2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Multimodal Material Classification for Robots using Spectroscopy and High Resolution Texture Imaging

Zackory Erickson, Eliot Xing, Bharat Srirangam, Sonia Chernova, and Charles C. Kemp

***Abstract*— Material recognition can help inform robots about how to properly interact with and manipulate real-world objects. In this paper, we present a multimodal sensing technique, leveraging near-infrared spectroscopy and close-range high resolution texture imaging, that enables robots to estimate the materials of household objects. We release a dataset of high resolution texture images and spectral measurements collected from a mobile manipulator that interacted with 144 household objects. We then present a neural network architecture that learns a compact multimodal representation of spectral measurements and texture images. When generalizing material classification to new objects, we show that this multimodal representation enables a robot to recognize materials with greater performance as compared to prior state-of-the-art approaches. Finally, we present how a robot can combine this high resolution local sensing with images from the robot's head-mounted camera to achieve accurate material classification over a scene of objects on a table.**

I. Introduction

When interacting with everyday objects, people frequently use material properties to inform their interactions [1]. We make sure not to place metal in the microwave, we take caution when carrying glass or ceramic objects, we look for styrofoam or paper cups to hold hot liquids, and we sort some paper, plastic, and metal objects into recycling bins. Robots can benefit from these same skills when operating in human environments.

In this work, we demonstrate how robots can use a non-contact multimodal sensing technique, based on spectroscopy and close-range texture imaging, to accurately estimate the materials of household objects prior to manipulation. This sensing approach collects near-infrared spectral measurements from a handheld micro spectrometer with a narrow field-of view camera for high resolution texture imaging. Both sensors are small and can be held by or directly integrated into a robot's end effector. Non-contact sensing can enable a robot to determine properties and use cases of objects without the intricacies of contact physics that can affect the performance of haptic touch-based sensing.

To evaluate this multimodal sensing technique, we have assembled and released a dataset of 14,400 high resolution texture images and corresponding spectral measurements.

*This work was supported by NSF award IIS-1514258 and AWS Cloud Credits for Research. Dr. Kemp owns equity in and works for Hello Robot, a company commercializing robotic assistance technologies.

Zackory Erickson, Eliot Xing, Bharat Srirangam, and Charles C. Kemp are with the Healthcare Robotics Lab, Georgia Institute of Technology, Atlanta, GA., USA.

Sonia Chernova is with the Robot Autonomy and Interactive Learning Lab, Georgia Institute of Technology, Atlanta, GA., USA.

Zackory Erickson is the corresponding author zackory@gatech.edu.

Fig. 1. The PR2 used a spectrometer and near-field camera to estimate the material of 144 everyday objects.

We collected this data with a PR2 mobile manipulator that interacted with 144 household objects, shown in Fig. 1, which spanned eight material categories: ceramic, fabric, foam, glass, metal, paper, plastic, and wood.

Using this dataset, we trained a neural network that learns a shared representation of spectral and visual sensory data. By learning a compact multimodal representation, our model achieves state-of-the-art material recognition performance of 80.0% when generalizing material classification to a new set of heldout objects across eight materials (12.5% baseline with a random classifier). We further investigate the role of texture image preprocessing by comparing several ImageNet-pretrained CNN models for generating lower-dimensional visual representations. Finally, using this spectral and visual sensing approach, we demonstrate that a robot can reliably classify a scene of objects on a table without direct contact. In this work, we make the following contributions:

- We introduce a near-infrared spectroscopy and high resolution texture imaging approach that surpasses prior state-of-the-art performance [2] for material classification.
- We release SpectroVision, a dataset of 14,400 high resolution texture images and spectral measurements collected from a PR2 mobile manipulator that interacted with 144 household objects from eight material categories.
- We demonstrate that our multimodal approach surpasses the performance of models trained on each independent modality.
- We show that a robot equipped only with our handheld sensors and an RGB-D camera can successfully use our approach to perform material classification on multiple objects casually arranged on a table.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

II. Related Work and Background

A. Material Recognition

Material recognition using haptic sensors, which require direct physical contact with objects, has been widely explored. Modalities such as force [3], [4], temperature [5], [6], [7], capacitance [8], vibration [9], [10], and radar [11], have been used in haptic perception for material recognition. The BioTac fingertip, capable of sensing force, temperature, and vibration, has been studied for multimodal haptic perception [12], [13], [14], [15]. Chin et al. introduced a compliant haptic sensor for robots to distinguish between plastic, metal, and paper during recycling [16]. Several works also use multimodal perception by combining data from multiple modalities for material recognition and outperforming single modality approaches [17], [18], [19], [20], [21]. Similarly, we find that non-contact material recognition approaches also benefit from multiple sensing modalities, and we demonstrate that visual sensing couples well with spectroscopy.

