Texture Classification by Audio-Tactile Crossmodal Congruence

Yawen Liu*, Shihan Lu*, and Heather Culbertson

Abstract— **This work presents a large-scale texture classification method using a novel texture feature *Projected Spectral Mapping* (PSM) based on audio-tactile crossmodal congruence in unconstrained tool-surface interactions. We describe a quick-computable extraction process for PSM from the proposed crossmodal inter-band spectral mapping (IBSM) that relates the frequency components in different bands between the modalities. We conducted a texture classification on the LMT Haptic Texture Dataset with 69 textures in 9 categories to evaluate the PSM feature by both random sampling train-test split and participant-specific cross-validation. Compared to a variety of texture features from previous work, the results showed that our PSM feature reached >74% classification accuracy by 3 out of 4 classifiers and outperformed all other features with significant improvement.**

I. Introduction

Texture classification is an important tool in the broader area of object identification and scene analysis. This task goes beyond the limits of vision, and accuracy is improved when both sound and touch are considered in the classification task. The simultaneous haptic and auditory feedback experienced during interactions with textured objects jointly help us determine the object's properties and distinguish differences between textures. The anatomically-linked neural substrates and cross-talk between somatosensation and audition enable humans to not only understand the texture they are touching, but also bidirectionally verify this understanding based on the crossmodal connection [1]. Additionally, the temporal frequency channels across touch and audition are shown to be associated and affect a human's perception [2]. However, it is still unclear how to use this correlation between feeling and hearing to create a texture classification method that does not require human input.

A variety of multimodal features have been explored for texture classification in tool-surface interactions [3], [4]. Researchers have explored different methods for combining these features including stacking the multimodal features [5], [3], [6], statistical sequential inference [7], and crossmodal coupled deep learning [8]. However, most research has focused on visual-tactile or visual-audio features, and the congruence between audio and tactile features has been neglected. Unlike visual-tactile or visual-audio features that are weakly paired and highly heterogeneous, the high temporal and spectral congruence between tactile and audio cues from the interactions can naturally play a significant role in identifying the texture.

In unconstrained interactions, the two scan-time parameters of applied force and scan speed significantly affect the multimodal data on which the texture classification relies. Strese *et al.* [3] showed preliminary success in texture classification using scan-free features from multimodal data and Burka *et al.* [6] demonstrated the inferiority of scan-dependent features, which produced lower accuracy when including the scan-time parameters in the model. However, a robust feature that mitigates the effect of scan-time parameters on texture classification has not been developed. We address this problem of unconstrained and scan-dependent data by focusing on the crossmodal relationship between the frequencies of audio and tactile data.

In this work, we conduct a large-scale texture classification task by analyzing the audio-tactile crossmodal congruence in the frequency domain and leveraging the features extracted from the crossmodal congruence. We propose a new congruence measure using the inter-band spectral mapping, from which we capture the low-dimension texture feature that is robust to the scan-time parameters in unconstrained tool-surface interactions. We assess the effectiveness of this feature in texture classification compared to other texture features and prior classification work.

II. Background

There has been significant prior work on texture classification covering single to multi modality, controlled to unconstrained interactions, and tool-mediated to direct contact.

Acceleration data has been used extensively in a number of experiment setups for texture classification as a single modality. Romano *et al.* [9] used features from the Discrete Fourier Transform (DFT) of acceleration data collected by a PR2 robot with a custom haptic recording tool to classify 15 textures under varying contact conditions. Strese *et al.* [5] presented a tool-mediated classification method for 69 textures using acceleration features invariant to scan-time parameters during freehand movements. This work was extended to multimodal texture classification by combining features from sound, friction, and image [3]. Based on the scan-free features of the acceleration data in [5], Burka *et al.* [6] further compared their scan-dependent features and revealed the superiority of scan-free features for texture classification in unconstrained tool-surface interactions.

Features from modalities beyond acceleration have also been explored. Mayol-Cuevas *et al.* [10] created an early texture recognition method using the sound produced when a pen was dragged over the surface. Classification using images has also achieved decent performance using the well-known Haralick textural features [11]. However, subjective

*These authors contributed equally to this work.

Y. Liu is with the Department of Electrical Engineering. S. Lu and H. Culbertson are with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA. {yawenliu,shihanlu,hculbert}@usc.edu

texture ordering results based on image-only and visual-touch features suggested that the image-based features alone were inadequate to match the human perception of texture roughness and that additional information such as acceleration and sound data was needed for better classification [3].

