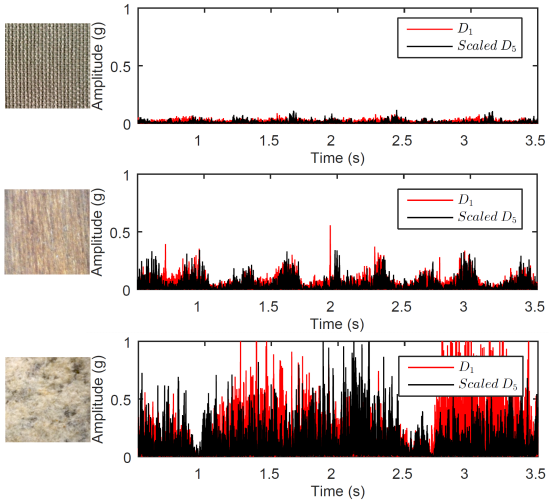

Fig. 9: Illustration of the AMTR feature. We calculate two wavelet detail levels for each acceleration signal, scale them to similar mean amplitude and calculate a difference vector. Rough surfaces have major differences in both detail levels and, hence, a larger mean absolute difference value, which we use as temporal surface roughness in this work. The scaling strongly decreases the effect of different scan forces and velocities, because both detail levels are affected by these variations.

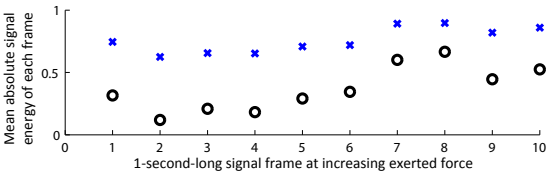

Fig. 11: Exemplary roughness comparison of an acceleration signal (roof tile) for linearly increasing scan force. The definition of roughness from [7] (black circles) shows higher variance compared to our roughness feature (blue crosses) in (3). This figure results from our controlled recordings and is used to demonstrate for one example that the AMTR roughness definition is more robust compared to the roughness definition from [7].

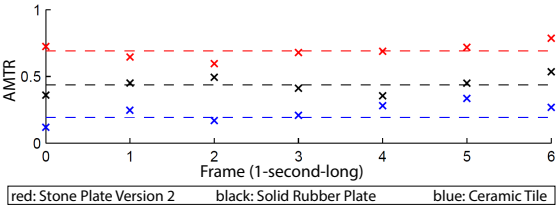

Fig. 12: Temporal roughness feature for different surfaces, tested on movement phase data under linearly increasing scan force.

4.1.4 Macroscopic Roughness Features

Macroscopic roughness comprises the existence of surface grating wavelengths above 1 mm [11] and, thus, is presumably represented by the low-frequency content of a surface acceleration signal. Unlike [8], we do not remove spectral components below 10 Hz that are supposed to contain hand movement frequency content. Although we do not explicitly model the human hand, it is still part of the tool-mediated textured surface interaction and can contain surface specific reactions with regard to the surface exploration. We observed, e.g., that coarse structures like stones or meshes

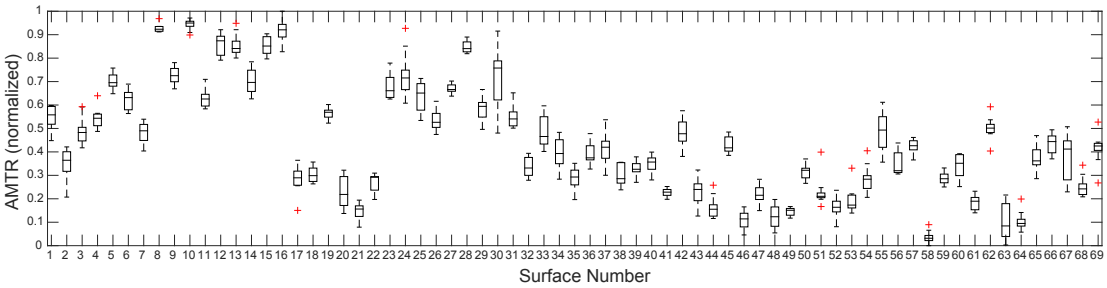

Fig. 10: AMTR Boxplot of all 69 textured surfaces, which shows high inter-class variance and low intra-class variance of our AMTR feature. The red crosses show extreme outliers.

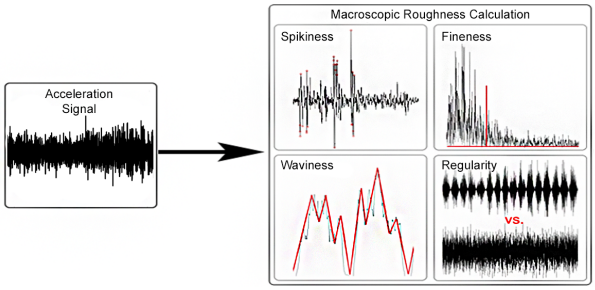

Fig. 13: Macroscopic roughness feature extraction block diagram. All features are described in Section 4.1.4.

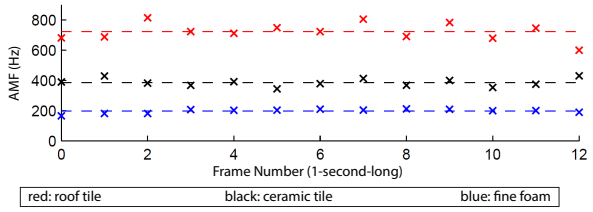

Fig. 14: Fineness of different surfaces during the movement phase at increasing movement velocity (from left to right).

include valuable information in these low-frequency vibrations. We identified four possible ways to capture the macroscopic roughness of a surface, which are shown in Fig. 13 and explained in the following.

4.1.4.1 Waviness: Waviness describes the deviation of the low-frequency signal slopes in the time domain. First, we divide our low-pass filtered (100 Hz cutoff frequency) movement data vector $\mathbf{x}_{filt}$ in frames and calculate the mean absolute value and store it in vector $\mathbf{m}$ for each frame. We choose 200 data points, in this work equivalent to 20 ms, as the frame size following the observation in [8] that the scan force and velocity variation are almost negligible in this range. For mitigating the effect of increasing scan force or velocity, we also calculate the Simple Moving Average (SMA) $\bar{\mathbf{s}}_{100}$ of 100 data points of $\mathbf{x}_{filt}$. Our acceleration movement waviness feature (AMWV) is calculated as

$$AMWV = \log_{10}(\sigma(\mathbf{m} - \bar{\mathbf{s}}_{100})) \tag{5}$$

where σ denotes the standard deviation.