Several studies have evaluated visual features for material recognition [22], [23], [24], [25], [26], [27], [28]. Extensive literature also exists for leveraging visual or depth imaging for vision-based tactile sensors, including the GelSight [29], [30], FingerVision [31], and TacTip [32], [33], to perform manipulation tasks [34], [35], texture recognition [36], [37], and estimation of material properties [38], [39]. Both [29] and [40] have used visual and haptic features to estimate object properties, such as hardness or haptic adjectives. Overall, we find that multimodal approaches overcome weaknesses in the ability of any individual modality to classify materials.

B. Spectroscopy

Spectroscopy [41] has found a number of practical applications such as for pharmaceutical manufacturing [42], food analysis [43], and recycled material separation [44]. Recently, a number of handheld spectrometers have been developed for performing spectral analysis outside of lab and manufacturing settings [45], [46]. These portable micro spectrometers have been demonstrated for pharmaceutical quality control [47] and food analysis [48], [49], [50].

Prior research has shown how a robot can use near-infrared spectroscopy with a commercial handheld SCiO spectrometer to recognize materials of household objects [2]. Near-infrared spectroscopy has since been used by robots to recognize the materials of household objects for informing semantic grasp predictions and for tool construction [51], [52], [53]. In this paper, we demonstrate that robots can more accurately recognize common household materials by leveraging both spectroscopy and close-range texture imaging.

C. Texture Representation

Several techniques have been introduced for extracting or learning texture representations from visual images, including convolutional neural network (CNN) based texture analysis [54] and handcrafted descriptors [55]. Recent work in texture analysis has primarily investigated CNN-based texture representations [54]. This is due in part to a collection of works in texture and material classification tasks that have shown learned CNN feature descriptors frequently outperform alternative, handcrafted approaches [25], [26], [54], [55], [56].

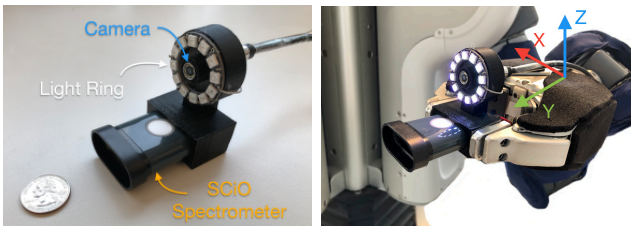

Fig. 2. Figures of the SCiO and camera/light ring sensor setup. (Left) On a table, with a quarter shown for sizing. (Right) Held by the PR2.

Research in texture synthesis [57], [58] has also provided insight into the ways in which CNNs capture and encode textures. Vision-based tactile sensing techniques for texture classification have frequently used texture features from pretrained ImageNet models [59]. The use of these models for extracting textural features is further supported by findings of Geirhos et al. [60] that ImageNet-trained CNNs are more biased towards recognizing and representing localized textures rather than global shape structure, similar to results by [57], [61], [62]. Building on these prior findings, we leverage pretrained ImageNet CNNs to extract robust visual texture features for material classification.

III. SpectroVision Dataset

A. Sensors

Our sensing approach consists of a micro handheld spectrometer for near-infrared spectral measurements and a narrow field-of view camera for high resolution texture imaging. Compared to haptic sensing, spectroscopy and imaging have advantageous properties for material recognition, including fast response times and no physical contact requirements.

Fig. 2 shows the SCiO spectrometer and camera, by themselves and when held in a PR2 robot's end effector. The SCiO is a near-infrared spectrometer that measures light spectra in the wavelength range of $\lambda = 740$ nm to $\lambda = 1{,}070$ nm. The 35 gram spectrometer is Bluetooth enabled and has a black pigmented cover around the sensor aperture which ensures there is an $\sim$1 cm minimum air gap between an object and the sensor aperture.