With the release of several open-source multimodal texture databases [4], [12], researchers have changed focus to the fusion of multimodal features, with many tackling the problem using deep learning. Liu *et al.* [13] investigated learning methods across visual-audio and visual-tactile features for mutual information pairing to enhance classification accuracy. Using a humanoid robot to explore the objects, Tatiya *et al.* [14] used a joint neural network to process visual, auditory, and haptic data combined with exploratory motions for object recognition. However, the multimodal deep learning can easily generate biases between the modalities and overly rely on a single modality in the prediction.

Various bio-inspired tactile sensors that offer the robot a human-like sense of touch have brought new opportunities for texture exploration. The BioTac, an early multimodal tactile sensor measuring force, vibration, and temperature, was used for texture discrimination by collecting data during varying combinations of force and velocity [15]. This work inspired related active tactile exploration research through tactile sensors [16]. Other tactile sensors, such as an artificial finger measuring applied strain, pressure, and vibration [17], SAPHARI tactile skin measuring contact position and force [18], and GelSight collecting tactile images [19], have also been used for texture recognition coupled with active exploration. However, most texture classification systems using tactile sensors attached to a robot finger require expensive sensors and complex exploratory motion planning, which can be an obstacle to unstructured and rugged surfaces. With those surfaces, although the scan-time parameters can be programmed, high variances in force and speed are still observed and must be carefully compensated for during classification tasks. In this work, we demonstrate that by using audio-tactile crossmodal congruence in the frequency domain, we can realize a high texture classification accuracy using the data from unconstrained tool-surface interactions.

III. Crossmodal Data

In this section, we introduce our audio-tactile data processing pipeline, including data alignment, normalization, and frequency-domain representation as preparation for the crossmodal congruence analysis in Sec. IV.

A. Dataset

We use the LMT Haptic Texture Dataset [3] for crossmodal congruence analysis and further texture classification. The LMT dataset consists of multimodal tool-surface interaction data, including acceleration (vibration), sound, friction, and images of 69 textures in 9 categories ranging from mesh to fabric. In this work, we mainly focus on the vibration and sound data. For each texture, the database includes a set of training data with 10 samples recorded by the authors and a set of testing data with 10 samples recorded by ten participants. All experimenters explored the textures in an unconstrained manner with a wide range of scan-time parameters. Since the sound and vibration data were not recorded from the same motion in the training set, we use only the testing set that does consist of multimodal data that was recorded from the same motion but with significant delay between sound and vibration data due to recording limits.

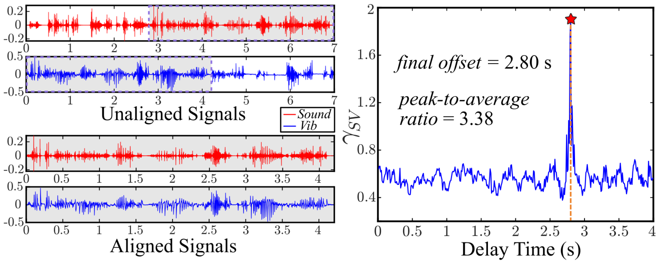


Fig. 1: Left: Unaligned sound and vibration signals, and aligned signals by feature-based CCA; Right: Canonical correlation index γ_{SV} between unaligned sound and vibration.

B. Data Preprocessing

1) Feature-based Data Alignment: Crossmodal Congruence refers to a match in frequency or intensity between the feedback from two modalities [20], and requires synchronized crossmodal signals. However, there was a delay in the data recording for vibration and sound in the LMT dataset, resulting in signals that are not temporally aligned. So, we need to align the data to obtain synchronized signals. Cross-covariance methods are widely used for this task [21], but they cannot handle the variation and size of the temporal offset between vibration and sound signals in LMT. Additionally, there is no ground truth of the synchronized output to guide the alignment.

Previous research has shown that touch-produced vibration and sound are tightly correlated in terms of their features in the frequency domain [22]. Using this correlation, we adapt a feature-based alignment method for vibration and sound signals using Canonical Correlation Analysis (CCA) [23]. The goal is to synchronize vibration and sound tracks by maximizing the mutual information.

The sampling rates in the LMT dataset are 10 kHz for single-axis vibration and 44.1 kHz for sound. We use Mel-Frequency Cepstral Coefficients (MFCC) as features for both sound and vibration, considering their effectiveness in audio recognition and possible applicability for vibration. MFCC approximates the human hearing system's response using the Mel scale for frequency bands. To match the feature size, both vibration and sound signals are processed with a time window of 30 ms with 20 ms overlap (see [9]). For each window, 14 MFCC coefficients are computed for both modalities. Based on the feature extraction, we assume a given delay time s and denote the MFCC from the sound and vibration as f_S^k and f_V^{k+s}, where k is the time reference. Both f_S^k and f_V^{k+s} are $l \times w$ matrices in the format of *feature* $\times$ *time* with $l = 14$ and w depending on the signal length. The whole canonical correlation between f_S^k and f_V^{k+s} is computed as $\gamma = (\gamma_1, \gamma_2, ..., \gamma_l) \in \mathbb{R}^l$ from joint covariance matrices, where γ_l is the correlation from the lth pair of canonical variables. The goal is to maximize the correlation index γ_{SV} between f_S^k and f_V^{k+s}, which is defined as a function of delay time s,