4.1.4.2 Spikiness: Spikes in otherwise smooth acceleration time-domain signals are a strong indicator for bumpy surfaces or meshes on which the recording device gets occasionally stuck. We apply a custom spike detection algorithm on the low-pass-filtered signals $\mathbf{x}$. With the SMA $\bar{\mathbf{x}}_{5000}$ (5000 data points) we calculate the threshold vector $\mathbf{x}_{th} = 2 * \sigma(\mathbf{x}) + \bar{\mathbf{x}} + \bar{\mathbf{x}}_{5000}$ and the difference vector $\mathbf{x}_{\Delta} = \mathbf{x} - \mathbf{x}_{th}$. All values smaller than zero are set to zero and we calculate our acceleration movement spikiness (AMSP) feature as

$$AMSP = \log_{10}(\bar{\mathbf{x}}_{\Delta}) \tag{6}$$

4.1.4.3 Fineness: In [7], the Spectral Centroid (SC) is introduced as a feature for textured surface classification, but not evaluated on freehand acceleration data. Applied on the movement phase, we consider it to represent the acceleration movement fineness (AMF) of the textured surface, which is assumed to increase at higher velocities [7]. When using our recording setup, which differs from the one in [7], this observation was not confirmed during our data collection procedure. As an example, we use our controlled recordings for three different materials from our database, where the scan velocity was linearly increased and calculate the SC for 1-s-long frames as introduced in Section 4.1.2.2, see Fig. 14. In the same way, we observe for freehand recording frames that the SC stays nearly constant[2]. The approach in [7] deliberately cuts off frequencies above 600 Hz, motivated by the perception threshold of the human skin. From our observations we conclude that the frequencies above 600 Hz include further information for the surface classification; thus, we do not apply such a filter.

4.1.4.4 Regularity: The amount of self-repeating patterns and regular structures of object surfaces, and hence in the acceleration signals resulting from the corresponding recordings, is a characteristic surface property. As shown in [4], a comparison to the voiced and unvoiced speech can be drawn because signals may either show the appearance of quasi-periodic patterns or similarity to random noise. Each movement signal value from $\mathbf{x}$ is range-normalized with the denominator $d = \max(\mathbf{x}) - \min(\mathbf{x})$ and stored in $\hat{\mathbf{x}}$. We calculate the auto-correlation vector $\mathbf{r}_k$ of $\hat{\mathbf{x}}$ and take the derivative $\Delta\mathbf{r}_k = \hat{\mathbf{r}}_k - \hat{\mathbf{r}}_{k-1}$. All values smaller than zero in $\Delta\mathbf{r}_k$ are set to zero since we assume that the amount of regularity is correlated to the positive slopes within the auto-correlation function $\mathbf{r}_k$ of a surface signal. We finally calculate our acceleration movement regularity feature (AMRG) as

$$AMRG = \overline{\Delta\mathbf{r}_k} \tag{7}$$

4.2 Friction Feature

Friction is a major tactile dimension and relevant for surface classification, as stated in [7] and [11]. It is estimated in [8] from

2. Further examples on http://www.lmt.ei.tum.de/texture/

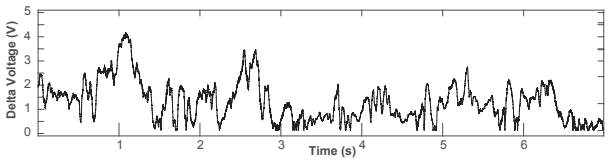

Fig. 15: The differential friction force values (absolute values) for the interaction with an exemplary foam surface. Note that smaller values correspond to phases, where the operator presumably changed the exploration direction and larger values to friction forces during the movement phases over a surface. The mean absolute value, e.g., represents one component of the friction feature of a surface.

the force signal, which was measured with a high-quality force sensor. As we want to rely on inexpensive sensors, we attached two FSRs on both sides of the gripping part of our haptic stylus. We use the differential voltage (Wheatstone Bridge, two FSR and two 10 KΩ resistors) as input data for the calculation of our friction feature.

We measure a differential voltage vector, exemplary shown in Fig. 15 and calculate the absolute value $\mathbf{v}_\Delta = [|v_1|,\ldots,|v_n|]^T$ from each entry with n being the number of sampled friction force values. We have observed that stickier surfaces have a larger range of differential voltage values. We assume this to be related to dynamic friction and calculate the mean value to represent the dynamic friction coefficient. For example, foam surfaces typically compel users to apply higher dragging force compared with metal surfaces. The positioning of both FSR decreases the dependency in which direction the user explores the surface, because we only use the absolute differential friction force values. Moreover, the use of differentially working FSR eliminates the effect of different gripping strength of different operators.

A second component of our friction-based feature is related to the vibrations in the friction force signals, known as stick-slip effect [11]. We calculate the derivative of the friction signal $\Delta\mathbf{v}_\Delta$ and use its variance σ as stick-slip component in our friction feature (FM)

$$FM = \frac{a\cdot\overline{\mathbf{v}_\Delta} + b\cdot\sigma(\Delta\mathbf{v}_\Delta)}{\overline{|\mathbf{x}|}} \tag{8}$$

where $\mathbf{v}_\Delta$ is the magnitude differential force vector, $\Delta\mathbf{v}_\Delta$ is its derivative and a and b are weighting factors to combine both components into a single feature. For now, both values are set equal to one. We divide them by $\overline{|\mathbf{x}|}$, which is the mean absolute signal energy of the corresponding acceleration signal, which is recorded in parallel to the friction force signal. The acceleration-based denominator increases the feature robustness with regard to different scan forces and velocities, because both the nominator and the denominator depend on these influences.

4.3 Sound Features

The interaction of a tool tip with a surface does not only lead to vibrations inside the tool but also to audible sound waves that allow humans to differentiate between various surfaces. While varying force and velocity conditions typically change the perceived sound intensity, we have observed that spectral properties, e.g., the pitch or temper of the sound signals, stay characteristic for different surfaces. We observed that soft or smooth materials (e.g., Styrofoam or artificial grass) generally have lesser discriminatory ability in their recorded acceleration signals but are characteristically different in their sound spectra.

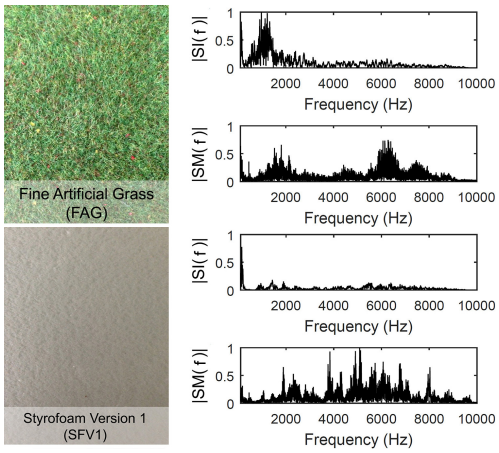

Fig. 16: $|SI(f)|$ and $|SM(f)|$ of different surfaces (fine artificial grass (top) and Styrofoam (bottom).