We capture texture images with a 2 megapixel endoscope camera. The 8.4 mm diameter camera has an optimal viewing distance of 6 cm to 10 cm and is capable of capturing images at 1600×1200 resolution. We placed a 12 LED light ring around the camera (see Fig. 2) to ensure consistent illumination of each object that the robot interacts with.

We attached the spectrometer and camera together with a grasping mount for the PR2's end effector. Note that there is an $\sim$3.5cm offset between the apertures of the two sensors.

B. Dataset and Data Collection

We have collected and released SpectroVision, a dataset[1] of 14,400 texture images and near-infrared spectral samples.

[1]SpectroVision dataset: `https://github.com/Healthcare-Robotics/spectrovision/releases`

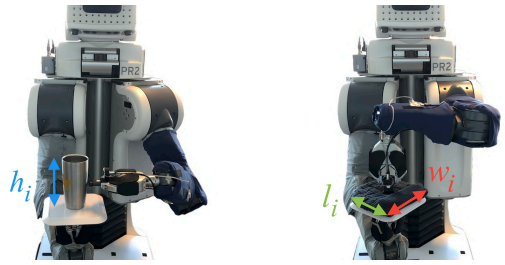

Fig. 3. Demonstration of data collection with the PR2. (Left) Interaction with a vertical object. (Right) Horizontal object interaction.

This data was captured from a PR2 robot that interacted with 144 household objects from 8 material categories, as shown in Fig. 1. These materials include ceramic, fabric, foam, glass, metal, paper, plastic, and wood, with 18 unique objects per material.

The robot performed 100 interactions with each object, sampling at random positions and orientations along an object's outer surface. To do this, the robot used its right end effector to hold a flat platter on which we rigidly mounted objects to, as demonstrated in Fig. 3. For measurements collected with vertically standing upright objects, the robot would rotate the platter, then randomly sample a roll orientation for the left end effector $\theta_y \in [-\frac{\pi}{9}, \frac{\pi}{9}]$ (see Fig. 2) and a vertical height to interact with the object at in $[0, h_i]$, where h_i represents the height of object i (see Fig. 3). For objects that lie flat on the platter, the robot would randomly sample an end effector roll orientation $\theta_y \in [-\frac{\pi}{6}, \frac{\pi}{6}]$ and a point of contact in $[0, l_i]$, $[0, w_i]$ along the top surface of the object, with length l_i and width w_i. Due to the random roll orientation of the robot's end effector and the $\sim$3.5 cm height offset between the spectrometer and camera (seen in Fig. 2), spectral and texture images captured at the same time are not co-located and hence pairings between these measurements are not strict. In early evaluations, we found that randomizing pairings between spectral and image samples from the same object did not have considerable impact on classification performance. Video sequences of the data collection process can be found in the supplementary video. We note that future research could extend the results in this work by generalizing to other common object sets [63], or evaluating multilabel classification with objects of non-homogeneous materials.

In comparison to some haptic sensing approaches that can take upwards of 15-20 seconds per measurement [9], [15], spectroscopy and imaging offer consistently fast sensing times. Capturing an image takes $\sim$1.5 milliseconds, whereas the SCiO has a 1-2 second sensing time, which consists of $\sim$1 second of light exposure, reflectance data processing, and Bluetooth communication. Data processing consists of normalizing the raw spectrum reading from the SCiO's optical head by the raw spectrum of a calibration apparatus (a high reflectance mirror material).

Fig. 4 depicts sample images from each material category, captured by the camera during the interactions. Fig. 5 shows the spectral measurements, which were captured alongside the images in Fig. 4. A raw spectral measurement consists of a 331-dimensional vector with a 1 nm wavelength step between the range of $\lambda = 740$ nm to $\lambda = 1{,}070$ nm. Prior works have shown that the difference quotient (numerical first order derivative) of spectral measurements can improve learning performance [2], [64]. Given this finding, we concatenate the difference quotient to each raw spectral measurement, resulting in a 662-dimensional spectral vector.