$$\gamma_{SV}(s) = \sum_{i}^{l} \gamma_i^2 \tag{1}$$

We increment the delay time s from 0 to 4 seconds to iteratively compute the γ_{SV} and determine the final offset by finding the time with the maximal γ_{SV}. We can control the precision of this method by the incremental step, which is set as 0.01 s. We also use the peak-to-average ratio of γ_{SV} over the delay time as an indicator to show the confidence of the alignment. An example of the data alignment for Squared Aluminum Mesh by feature-based CCA is shown in Fig. 1.

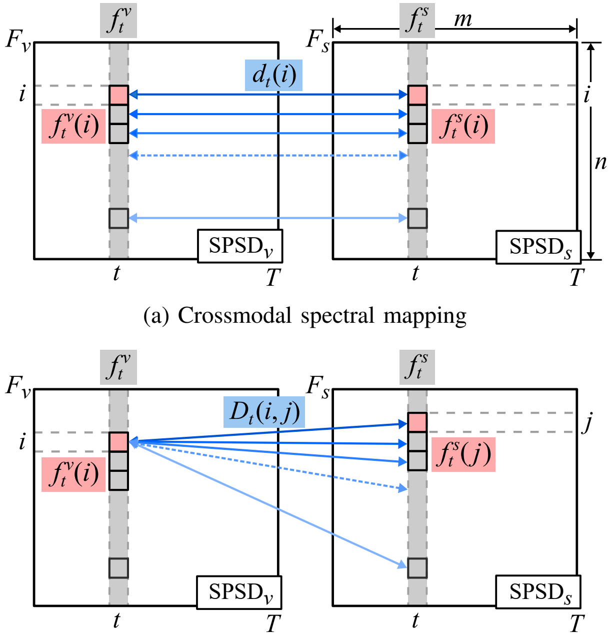
2) Data Normalization: Due to the difference in amplitude between the vibration and sound recordings, we normalize the synchronized data to better compare the signals. Both sound and vibration signals of each texture are normalized by computing the vector-wise z-score with mean 0 and standard deviation 1. This constrains the difference in the spectral intensity across modalities in later steps.

3) Data Representation: Considering the way that the feedback is generated from the interaction, we hypothesize that the synchronized sound and vibration signals have strong correlations in the time-frequency analysis for a specific tool-surface pair. Based on this hypothesis, we choose to represent our sound and vibration signals using the Spectrogram of Power Spectral Density (SPSD), a common method for describing a signal's frequency components over time. SPSD converts the temporal signal into a time-frequency structure, so it retains the variation caused by the scan-time parameters in unconstrained interactions.

The synchronized signals have a duration ranging from 3.09 to 6.97 seconds. To allow direct comparison between textures, we select the first 3 seconds from each set of synchronized signals. To match the output dimension of SPSD for sound and vibration signals, we use a 100 ms Hamming window with 25% overlap and a frequency vector with 200 bands in the SPSD computation for both signals. The frequency vectors for both modalities uniformly span their half sampling frequency. The computation results in $n \times m$ SPSD matrices, where n is the number of frequency bands and m is the number of time windows, which are 200 and 39, respectively (Fig. 2). The selection of Hamming window, overlap, and the number of frequency bands will affect the SPSD resolution and the congruence analysis. The above parameters are chosen for the best trade-off between computation time and SPSD resolution. The SPSDs from both modalities serve as the basis of our congruence analysis.

IV. Congruence in Frequency Domain

Using the SPSDs from sound and vibration signals in tool-surface interactions, we introduce inter-band spectral mapping (IBSM) as a metric that describes the crossmodal congruence in the frequency domain. As an extension of the spectral mapping, IBSM is a more robust and effective measure that reflects the characteristics of a surface.



(a) Crossmodal spectral mapping

(b) Crossmodal inter-band spectral mapping

Fig. 2: Two crossmodal spectral mapping methods

A. Crossmodal Spectral Mapping

Spectral difference analysis has been used to rate the similarity between real and virtual vibration signals in tool-surface interactions [24]. The spectral difference is defined as the sum of the normalized spectral error and power error of the discrete Fourier transform of two signals. These errors are calculated in short time windows due to the non-stationarity caused by the variance of scan-time parameters, and are averaged across all windows. Other works have also explored the spectral correlation using the Euclidean distance between frequency components [25]. But the signal's non-stationarity, a key factor in our case, has not been considered in the correlation. In addition, the above works are limited to the spectral comparison between signals from the same modality.