Intuitively, this is supported by the larger frequency range of sound signals compared to acceleration signals. We take advantage of these characteristic spectral differences in the sound features in the surface classification task. Hence, we adapt common audio features from [27], described in the following section, to extract the surface properties from the recorded sound signals.

Figure 16 shows the normalized sound signal impact impulse $SI(f)$ and movement $SM(f)$ magnitude spectra of different surfaces, such as fine grass or Styrofoam. Normalization means that all values are divided by the largest value in the spectrum, which constrains the spectral intensity range between 0 and 1. The normalization reduces the dependency of the spectral intensity on the applied normal force and tangential velocity.

4.3.1 Sound Impact Hardness

The impact hardness calculation based on sound signals is similar to our acceleration-based hardness calculation. From the sound impact signal $\mathbf{i}_{Snd}$ alone, we cannot estimate the hand approach velocity toward the surface. However, our setup allows us to simultaneously record acceleration and sound data, and hence, we use, similar to Section 4.1.2, the low-pass filtered hand acceleration data $\hat{\mathbf{h}}_{Acc}$ to estimate the hand velocity toward the surface. We define our sound impact hardness feature (SIH) as

$$SIH = \frac{\max(\mathbf{i}_{Snd})}{\bar{n}} \cdot \frac{1}{\sum_{i=1}^{n}\hat{h}_{Acc_i}} \tag{9}$$

where $\bar{n}$ and $\hat{h}_{Acc}$ are the temporal centroid of the impact impulse and the hand acceleration data, respectively.

4.3.2 Spectral Low-High-Ratio

We have observed a characteristic spectral energy distribution of sound recordings (compare the exemplary sound spectra in Fig. 16). We define lower and upper spectral thresholds and calculate the Low-High-Average ratio (SMLH) from the absolute sound movement signal spectrum $|SM(f)|$

$$SMLH = \frac{\overline{|SM(f)|_{1-4000Hz}}}{\overline{|SM(f)|_{4000-10000Hz}}} \tag{10}$$

For example, foams and artificial grass typically generate fine cracking noise, whereas ceramics tend to clang during the tool-mediated interaction according to their characteristic spectral distribution. The use of a ratio mitigates the effect of scan force and

scan velocity variations, because both nominator and denominator are effected in the same way by, e.g., a larger spectral intensity due to larger scan force.

4.3.3 Spectral roll-off Frequency

We set a spectral roll-off threshold of 95% of $|SI(f)|$ of the current sound signal spectrum. We use the spectral roll-off frequency as the sound spectral roll-off (SISR) feature that is defined as the frequency for which the sum of all intensity values of prior frequencies exceeds the spectral roll-off threshold. When the roll-off value in (11) is equal to the 95% threshold, the upper limit of the sum in the equation is our SISR feature that captures the global spectral characteristics. We observed, e.g., that sound-damping materials have the largest part of their spectral distribution in low-frequency energy bands.

$$roll-off = \sum_{k=1}^{SISR} \mathbf{P}(f_k) \stackrel{!}{=} 0.95 \cdot P_{total} \tag{11}$$

4.3.4 Spectral Harmonic

For the impact spectral harmonic feature (SISH), we first calculate the auto-correlation function $\mathbf{X}_I$ of the DCT of the sound impact data $\mathbf{i}_{Snd}$. We then take the derivative of $\mathbf{X}_I$ to obtain $\hat{\mathbf{X}}_I$ and set all values smaller than zero in $\hat{\mathbf{X}}_I$ to zero. Similar to the definition of regularity in Section 4.1.4.4, the positive values of $\hat{\mathbf{X}}_I$ can be understood as the degree of repetition inside the spectral sound domain. These repetitions can be used to approximate the harmonics of the sound signals. We calculate the centroid of $\hat{\mathbf{X}}_I$ as the SISH feature

$$SISH = \frac{\sum_{k=\frac{i}{2}}^{K} |\hat{\mathbf{X}}_I(f_k)| \cdot f_k}{\sum_{k=\frac{i}{2}}^{K} |\hat{\mathbf{X}}_I(f_k)|} \tag{12}$$

where K denotes the number of elements in $\mathbf{X}_I$. Because of its symmetric nature, we only use half of $\hat{\mathbf{X}}_I$ in the calculation. This feature is related to the audible characteristic pitch generated when the tool tip impacts the surface.

4.3.5 Spectral Spread

In (2) we use the definition of the SC in our acceleration signal-based hardness feature that is related to the stiffness of a surface. Similarly, the SC of the sound spectra can be calculated and used as a Sound Impact SC (SISC) feature. In our experiments, we observed that the closely related spectral spread captures further spectral characteristics of the sound signals. For example, Fig. 16 shows strong differences in the spread around the SC for the two given surfaces. Hence, we consider the sound impact spectral spread (SISS) as an additional feature

$$SISS = \frac{\sum_{k=1}^{\frac{m}{2}} (f_k - SC)^2 \cdot \mathbf{I}(f_k)^2}{\sum_{k=1}^{\frac{m}{2}} \mathbf{I}(f_k)^2} \tag{13}$$

that captures the spread around the calculated SISC of all frequencies f_k.

4.4 Image Features

The human perception of surface materials relies not only on touch but also on visual inspection. We argue that a technical classification system benefits from the use of various modalities because relying only on one source of information leads to strong limitations in such a system. Image-based features can be used to detect textural information, such as contrast or directionality on a surface or its color information. However, without the evaluation of acceleration data in this context, image features alone may confound the classification result, if, e.g. different printed patterns for the same material are present. We assume a multimodal surface classification approach to outweigh the disadvantages of single modalities.

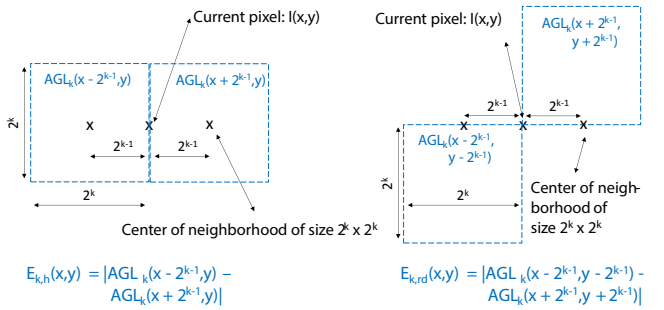

Fig. 17: Image Coarseness: Graphical illustration of the adjacent windows with respect to the current pixel in the horizontal (left) and right-diagonal direction (right).