C. Multimodal Learning Architecture

We construct a multimodal network that learns independent representations for each modality and fuses layers at the end for multimodal classification. We begin by building separate networks to learn low-dimensional representations for the spectral and image modalities. The spectral network, Fig. 6 (A), takes as input a 662-dimensional spectral sample and outputs material probability estimates from a softmax function. The model has two 64 node hidden layers followed by two 32 node layers, with batch normalization and a leaky ReLU activation applied after each layer. We apply a dropout of 0.25 after all but the last 32 node hidden layer.

Prior to training a model over texture images, we first feed images through a DenseNet-201 CNN pretrained on ImageNet [65]. We remove the 1000-class output layer such that the network outputs a feature vector of length 1920, resulting from global average pooling on the output of the preceding convolutional block. Models trained on ImageNet often learn strong representations for texture within an image [60]. Given this, in Section IV-C, we compare material recognition results across various ImageNet-trained models used for computing texture image embeddings.

Fig. 6 (B) shows our image network, which takes as input the 1920-dimensional features from DenseNet-201. The network has three hidden layers of size 128, 64, and 32 nodes, with batch norm and leaky ReLU applied after each layer. We apply a dropout of 0.1 after the first two hidden layers. During evaluation (Section IV), we train the spectral and image networks each for 50 epochs with a batch size of 128.

Given trained spectral and texture image models, we then define our multimodal network architecture. We freeze the weights in both networks and remove the final 8 node output layer (depicted by the orange dotted lines in Fig. 6). Both models output a 32-dimensional representation for their respective sensory modality. As depicted in Fig. 6 (C), these two outputs are concatenated and fed to a 32 node hidden layer followed by a leaky ReLU activation. We use a softmax activation after the final 8 node output layer to compute probability estimates for each of the 8 material categories. Since the spectral and texture image models are pretrained, we train only the weights for layers after the concatenation for 10 epochs. We trained all models with the Adam optimizer, using $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 0.0005.

Learning separate representations for each modality and combining into a shared representation for classification is commonly used for multimodal learning [66], [67], [68], [69]. From initial tests, we found that this late fusion

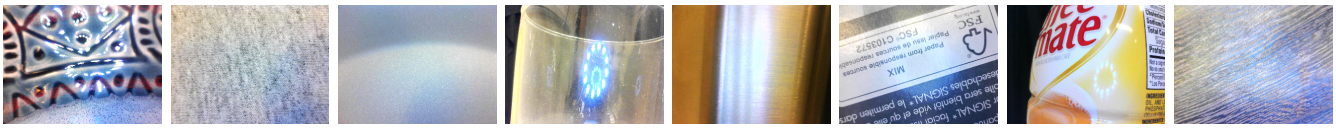

Fig. 4. Examples of texture images from each material category.

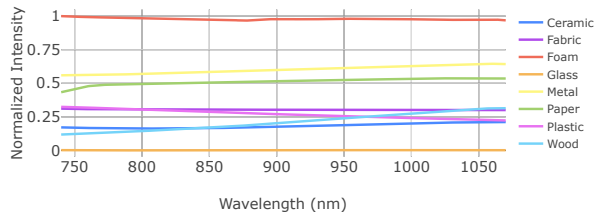

Fig. 5. Example raw spectral samples for each of the objects in Fig. 4.

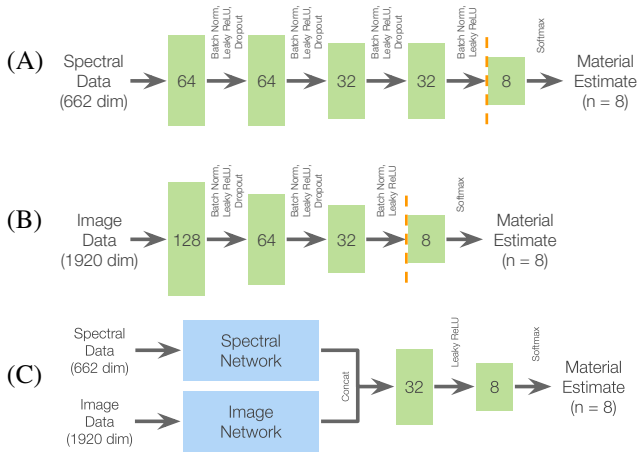

Fig. 6. The spectral, image, and multimodal network architectures. For the multimodal network (C), the spectral and image networks are first pretrained and then trimmed at the dashed orange line shown above.

approach performs better than directly learning a joint representation with early fusion. Overall, our results in Section IV show that combining modalities improves generalization to recognize materials of unseen objects, with close-range texture imaging and near-field spectroscopy providing strong individual baselines.