Inspired by the works above, we first propose a spectral mapping to represent the audio-tactile crossmodal congruence in the frequency domain. Based on the computed SPSDs from both modalities, we analyze the congruence through a direct comparison of frequency components at the same band using the Euclidean distance. The frequency components at the same band correspond to the different frequencies across modalities due to their different frequency ranges.

To mitigate the effect of the scan-time parameters, we compare the two modalities first within corresponding time windows and then average the results across all windows. The frequency components from the SPSDs at time t are denoted as f_t^v for vibration and f_t^s for sound. We define the spectral mapping at time t by computing a Euclidean distance vector d_t as:

$$d_t(i) = | f_t^v(i) - f_t^s(i) | \tag{2}$$

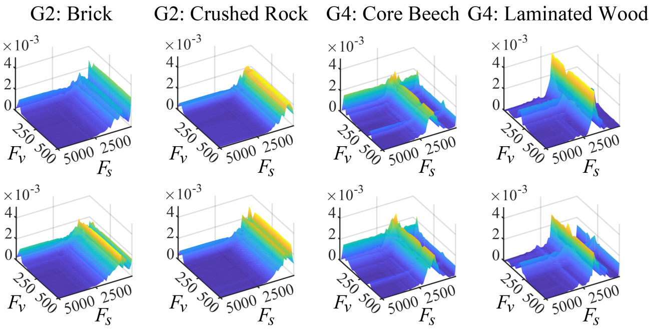


Fig. 3: Surface plots of IBSM outputs of selected surfaces from categories G2 (stone) and G4 (wood). Two rows show two different recordings of the surface. F_v and F_s are in Hz.

where $i = 1,\dots,n$ is the band index in the frequency vectors for SPSDs, and $f_t^v(i)$ and $f_t^s(i)$ are the frequency components of vibration and sound at the ith band, respectively. $d_t \in \mathbb{R}^n$ is the spectral mapping at time t, which covers all the frequency bands. The number of frequency bands is 200, as discussed in Sec. III. The procedure is shown in Fig. 2a.

We average d_t across all time windows to obtain the congruence measure based on the crossmodal spectral mapping. Crossmodal spectral mapping shows distinguishable characteristics across categories, but the difference between textures is small since the information captured by the flattened 1-D structure of the spectral mapping is still limited.

B. *Crossmodal Inter-Band Spectral Mapping (IBSM)*

To overcome the limitations of spectral mapping, we further propose a novel inter-band spectral mapping (IBSM) for measuring audio-tactile crossmodal congruence. In contrast to the spectral mapping above, which relates the frequency components in the same band between crossmodal signals, IBSM leverages the correlation of inter-band frequency components to enrich the congruence analysis.

We define the inter-band spectral mapping between modalities at time t as a matrix D_t:

$$D_t(i,j) = \mid f_t^v(i) - f_t^s(j) \mid \tag{3}$$

where $i = 1,\dots,n$ and $j = 1,\dots,n$ are the indices of the frequency band for vibration and sound, respectively. As shown in Fig. 2b, $D_t(i,j)$ represents a mapping between frequency components at the ith and jth frequency band of vibration and sound signals. $D_t \in \mathbb{R}^{n\times n}$ contains the crossmodal inter-band spectral mapping at time t.

Similarly, we calculate the crossmodal congruence measure for a texture with respect to IBSM as the average D_t over all time windows. IBSM captures the full-range interleaved frequency comparison between modalities. Compared to the 1-D structure of spectral mapping, IBSM sufficiently leverages the relationship between frequency components in different bands to realize a 2-D structure representation. This 2-D structure conveys more information about the interplay of touch-produced sound and vibration in the frequency domain and the micro features in the tool-surface pair.

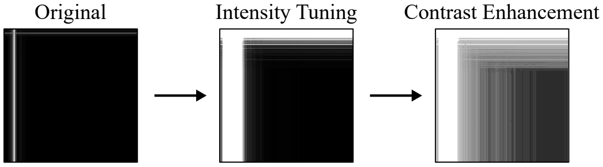


Fig. 4: Example of enhanced output of IBSM for Brick.