4.4.1 Image Coarseness

Coarseness is, along with contrast and directionality, one of the most fundamental visual features and is related to the appearance of large structured elements in an image [16]. We commence by computing and storing the Average Gray Level (*AGL*) of subwindows (computed as the mean of pixel values) of size $2^k \times 2^k$, with $k \in \{1,2,3,4,5,6\}$, which are centered at every pixel within the image. As illustrated in Fig. 17, we perform the symmetric subtraction of AGL of non-overlapping, adjacent subwindows in the horizontal, vertical and diagonal directions with respect to the current pixel for the complete range of sizes and store the results. In the horizontal direction, e.g., it is computed as $E_{k,h}(x,y) = |AGL_k(x+2^{k-1},y) - AGL_k(x-2^{k-1},y)|$. We use a 45° resolution for the *AGL* difference, unlike the 90° one used in [16], to detect textured elements that have other orientations as well and to make this feature less dependent on image rotations. For each pixel, we select the k value that corresponded to the maximum gradient of average gray levels of the adjacent subwindows in all directions we obtain $S_{best}(x,y) = 2^k$. Image coarseness is determined as the average of S_{best} over the entire picture

$$ICOA = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} S_{best}(i,j) \tag{14}$$

where m and n are the width and height of the picture, respectively. For example, coarse surface images typically contain large-structured areas, which correspond to large gradients of large AGL elements.

4.4.2 Edginess

To calculate a feature capturing the edginess of an input image I, we use a low-pass filter with a cutoff frequency of 1% of the width of the Fourier transform of I to obtain the blurred image I_{LPF}, see Fig. 18. We compute the difference image I_{diff}, which evens out the intensity levels of the initial image and enhances the edges, if present. Finally, the edginess feature (IE) is obtained as the ratio

Fig. 18: Edginess calculation. The difference between the original image (left) and the blurred image (middle) exhibits the large image energy for strongly edged surfaces (right).

between the energy (sum of all pixels) of the I_{diff} image and the energy of the initial image:

$$IE = \frac{\sum_i \sum_j |I_{diff}(i,j)|}{\sum_i \sum_j I(i,j)} = \frac{\sum_i \sum_j |I(i,j) - I_{LPF}(i,j)|}{\sum_i \sum_j I(i,j)} \quad (15)$$

A large value of this ratio corresponds to a high number of edges, leading to a high IE feature value. The denominator is used to normalize the nominator to the original energy of $I(i,j)$.

4.4.3 Glossiness

The appearance of surfaces is strongly affected by illumination, reflection properties, and the surface geometry. The camera-captured gloss of a surface is directly connected to its capability of reflecting light and thus its micro-structure. We assume that two components, related to diffuse and specular reflection, are important in the surface classification task at hand.

We convert the RGB image into the YCbCr color space and we only retain the Y component because of its human perception-related luminance definition. We map these values from the image I to the new values in $I_{adjusted}$ to correct its contrast so that 1% of the data is saturated at low and high intensities of I. Afterwards, we compute the histogram H_I of $I_{adjusted}$ and the histogram H_Y of the initial image luminance layer. Subsequently, we compute the skewness s of H_Y, which is correlated to the general surface gloss, as discussed in [28]. As for the specular reflection, we compute the area under the histogram H_I from gray levels 250 to 255, obtaining $A_{specular}$ and divide this value by the area under the entire histogram, A_{total}. According to our observation, high values in the intensity levels correspond to specular reflections and the existence of spotlights. This component allows us to differentiate between glossy surfaces and surfaces with large white areas. Our image glossiness (IG) feature is defined as

$$IG = \frac{A_{specular}}{A_{total}} + s \quad (16)$$

with the skewness s of H_Y. If the luminance histogram H_Y is negatively skewed, the corresponding surface is assumed to be less reflective and has lower diffuse and ambient light reflection, whereas when H_Y is positively skewed, the surface is assumed to be more reflective.

Characteristically high values for $A_{specular}$ can only be achieved when a surface has a strong tendency for specular reflection, whereas a whitish surface still has most of its high-intensity values below the above-mentioned threshold.

4.4.4 Line-likeness

This feature determines the degree of line occurrences in an image. It distinguishes between irregular, coarse structures and those that may have visual lines e.g., fur, wooden, or fiber-containing surfaces. Figure 19 compares line- and blob-like textures. As suggested in [16], we use the gray-level co-occurrence matrix (GLCM) in the feature calculation. The basic idea is proposed

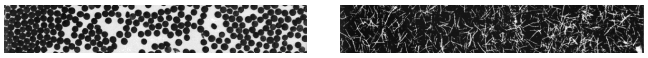

Fig. 19: Line-likeness: The left surface image is blob-like, whereas the right surface image has strong line components.

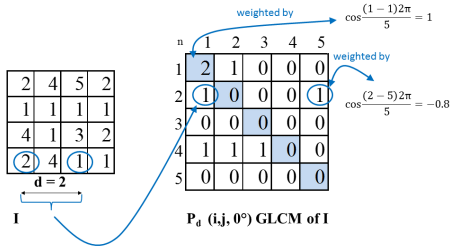

Fig. 20: Illustration of the GLCM P_d matrix for $d = 2$ that is computed for an image I. The 0° argument of P_d suggests that the GLCM is computed in the horizontal direction and its element in position (1,1) represents the number of horizontally adjacent pixels in I that are separated by the distance $d = 2$ and have the values (1,1). Using this matrix, a line is modeled as a pair of neighboring pixels with the same values, which represents the entries on the main diagonal of P_d, or with similar values, which represent the rest of the entries. The values of the weighting factor represented by the cosine function decrease as the elements of the GLCM move away from each other, thus decreasing the contribution of the pairs of pixels with radically different intensity values, which are very likely not to represent a line.

by [17]. The element $P_d(i,j)$ of the GLCM matrix is defined as the relative frequency, with which two image intensities (at positions i and j) are separated by distance d as shown in Fig. 20. We define the image line-likeness (IL) as

$$IL = \sum_{k=1}^{4} \sum_{d=1}^{10} \frac{\sum_i^n \sum_j^n P_d(i,j,k) \times cos(\frac{(i-j)2\pi}{n})}{\sum_i^n \sum_j^n P_d(i,j,k)} \quad (17)$$

where n is the width (and height) of the P_d matrix. Instead of selecting a certain value for d as in [16], we select a range of values for d ($d = 1...10$). We believe that a range is a better option because with fixed d values greater than one, [16] does not consider the pixels which fall between the ends of the pixel-pair of interest. The assumption that no sudden changes between the values of neighboring pixels is not sufficient for the task of surface classification. Unlike [16], we use a range of values for d and search for consistent and coherent line patterns inside the image.

We have chosen a 45° quantization level of four directions (horizontal, vertical, both diagonals). This leads to $k \in \{1...4\}$ and increases the independence of different image rotations.