IV. Evaluation

Our dataset contains 14,400 spectral and image measurements from 144 distinct household objects. Prior to training and hyperparameter optimization for the models defined in Section III-C, we split these data into a training set of measurements from 104 objects, and a heldout test set of 40 objects (5 objects per each of the 8 material categories). This heldout data was not used for optimizing our models' hyperparameters. This heldout test set also includes the same test set objects used in [2], shown in Fig. 7, for a direct comparison to prior work that used only idealized spectral measurements[2]. To reduce the influence of random weight initialization when training models, we report all results averaged over 10 random seeds.

[2]Idealized measurements are collected with flat material objects to block out environmental light and reduce noise in spectral measurements.

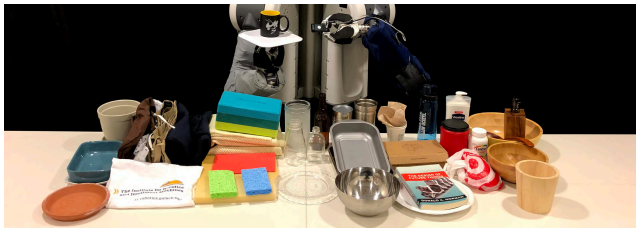

Fig. 7. The 40 heldout test set objects.

A. Recognizing Materials of New Objects

When deployed in real-world environments, robots are likely to encounter new objects which they have not yet been exposed to. Similar to prior works in material classification, we begin by evaluating our multimodal sensing approach when recognizing the materials of new objects not found in the training data [2], [19]. We first assess generalization across all 104 training set objects using leave-one-object-out cross-validation. To do so, we train a model on 103 objects (10,300 measurements) and evaluate material classification accuracy on the 100 samples from the one left-out object. We then repeat this process for each object and compute the average accuracy over the 104 splits.

As shown in Table I, when using only spectral measurements with our spectral model (model A), we achieved an accuracy of 65.1% averaged over 10 random seeds. When training on visual data, our image model (model B) achieved a material classification accuracy of 70.5%. In comparison, our multimodal approach (model C) achieved an accuracy of 74.2%, a ∼4% improvement using low-dimensional representations of both image and spectral samples.

Prior research has investigated how a robot can use near-infrared spectroscopy to recognize object materials with leave-one-object-out cross-validation over five material categories: fabric, metal, paper, plastic, and wood [2]. For a direct comparison of results, we evaluate our performance given only these materials (excluding ceramic, foam, and glass objects). Table I also depicts the performance of our models over these five materials (65 objects, 13 objects per material). Notably, our model trained on SCiO spectral measurements achieved 79.1% accuracy, which is identical to the 79.1% leave-one-object-out accuracy presented in prior work that used flat material objects [2].

As a final assessment of how spectroscopy and texture imaging enables generalizing material classification to new objects, we evaluate results over the heldout test set consisting of five objects from each material category. Prior work has evaluated how a model trained on idealized spectral measurements from flat objects can be used to recognize the materials of 25 household objects from five material categories (fabric, metal, paper, plastic, and wood) [2]. We

TABLE I

LEAVE-ONE-OBJECT-OUT ACCURACY WITH ALL 8 MATERIALS AND THE 5 MATERIALS FROM [2].

	Spectral (A)	Image (B)	Multimodal (C)
5 Materials	**79.1**	76.8	**79.1**
8 Materials	65.1	70.5	**74.2**

TABLE II

ACCURACY OVER THE HELDOUT TEST SET, WITH ALL 8 MATERIALS AND THE 5 MATERIALS FROM [2].

	Spectral (A)	Image (B)	Multimodal (C)
5 Materials	85.9	80.1	**90.8**
8 Materials	77.2	69.6	**80.0**

include these same objects in our heldout test to enable a direct comparison with our multimodal sensing approach. By training a multimodal model on both spectral and texture images from the training set (65 objects from five materials), our resulting model recognizes the materials of the 25 heldout test objects with 90.8% accuracy, as shown in Table II. When generalizing to new household objects, this is a ∼9% improvement compared to the 81.6% accuracy achieved in [2], which trained a neural network model on only spectral measurements from flat material samples, rather than from household objects. When compared to the leave-one-object-out results, we note that our multimodal approach performs significantly better on the paper and plastic heldout objects, 98.3% and 77.6% accuracy respectively, leading to higher overall performance on the heldout dataset.