The difference between the IBSM for different textures is conspicuous in the surface plots (Fig. 3). Here we show only the region below 5000 Hz for sound and 500 Hz for vibration for comparison purposes. The surface plots exhibit similar visual patterns within a texture or category, and fairly different patterns across textures and categories, which provides the potential for conducting texture classification. For instance, the textures in category G4 have dominant peaks at 1700-2000 Hz for sound and 50-200 Hz for vibration in the surface plot. In contrast, the peaks for the textures in category G2 concentrate in 200-700 Hz for sound and 50-100 Hz for vibration. In addition, smooth surfaces (*e.g.*, foils and papers) generally have a single dominant peak observed in the surface plot of IBSM while rough surfaces (*e.g.*, meshes and stones) have multiple but smaller distinct peaks.

V. Feature Extraction

This section details the texture feature extraction based on IBSM. Leveraging its 2D structure, we view IBSM as an image and conduct singular value decomposition to obtain low-dimension features using dominant eigenimages. To improve the feature quality, we employ preprocessing steps, including intensity tuning, rescaling, and contrast enhancement.

A. *Quality Enhancement*

We use several techniques to enhance the IBSM outputs before feature extraction. This allows us to approximate the IBSM outputs as images and to identify the regions that contribute the most to the crossmodal congruence.

We first scale the IBSM outputs by multiplying by a factor of 7×10^6 to emphasize the peaks with a small magnitude. Next we adjust the intensity by capping the entries at 72% of the magnitude of all entries in the IBSM outputs. The scaling factor and intensity bound are determined empirically. Then we rescale the IBSM output to the interval of $[0,1]$ for visualization. We also conduct histogram equalization to adjust the visual contrast and reveal hidden details. These steps improve the differentiability of the IBSM output without changing its original characters, as shown in Fig. 4.

B. *Projected Spectral Mapping (PSM) Feature*

Eigenimages, widely used in image recognition, are defined as a set of orthogonal images that can be linearly combined to form the original image [26]. More importantly, we can project the original images to the space of dominate eigenimages to extract low-dimension features. Singular value decomposition (SVD) is a powerful tool to compute eigenimages for further feature extraction [27].

For feature extraction from the enhanced outcomes of IBSM, $\mathbf{I} \in \mathbb{R}^{n \times n}$, we first normalize the data by subtracting the mean from each sample. Then we vectorize $\mathbf{I}$ to a column vector, $\mathbf{x} \in \mathbb{R}^{p}$, and stack $\mathbf{x}$ column-wise in a matrix $\mathbf{X} \in \mathbb{R}^{p \times q}$, where q is the number of samples.

We decompose matrix $\mathbf{X}$ using SVD such that

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \tag{4}$$

where $\mathbf{U} \in \mathbb{R}^{p \times p}$ and $\mathbf{V} \in \mathbb{R}^{q \times q}$ are unitary matrices and $\mathbf{\Sigma} \in \mathbb{R}^{p \times q}$ is a matrix with non-negative entries on the diagonal and zeros off the diagonal. The columns of $\mathbf{U}$ are eigenimages, which are essential for feature extraction. The diagonal entries of $\mathbf{\Sigma}$ are singular values ranked from largest to smallest, indicating the variance of the associated eigenimages.

The eigenimages from the columns of $\mathbf{U}$ form an eigenspace. By projecting a vectorized sample $\mathbf{x}$ onto the eigenspace of the first r eigenimages, $\mathbf{U}_r \in \mathbb{R}^{p \times r}$, we obtain a set of coordinates α in this space:

$$\alpha = \mathbf{U}_r^{\mathrm{T}}\mathbf{x} \tag{5}$$

This *Projected Spectral Mapping* (PSM) feature, $\alpha \in \mathbb{R}^{r}$, captures the low-dimension features from $\mathbf{x}$. The number of eigenimages used in the eigenspace, r, determines the compression rate in the projection. We set r as 20 by observing the significant variance captured in the first 20 eigenimages when comparing their singular values with others.

We construct the eigenspace from the training set for our texture classification task, and extract PSM features for both training and testing sets using this eigenspace. The train-test split is discussed in the following section.

VI. Evaluation

To evaluate the PSM feature, we conducted a texture classification task based on recordings from unconstrained tool-surface interactions with a large set of materials.

A. Train-Test Split and Classification Algorithms

We only used the recordings from the testing set of LMT Haptic Texture Database for our evaluation because the training set of the database was recorded using different motions for sound and vibration, which violates the assumptions needed for crossmodal congruence. The testing set of the LMT database includes 69 textures with 10 recordings by ten participants for each texture, from which we extract the synchronized data to create the dataset for this evaluation. We conducted a train-test split to randomly select 7 recordings of each texture for training and use the remaining 3 recordings for testing. By this random sampling data split, a total of 483 recordings were used for training, and 207 recordings were used for testing in a supervised multi-class classification.