4.4.5 Image Roughness

The roughness of a surface is normally measured using acceleration sensors. However, humans are also able to examine a given structure and evaluate its roughness visually. We define our image-based roughness feature on the outcome of a subjective experiment, which is described in Section 4.4.5.1.

Our roughness feature is based on the input image I as well as its GLCM. As described in [17], a GLCM represents a statistical method of examining textured patterns that consider the spatial relationship among pixels, as used in Section 4.4.4. We reduce

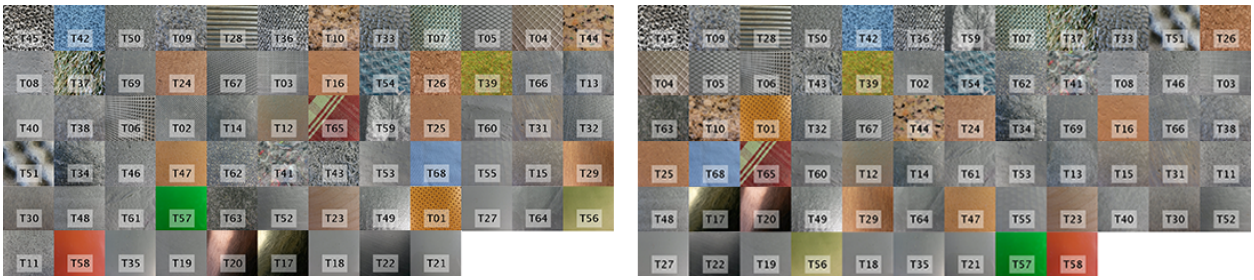

Fig. 21: Visual-touch-based ordering result for subjective test. We applied a mean ordering procedure on the experiment results across all subjects. For example, all subjects ranked surface number 45 as most - or second most roughest surface ($rank = 1$), so its mean ordering rank is equal to the most roughest surface in this result list sorted in ascending order.

the number of gray levels of the image to 32 instead of using all 256 values for faster processing and normalize the GLCM values between 0 and 1. In addition, as in Section 4.4.4, the GLCM is computed for 0°, 45°, 90°, 135° and averaged in these four directions, resulting in a reduced rotation dependency.

4.4.5.1 Subjective Experiment - Procedure and Task: We performed a subjective test with 20 subjects, four female and 16 male subjects to obtain the roughness feature and to find the regression coefficients. The experiment comprised two parts, which were conducted on different days.

In the visual part of the experiment, we used two monitors, where the subjects could order the files containing the corresponding surface image on the desktop according to their visually perceived roughness level. The subjects were instructed to base their decisions only on what they perceive in terms of roughness when looking at the database images, thus keeping the recreation of the feeling of the texture out of the equation when performing the ordering.

In the visual-touch experiment, we sliced the textured surfaces in stripes and placed them on two tables to let the subjects freely perform the ordering procedure. The subjects were not blindfolded to make it feasible to arrange all the 69 surfaces in reasonable time. This procedure required the subjects to inspect the surfaces with their fingers to determine the roughness level of each texture. We selected bare finger interaction over the tool-mediated interaction for roughness examination to obtain a device-independent tactile ground truth data. This is motivated by the fact that humans have similar capabilities of discriminating between textures either via a tool or a bare finger [29]. In addition, before the visual-touch-based experiment, the subjects were asked to keep the inspection velocity and pressure applied on the surface roughly consistent throughout the entire set of textured surfaces.

4.4.5.2 Subjective Experiment - Results and Discussion: We used the mean ordering results to approximate our image roughness definition. All 14 Haralick parameters, e.g., entropy or mean image energy as detailed in [17], were computed over the image database and its GLCMs. Most of these features have a visual meaning - contrast measures the local gray-level variation, the inverse difference moment measures how uniform the entries in the GLCM are, and entropy measures the spatial disorder of the image, whereas energy is a measure of local homogeneity. Multivariate normal regression was performed to obtain the coefficients column vector $\mathbf{b}$ that would yield an ordering similar to the one obtained in the experiment, denoted by the column vector $\mathbf{y}$. If we denote the Haralick coefficients matrix as $\mathbf{H} = [\mathbf{w}_1\ \mathbf{w}_2\ ...\ \mathbf{w}_{14}]$ [17], where $\mathbf{w}_k$, for $k \in \{1,...,14\}$ are the column vectors of the Haralick coefficients corresponding to all images in our database, our regression equation can be written as $\mathbf{y} = \mathbf{Hb}$.

It is extremely unlikely that two rankings will match completely. Hence, we considered a tolerance value d for determining the ranked orderings (experimental and regression result) correlation. If an image gets a rank i in one of the two ranked lists while it gets the rank $i \pm d$ in the other, it is assumed that there is a strong correlation between the two ranking lists for that image, and we considered this a match.

The precision p of the two ranked lists is computed as the ratio between the number of matches divided by the total number of surfaces. We wanted to obtain a certain mismatch when ranking the top roughest images, and for the middle and least roughness orderings, there may be considerable mismatch between the two rankings across all subjects. To use a position-dependent value for d, as it is suggested in [30], we use the rounded value from the exponential equation $d(r) = 3 + 4.5(1 - e^{-2r/\mu})$, where μ is the mean of ranks from 1 to 69, and $d(r)$ is the recommended tolerance for rank r.

The highest regression precision, which was obtained considering all possible combinations of Haralick parameters is $p = 0.3623$, which uses the following Haralick parameters: W_3 (Correlation), W_5 (Inverse Difference Moment), W_6 (Sum Average), W_8 (Sum Entropy), W_{12} and W_{13} (Information Measures of Correlation). Hence, we define our image roughness feature (IR) as

$$\begin{aligned} IR = {} & 183.25W_3 + 8.03W_5 - 0.66W_6 \\ & -66.85W_8 - 510.32W_{12} - 51.01W_{13} \end{aligned} \tag{18}$$

We also compare the image-based and visual-touch-based orderings directly, leading to a second result in our subjective experiment. We again employ the above-mentioned precision p measurement to see the extent to which the two average rankings across the subjects correlate with each other. However, we calculated $p = 0.4638$, which suggests an existing positive correlation of visual and tactile orderings, with several mismatches during the ordering procedure. This suggests that image-based features alone are not sufficient. Additional information, such as acceleration and sound signal features, is necessary for successful surface classification. The order results for both the visual only and visual-touch-based experiments are shown in Fig. 21. In our opinion, the main reason for the visual-haptic mismatch can be found in the missing surface hardness and high-frequency acceleration (roughness) information, which cannot be conveyed by images alone. Surface images of soft surfaces, e.g., may reveal strong contrast information but do not necessarily lead to strong tactile