To assess the model's capability in classifying textures based on unseen recording styles, we also conducted a ten-fold participant-specific cross-validation on the dataset for the train-test split. In each fold, we hold out the recording data for each texture from one participant as the testing set, and then use the recordings from the remaining nine participants as the training set. There is no overlap in the participant selection for the testing sets across the folds. This cross-validation also allows us to better evaluate our PSM feature using the limited synchronized data in LMT.

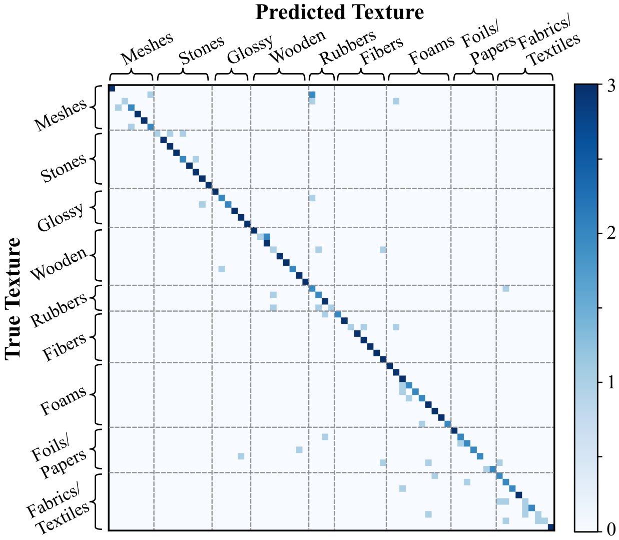


Fig. 5: Confusion matrix for the classification using PSM feature and SVM classifier with random sampling data split.

To compare our method with features in [3], [28], [9], [29], we predicted the classification accuracy using Naive Bayes (NB), Support Vector Machines (SVM), Decision Tree, and Multi-Layer Perceptron (MLP) classifiers. We used the `sklearn` library [30] in Python for the implementation of all classifiers. For NB, we set the variance as 10^{-9}. For SVM, we used C-support vector with a linear kernel considering the large number of classes. For the decision tree, we used the Gini entropy criteria for the information gain with a minimum sample leaf of 3. For MLP, we set the learning rate as 0.002 and iterations as 500 for the classifier training. We chose the logistics activation function for the output layer to support a multi-class classification. We used default values in `sklearn` for unmentioned parameters and same random states for all classifiers, if applicable. We calculated the classification accuracy by dividing the number of correctly predicted samples by the total number of samples in the testing set for the random sampling data split, and computed the average accuracy across ten folds as a measure of the classification performance for the cross-validation.

B. Experimental Results

The confusion matrix for the classification using the proposed PSM feature and SVM classifier with the random sampling data split is shown in Fig. 5. The overall texture classification accuracy was promising, though the accuracy for foils/papers and fabrics/textiles was slightly lower than that for relatively rough textures (*e.g.*, stone, fibers). This can be explained by the relatively similar perceptual feedback from interactions with different smooth surfaces. In addition, based on the texture classification results, the overall 9-category classification reached 88.9% accuracy.

TABLE I: Texture Classification Performance Using Different Features

Features	Modality	Dimension	Naive Bayes(%)	SVM(%)	Decision Tree(%)	MLP (%)	Reference
AMSP, AIH, AMTR, AMSR	Acc	4	8.70	6.28	13.53	9.18	[3]†
SIH, SISS, SISR, SISC, SMLH	Snd	5	26.57	6.76	47.83	1.45	[3]†
ICD, ICOA, IL, IG, IR	Img	5	45.89	26.09	28.02	12.08	[3]†
AIH, AMTR, SIH, ICOA, IG	Acc, Img, Snd	5	14.01	15.46	32.37	2.90	[3]†
			12.33	16.38	33.62	2.03	
10 LPC Coefficients	Acc	10	65.70	49.76	50.72	1.93	[28]
			64.35	48.09	46.67	1.45	
30 Energy Values	Acc	30	13.53	*	*	2.90	[9]
			12.03	7.39	7.83	1.45	
Spectrogram	Acc	125	*	*	7.73	*	[29]
			*	*	*	26.23	
PSM	Acc, Snd	20	74.40	77.78	43.96	79.23	Sec. V
			74.20	77.25	47.25	78.02	
PSM, ICD, ICOA, IL, IG, IR	Acc, Img, Snd	25	85.02	81.64	54.59	82.61	Sec. V
			87.83	82.18	55.07	83.34	

Gray rows show the cross-validation accuracy of the selected features. Otherwise, the classification accuracy is resulted from random sampling.

* The classification result is over-fitting, which means the training accuracy is $\geq 60\%$ higher than the testing accuracy.

† Experiments in [3] used 10 samples recorded by authors for training and 10 samples recorded by ten participants for testing. But their training data was not recorded from the same motion and cannot be aligned between modalities, so we excluded those data. To compare with our features, we used only synchronized testing data as the dataset for reproducing their tasks (see Sec. VI-A). So, our results may be different from theirs.

We also compared the classification results with other features using different classifiers, and the results are shown in Table I. Feature descriptions can be found in the listed references, and they capture a variety of spatial, temporal, and frequency characteristics of the signals. For the random sampling, we achieved an accuracy between 74.4–79.2% using our PSM feature by NB, SVM, and MLP classifiers, with the highest accuracy being 79.23% by MLP. The decision tree reached a lower accuracy of 43.96% due to its incapability of handling a large number of classes. Compared to other features from prior work with the random sampling data split, PSM outperformed all with significant accuracy improvement by NB, SVM, and MLP classifiers.

For the participant-specific cross-validation, our PSM feature achieved a comparable accuracy with that from the random sampling, reaching the highest accuracy of 78.02% by MLP. In addition, we selected a multimodal feature (AIH, AMTR, SIH, ICOA, IG) from acceleration, image and sound modalities, a highest-accuracy feature in the random sampling (10 LPC Coefficients) and two over-fitting features (30 Energy Values, Spectrogram) for the cross-validation test. Again, PSM consistently exceeded the accuracy of all other selected features based on the cross-validation results.

PSM feature relies on the modalities of acceleration and sound, so we further evaluated its effectiveness of integrating with image features. Five common textural image features, image color distance (ICD), image coarseness (ICOA), line-likeness (IL), image glossiness (IG) and image roughness (IR), were stacked to PSM for evaluation. Compared to only PSM, PSM with image features boosted the accuracy by 3–13% in the random sampling and cross-validation. Moreover, PSM with image features produced at least 26% accuracy improvement over the image-only features.

VII. Discussion

Based on the classification results from both random sampling and cross-validation, our proposed PSM feature greatly outperformed related features from prior work. By adding the image features to the PSM, we achieved the highest accuracy 85.02% and 87.83% in this evaluation, which shows PSM's effectiveness in the large-scale texture classification and its flexibility of fusing with features from other modalities. Compared to popular end-to-end deep learning method [16], the low dimension and simple structure of the PSM allow us to avoid over-fitting with a much smaller training set (<10 recordings of 3 seconds for each texture) and prevent the data augmentation. The PSM extraction is directly affected by the resolution of SPSDs of the signals. A smaller Hamming window, larger overlap, and more bands in the frequency vector generate a higher SPSD resolution, and preliminary tests show that it leads to a higher classification accuracy. The computation time for PSM averages at 1.53 s per sample with our SPSD resolution setting, showing its quick computability.

In the cross-validation, we also investigated the differences between the folds. In each fold, we held out the recordings from one participant for testing and used the recordings from the remaining participants for training. The results of PSM feature in ten folds with different classifiers are shown in Fig. 6. The accuracy was consistent across the folds with the highest accuracy around 80%, although the testing using the data from the tenth participant produced a relatively low accuracy around 60% (P10 in Fig. 6). The tenth participant used short and fast-turning motions in the recording, which is perceivably different from most other participants. But the accuracy of 58.0% and 63.8% by SVM and MLP in this fold still exceeded the performance of all other features with the same classifiers in the cross-validation.

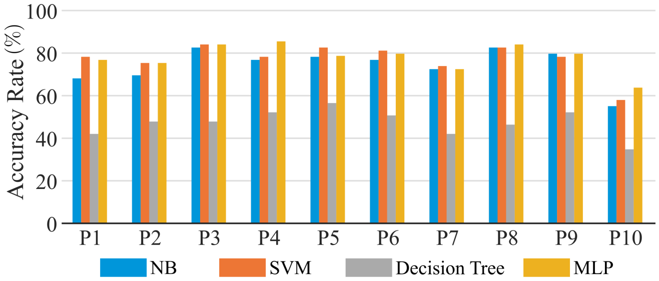


Fig. 6: Cross-validation accuracy of PSM when recordings from a specific participant are held out for testing.

For the texture features from prior work in Table I, the low-dimension features from single or multiple modalities (first four in table) did not reach 50% accuracy in any test. The high-dimension features from acceleration signals (30 Energy Values, Spectrogram) suffered from the over-fitting in both data split methods, which matches the observation in [3]. In contrast, no over-fitting was observed for PSM and PSM with image features. In addition, 10 LPC Coefficients achieved a relatively high accuracy by the Naive Bayes classifier, but its performance was inconsistent across classifiers.