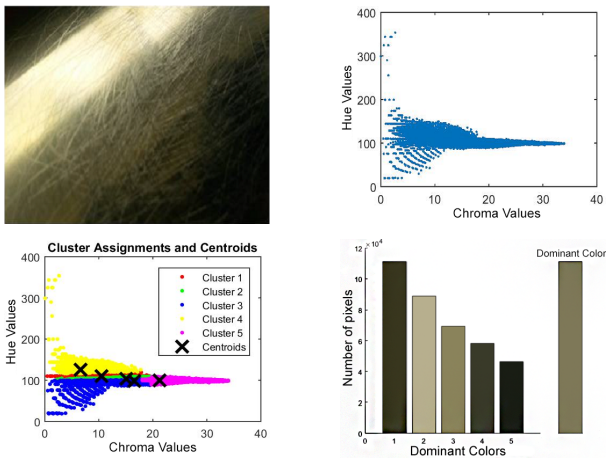

Fig. 22: Dominant color example. Given the texture (upper left), k-means clustering was performed (upper right and lower left) to determine the dominant color as the average of the three colors at the cluster centroids (lower right). The ranking of the dominant colors was based on the amount of pixels with that dominant color (lower right). Each cluster has a centroid, which has a chroma and a hue value. We obtain the color of each histogram entry by combining these values with the L value and going from the LCH to RGB space.

roughness due to their compressibility or the absence of high-frequency tactile vibrations.

4.4.6 *Dominant Color Distance*

Although strictly speaking, color is not a tactile surface feature, it becomes necessary for surface classification to differentiate among different types of surfaces, e.g., metals. In fact, the human tactile perception system is not able to distinguish metals, such as aluminum, brass, and gold, by touch only but by the dominant color during visual inspection. To determine the dominant color of the surface (see Fig. 22), we first convert the RGB space to the CIE L*c*h* color space. The RGB space is perceptually nonuniform, which means that a distance metric in this color space will not match the human-perceived distance between colors. In the next step, we form a cluster of points from this color space with all chroma values on the X axis and all hue values on the Y axis. We disregard the luminance component, because it is covered by our IG feature. Afterwards, we perform k-means clustering with k = 5 clusters and identify the center of each cluster, which corresponds to the five most dominant colors within the image, and then, we count how many pixels belong to this cluster. We then take each color as a reference and measure the difference between this and all the other four colors. The difference between two colors is expressed as ΔE and is computed as described in [31]. The resulting number describes the distance between the two colors in the color sphere, where the reference color is pure black, and is used as image color distance feature (ICD).

5 EVALUATION

We evaluated the different features using strictly separated training and testing sets for standard ML approaches, implemented in Matlab. Figure 23 shows the surface classification pipeline used in this work.

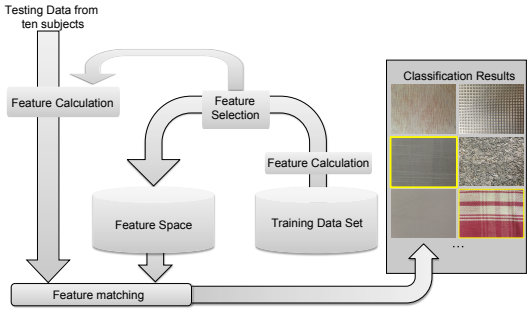

Fig. 23: Surface classification pipeline used in this work.

5.1 Parameter Identification

Before we applied the sequential feature selection, we tuned the parameter settings (e.g., thresholds) of each feature, independent of any ML approach. We used statistical analysis on single features to distinguish between well-defined and poor features. Each feature vector (calculated from ten signals for each of the 69 textured surfaces) was tested on a normal distribution (Kolmogorov - Smirnov test). The features do not follow a normal distribution, and hence, we applied a nonparametric Kruskal - Wallis test to determine if there was one group, in our case a surface, that is significantly different from the other groups. Post-hoc statistical analysis using the Bonferroni correction was used to identify significantly different groups. We calculated the corrected alpha threshold (Bonferroni) α_{bonf} and counted, how many comparisons are statistically significantly different for all p-Values ($\mathbf{p}$). A good feature shows a small intra-class variance and a big inter-class spread. The ratio of this sum to all possible comparisons C defines the goodness of features criterion (GFC)

$$GFC = \frac{\sum_{i=1,j=1}^{69,69} \mathbb{1}_{(\mathbf{p}_{i,j}<\alpha_{bonf})}(\mathbf{p}_{i,j})}{C} \quad , where\ i \neq j \qquad (19)$$

where $\mathbb{1}$ is the indicator function.

5.2 Feature Selection, training and testing procedure

We only use the training data (recorded by the authors, see 3.3.2) for a sequential forward feature selection (MATLAB Statistics Toolbox). After the identification of the n best features in the training set, we use the testing data (testing data originates from the ten subjects, see Section 3.3.3) and predict the accuracies using three supervised ML approaches from [33] (Euclidean distance-based approach, Naive Bayes, Classification Trees). For the Euclidean distance-based classifier, we use a k-means inspired classifier. Similar to k-nearest neighbor, we calculate the Euclidean distance from each testing feature vector to the training feature vectors, but not to the closest training vector but the averaged 10 training vectors of each class (surface). For the other used classifiers, Naive Bayes and classification tree, we use the Matlab implementations of the Machine Learning Toolbox.

We calculate the accuracy (A) of the resulting confusion matrices $\mathbf{M}_{CM}$ as

$$A = \frac{\sum_{k=1}^{69} \mathbf{M}_{CM}(k,k)}{N} \quad , where\ N = 69 \cdot 10 \qquad (20)$$

This equation sums up the entries for the correctly classified surfaces on the main diagonal of $\mathbf{M}_{CM}$ as a fraction of all the entries in $\mathbf{M}_{CM}$ and measures the performance of all selected features on the database. Since the training and testing data sets

TABLE 1: Comparison of the classification performance for different surfaces. We applied the sequential feature selection on selected parts of the feature space, e.g., on sound and acceleration features, and this selection identified the best features leading to the displayed ML accuracy values. Note that we applied features from related work only on the acceleration movement phase, since they are defined like this in the corresponding papers.