In this work, the sound and vibration recordings are synchronized to compensate the time lag and ensure the congruence. How the time lag between the recordings impacts the performance has not been investigated. The generalization of proposed method to unseen textures or different scanning setup would also be a promising future direction.

VIII. Conclusion

This paper presented a large-scale texture classification method using the feature from audio-tactile crossmodal congruence based on unconstrained tool-surface interactions. We described a robust texture feature PSM extracted from the novel crossmodal inter-band spectral mapping. We achieved the highest 79% and 87% classification accuracy by PSM and PSM with image features, respectively, on a dataset of 69 textures using common classifiers with two different data split methods. PSM and PSM with image features greatly outperformed related texture features from prior work. To the best of our knowledge, this is the first attempt exploring texture classification using crossmodal congruence. This method can be applied to many tool-mediated texture classification tasks such as on-board robot tactile sensing, considering its quick computability and high robustness to scan-time parameters. It can also be particularly valuable to the scenarios where the visual feedback is limited or occluded, such as weak lighting and moving surfaces.

References

[1] T. Ro, T. M. Ellmore, and M. S. Beauchamp, "A neural link between feeling and hearing," *Cerebral cortex*, vol. 23, no. 7, pp. 1724–1730, 2013.

[2] J. M. Yau, J. B. Olenczak, J. F. Dammann, and S. J. Bensmaia, "Temporal frequency channels are linked across audition and touch," *Current biology*, vol. 19, no. 7, pp. 561–566, 2009.

[3] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, "Multimodal feature-based surface material classification," *IEEE Trans. on Haptics*, vol. 10, no. 2, pp. 226–239, 2016.

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

[5] M. Strese, C. Schuwerk, and E. Steinbach, "Surface classification using acceleration signals recorded during human freehand movement," in *IEEE World Haptics Conference*, 2015, pp. 214–219.

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

[7] W. He, H. Guan, and J. Zhang, "Multimodal object recognition from visual and audio sequences," in *IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems*, 2015, pp. 133–138.

[8] W. Zheng, H. Liu, B. Wang, and F. Sun, "Cross-modal material perception for novel objects: a deep adversarial learning method," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 2, pp. 697–707, 2019.

[9] J. M. Romano and K. J. Kuchenbecker, "Methods for robotic tool-mediated haptic surface recognition," in *IEEE Haptics Symposium*, 2014, pp. 49–56.

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

[11] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[12] J. Sinapov *et al.*, "Grounding semantic categories in behavioral interactions: Experiments with 100 objects," *Robotics and Autonomous Systems*, vol. 62, no. 5, pp. 632–645, 2014.

[13] H. Liu, F. Wang, F. Sun, and B. Fang, "Surface material retrieval using weakly paired cross-modal learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 2, pp. 781–791, 2018.

[14] G. Tatiya and J. Sinapov, "Deep multi-sensory object category recognition using interactive behavioral exploration," in *IEEE Int. Conf. Robot. Autom.*, 2019, pp. 7872–7878.

[15] J. A. Fishel and G. E. Loeb, "Bayesian exploration for intelligent identification of textures," *Frontiers in neurorobotics*, vol. 6, p. 4, 2012.

[16] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," in *IEEE Int. Conf. Robot. Autom.*, 2016, pp. 536–543.

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

[18] P. Falco *et al.*, "Cross-modal visuo-tactile object recognition using robotic active exploration," in *IEEE Int. Conf. Robot. Autom.*, 2017.

[19] G. Cao, Y. Zhou, D. Bollegala, and S. Luo, "Spatio-temporal attention model for tactile texture recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 9896–9902.

[20] E. Hoggan, T. Kaaresoja, P. Laitinen, and S. Brewster, "Crossmodal congruence: the look, feel and sound of touchscreen widgets," in *Proc. Int. Conf. Multimodal Interfaces*, 2008, pp. 157–164.

[21] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE Trans. on Image Processing*, vol. 11, no. 3, pp. 188–200, 2002.

[22] T. Kassuba, M. M. Menz, B. Röder, and H. R. Siebner, "Multisensory interactions between auditory and haptic object recognition," *Cerebral Cortex*, vol. 23, no. 5, pp. 1097–1107, 2013.

[23] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

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

[25] H. Kekre *et al.*, "Speaker identification using spectrograms of varying frame sizes," *Int. J. Comput. Appl.*, vol. 50, no. 20, 2012.

[26] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Josa a*, vol. 4, no. 3, pp. 519–524, 1987.

[27] R. Kakarala and P. Ogunbona, "Signal analysis using a multiresolution form of the singular value decomposition," *IEEE Transactions on Image Processing*, vol. 10, no. 5, pp. 724–735, 2001.

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

[29] 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.

[30] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.