Feature Name	Feature Origin	Dimensionality	Naive Bayes (%)	Classification Tree (%)	Eucl. Dist. (%)	Reference
AMSP, AIH, AMF, AMTR, AMSR	Acc	5	19 ± 1.2	20 ± 2.1	19 ± 1.0	Section 4.1
FM	Frc	1	13 ± 1.4	20 ± 2.5	10 ± 1.3	Section 4.2
SIH, SISS, SISR, SISC, SMLH	Snd	5	45 ± 1.3	43 ± 2.3	46 ± 1.4	Section 4.3
ICD, ICOA, IL, IG, IR	Img	5	60 ± 1.2	62 ± 2.2	41 ± 1.5	Section 4.4
AIH, AMTR, SIH, ICOA, IG, FM	Acc, Img, Snd, Frc	6	$\mathbf{74} \pm 1.5$	67 ± 2.2	70 ± 1.3	Section 4, 4.4, 4.3,
AMTR, SIH, IG, FM	Acc, Img, Snd, Frc	4	56 ± 1.5	56 ± 2.0	58 ± 1.3	Section 4, 4.4, 4.3
30 Energy Values	Acc	30	29 ± 1.1	30 ± 2.2	26 ± 1.2	[8]
Spectrogram	Acc	125	39 ± 1.2	29 ± 2.3	10 ± 1.3	[5]
Spectral Centroid	Acc	1	9 ± 1.4	7 ± 2.1	7 ± 1.3	[7]
Standard Roughness	Acc	1	11 ± 1.1	10 ± 2.1	10 ± 1.3	[7]
10 LPC Coefficients	Acc	10	5 ± 1.0	4 ± 2.0	4 ± 1.2	[13]
Mean, σ, Wavelet Energy	Acc	3	12 ± 1.0	9 ± 1.9	7 ± 1.1	[32]
5 Relevant Spectral Peaks	Acc	5	9 ± 0.9	9 ± 1.9	7 ± 1.1	[12]

are strictly separated and recorded by different human subjects, the accuracy value represents a metric to evaluate the robustness of our features in the freehand surface classification, which is the major motivation for this work.

5.3 Classification results

The results for the proposed features, when applied to our database, are given in Table 1. According to the sequential feature selection, the combination of two acceleration, one sound, two image features and the FM feature works best with the Naive Bayes classifier with an overall accuracy of 74% and a dimensionality of only 6 features (see middle row in Table 1). If we select only the four best features (one from each domain), we achieve 56% - 58% accuracy, depending on the classifier. Since approaches using a higher number of features potentially suffer from over-fitting, we aimed to reduce the dimensionality of our finally proposed feature set; by using only three to six features from different modalities we provide a generalizable approach for the task of surface classification.

In the last rows in Table 1, we applied related work features on our acceleration movement signals. Note that the results from related work features use only acceleration movement-based features. Without scan-parameter robust features and using acceleration movement data, only low accuracies can be achieved during tool-mediated freehand movement explorations. This observation supports our assumption that a multimodal sensor fusion needs to be applied to successfully differentiate between surfaces.

5.4 Discussion

Several conclusions can be drawn from Table 1. If five of our acceleration features are used, 20% accuracy can be achieved. Compared to the last rows from the related work features, our acceleration features perform better while using the same dimensionality. Note that the other presented related work approaches use a larger amount of features and are thereby more likely to lead to an over-fitted accuracy. However, if only our sound or images features are used, even larger accuracies can be achieved. As for the sound features, we assume that the larger spectral content (about 15 kHz) of the sound signals is responsible for their larger accuracy. Compared to the acceleration spectral content (about 1 kHz), sound signals appear to contain more distinguishing information for surfaces, even if they are exposed to variations in scan force and velocity as well. As for the images, many surfaces reveal characteristic features even under, e.g., recording distance or illumination variations, and hence image-based features lead to a high accuracy. Note, however, that homogeneous surfaces, such as plastics or metals, may reveal no visually distinguishable clues but generate different acceleration signals during the tool-surface interaction. Also, the outcome of the experiment in section 4.4.5.1 supports the assumption that image features alone are not sufficient for surface classification. That is why we assume the combination of these features from different modalities performs best as shown in the middle row section of Table 1.

We used 69 surfaces for this study, which we considered as a representative set of materials. Note, however, that this selection is only a subset of the many different surfaces observable in the real world.

Features from various modalities can be used for the classification of textured surfaces and, presumably, to retrieve haptically similar surfaces in a surface retrieval engine in future work. During the proposed classification procedure, we observed that the n best results of the Naive Bayes or Euclidean distance classifier tend to be haptically similar to the query, especially if we apply surface structure-motivated features like our AIH or AMTR features. We have shown in this work that a supervised classification scheme is possible using different sensors. As next step, we would like to construct a classification system that retrieves the most similar known surface from our recorded database, if a new or unseen surface is explored. For this specific task, human ground truth data in terms of surface similarity needs to be collected. The features in our classification system need to be tuned to match these similarities to retrieve truly similar surfaces in terms of tactile perception.

Another point we have to mention about the proposed scheme considers extreme combinations of scan force and velocity. No operator deliberately applied low or large scan velocities or forces during the recordings, which may decrease the classification performance. Note, however, that in a possible real life scenario a human operator generally applies reasonable combinations of these two parameters. Moreover, surfaces such as foams or sponges compel the users to specific exploration patterns, like applying a slower scan velocity. If smartphones are used for surface classification, the graphical user interface can visually notify the user to apply a different scan velocity or to change the device inclination through its sensor capabilities.

Convolutional Neural Networks are known to perform well on images [34] and have initially been used on surface acceleration

signals and surface image data in [18], [19]. Hand-crafted features, as used in this and related works, may not discover all possible but hidden features in the recorded signals. We believe that a combinational approach of hand-crafted features and a deep learning structure, however, may further improve the performance and robustness of the classification system in future works.

6 CONCLUSION

This study extends the surface material database in [9], which comprised controlled and freehand recorded acceleration signals, by corresponding sound signals and surface images. New and adapted perceptual features that mitigate the dependency on the force and velocity variations are presented and evaluated using three standard ML approaches. A low-dimensional combination of acceleration, sound, friction and image features as input in a Naive Bayes classifier yields an accuracy of 74% from a set of 69 different surfaces. As imitation of the human surface perception system, which applies different modalities such as acceleration and image-based features, we have created a low-cost system for successful freehand surface classification.

ACKNOWLEDGMENTS

This work has been supported by the German Research Foundation (DFG) under the project STE 1093/6-1 and, in part, by the European Research Council under the European Unions Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement no. 258941.

REFERENCES

[1] K. Kuchenbecker, J. Romano, and W. McMahan, "Haptography : Capturing and Recreating the Rich Feel of Real Surfaces," in *The 14th International Symposium on Robotics Research (ISRR)*, 2009.

[2] W. McMahan, J. M. Romano, A. M. A. Rahuman, and K. J. Kuchenbecker, "High frequency acceleration feedback significantly increases the realism of haptically rendered textured surfaces," in *2010 IEEE Haptics Symposium*. IEEE, 2010, pp. 141–148.

[3] J. M. Romano and K. J. Kuchenbecker, "Creating realistic virtual textures from contact acceleration data," *IEEE Transactions on Haptics*, vol. 5, no. 2, pp. 109–119, April 2012.

[4] R. Chaudhari, B. Cizmeci, K. J. Kuchenbecker, S. Choi, and E. Steinbach, "Low Bitrate Source-filter Model Based Compression of Vibrotactile Texture Signals in Haptic Teleoperation," in *ACM Multimedia 2012*, October 2012, pp. 409–418.

[5] J. Sinapov, V. Sukhoy, R. Sahai, and A. Stoytchev, "Vibrotactile recognition and categorization of surfaces by a humanoid robot," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 488–497, 2011.

[6] C. M. Oddo, M. Controzzi, L. Beccai, C. Cipriani, and M. C. Carrozza, "Roughness encoding for discrimination of surfaces in artificial active-touch," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 522–533, 2011.

[7] J. A. Fishel and G. E. Loeb, "Bayesian exploration for intelligent identification of textures," *Frontiers in neurorobotics*, vol. 6, no. 4, pp. 1–20, June 2012.

[8] J. M. Romano and K. J. Kuchenbecker, "Methods for Robotic Tool-Mediated Haptic Surface Recognition," in *IEEE Haptics Symposium (HAPTICS)*, February 2014, pp. 49–56.

[9] M. Strese, C. Schuwerk, and E. Steinbach, "Surface classification using acceleration signals recorded during human freehand movement," in *IEEE World Haptics Conference (WHC) 2015*, June 2015.

[10] H. Culbertson and K. J. Kuchenbecker, "Should haptic texture vibrations respond to user force and speed?" in *IEEE World Haptics Conference (WHC) 2015*, June 2015, pp. 106–112.

[11] S. Okamoto, H. Nagano, and Y. Yamada, "Psychophysical Dimensions of Tactile Perception of Textures." *IEEE Transactions on Haptics*, vol. 6, no. 1, pp. 81–93, March 2013.

[12] N. Jamali and C. Sammut, "Majority voting: material classification by tactile sensing using surface texture," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 508–521, June 2011.

[13] H. Culbertson, J. Unwin, and K. J. Kuchenbecker, "Modeling and Rendering Realistic Textures from Unconstrained Tool-Surface Interactions," *Transactions on Haptics*, vol. 7, no. 3, pp. 381–393, July 2014.

[14] W. Mayol-Cuevas, J. Juarez-Guerrero, and S. Munoz-Gutierrez, "A first approach to tactile texture recognition," in *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, vol. 5. IEEE, 1998, pp. 4246–4250.

[15] W. He, H. Guan, and J. Zhang, "Multimodal object recognition from visual and audio sequences," in *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2015 IEEE International Conference on*. IEEE, 2015, pp. 133–138.

[16] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 8, no. 6, pp. 460–473, June 1978.

[17] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, November 1973.

[18] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," *arXiv preprint arXiv:1511.06065*, 2015.

[19] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Ozer, and E. Steinbach, "Deep learning for surface material classification using haptic and visual information," *arXiv preprint arXiv:1512.06658*, 2015.

[20] 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 *IEEE Haptics Symposium (HAPTICS)*, February 2014, pp. 319–325.

[21] 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*. Springer, 2010, pp. 79–86.

[22] M. Strese, J.-Y. Lee, C. Schuwerk, Q. Han, H.-G. Kim, and E. Steinbach, "A haptic texture database for tool-mediated texture recognition and classification," in *Proc. of IEEE HAVE*, Dallas, Texas, USA, October 2014.

[23] H. Culbertson, J. Unwin, B. E. Goodman, and K. J. Kuchenbecker, "Generating haptic texture models from unconstrained tool-surface interactions," in *World Haptics Conference (WHC)*. IEEE, April 2013, pp. 295–300.

[24] K. Wojcicki. (2011) Htk mfcc matlab. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab/content/mfcc/mfcc.m

[25] K. J. Kuchenbecker, J. Fiene, and G. Niemeyer, "Improving contact realism through event-based haptic feedback," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 2, pp. 219–230, 2006.

[26] L. A. Jones, "Kinesthetic sensing," in *Human and Machine Haptics*. Citeseer, 2000.

[27] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press, 2014.

[28] I. Motoyoshi, S. Nishida, L. Sharan, and E. H. Adelson, "Image statistics and the perception of surface qualities," *Nature*, vol. 447, no. 7141, pp. 206–209, May 2007.

[29] S. Lederman and R. Klatzky, "Haptic perception: A tutorial," *Attention, Perception, & Psychophysics*, vol. 71, no. 7, pp. 1439–1459, October 2009.

[30] M. M. Islam, D. Zhang, and G. Lu, "A geometric method to compute directionality features for texture images," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, May 2008, pp. 1521–1524.

[31] G. Sharma, W. Wu, and E. N. Dalal, "The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color research and application*, vol. 30, no. 1, pp. 21–30, February 2005.

[32] D. S. Chathuranga, Z. Wang, V. A. Ho, A. Mitani, and S. Hirai, "A biomimetic soft fingertip applicable to haptic feedback systems for texture identification," in *Proc. of IEEE HAVE*, 2013, pp. 29–33.

[33] P. Flach, *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

Matti Strese studied Electrical Engineering at the Technische Universität München (Germany). He received the degree Master of Science in July 2014. After this he joined the Media Technology Group at the Technische Universität München in September 2014, where he is working as a member of the research staff and working toward the PhD degree. His current research interests are in the field of analysis of haptic texture signals, surface classification and artificial surface synthesis devices.

Clemens Schuwerk studied Electrical Engineering at the Technische Universität München (Germany) and University of Edinburgh (Scotland). He received the degree "Dipl. Ing. (Univ)" in November 2010. After this he joined the Media Technology Group at the Technische Universität München in March 2011, where he is working as a member of the research staff. His current research interests are in the field of networked haptic virtual environments and Robotic Vision (RoVi), where he is developing novel visuo-haptic sensor systems for robotic manipulators.

Albert Iepure studied Electrical Engineering at the Technische Universität München (Germany). He received the degree Master of Science in December 2015. After this he joined Telefonica GmbH, where he is currently working as a Network Performance Management Engineer. His major interests are in the fields of self-organizing networks, as well as image and video compression algorithms.

Eckehard Steinbach (IEEE M96, SM08, F15) studied Electrical Engineering at the University of Karlsruhe (Germany), the University of Essex (Great-Britain), and ESIEE in Paris. From 1994-2000 he was a member of the research staff of the Image Communication Group at the University of Erlangen-Nuremberg (Germany), where he received the Engineering Doctorate in 1999. From February 2000 to December 2001 he was a Postdoctoral Fellow with the Information Systems Laboratory of Stanford University. In February 2002 he joined the Department of Electrical and Computer Engineering of the Technical University of Munich (Germany), where he is currently a Full Professor for Media Technology. His current research interests are in the area of haptic and visual communication, teleoperation over the Tactile Internet, indoor mapping and localization.