B. Spectral vs. Image Sensing

In this section, we provide insight and case studies into what materials the two sensory modalities (spectral and image) perform best with and how a multimodal network architecture can leverage the strengths of each modality.

Fig. 8 shows how our models trained on different modalities performed across material categories during leave-one-object-out cross-validation. We observe that it is easier to recognize fabrics with visual texture information, yet easier to recognize paper and glass with spectral data. Furthermore, some materials, such as plastic, remain difficult for both spectral and image data, in part due to large variation among plastic objects and difficulty distinguishing translucent plastics from glass. In addition, we observe that in many cases, a multimodal model that leverages both spectral and visual data can more accurately recognize materials than when using either modality independently. One example of this occurring is with foam objects, where the spectral and image models achieved 48.2% and 53.7% accuracy, respectively, yet our multimodal model attained 64.6% accuracy, ∼16% higher than the spectral model and ∼11% higher than the image modality.

Beyond averages over entire material categories, we also investigate examples of specific objects and how the different modalities compare, as shown in Table III. A gray

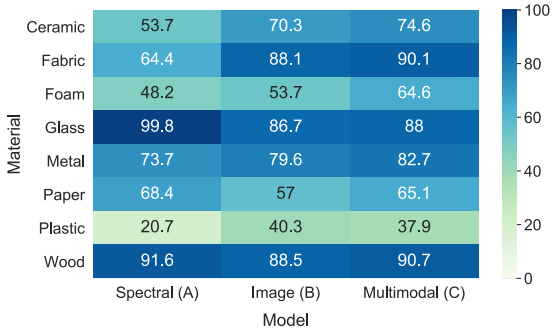

Fig. 8. Leave-one-object-out accuracy for each material.

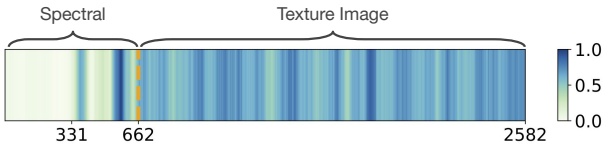

Fig. 9. A saliency map from our multimodal model for a single spectral and image sample of the wood tray object. Gradients of the model output are backpropagated to the input vector to compute this saliency map. The range between 331 and 662 represents the difference quotient (derivative) of the spectral measurement.

cotton fabric shirt was challenging for our spectral model to classify, achieving only 2.4% accuracy. In comparison, the image modality recognized this object as fabric with 99.7% accuracy. Our multimodal model also achieved 99.7% accuracy by learning to leverage the visual information to make its decision. Conversely, our image model struggled to accurately classify a foam plate with only 43.3% accuracy. By incorporating spectral data, our multimodal model correctly recognizes the foam plate with 99.4% accuracy.

As indicated in the previous examples, a multimodal model frequently matches or outperforms models trained on independent modalities. Another example of this is a wood tray for which the spectral and texture image models reach 0.2% and 63.6% accuracy, respectively. Yet our multimodal model recognizes this object as wood with 78.4% accuracy, a ∼15% improvement over the image model. Fig. 9 shows a saliency map from our multimodal model (visualization of what input features would affect the material estimate most if changed) [70] for a single measurement from the wood tray. As depicted, our multimodal model uses the entire image modality to make its classification, but also uses a small portion of the spectral data to further improve its estimate.

A few limitations still remain with using a multimodal network architecture. Namely, there are instances where

TABLE III

CASE STUDIES OF OBJECTS WITH LEAVE-ONE-OBJECT-OUT ACCURACY. IMAGES OF EACH OBJECT ARE SHOWN IN FIG. 4.

Object	Spectral (A)	Image (B)	Multimodal (C)
Fabric gray shirt	2.4	99.7	99.7
Foam plate	100.0	43.3	99.4
Wood tray	0.2	63.6	78.4
Plastic coffee-mate	0.6	81.5	50.5
Paper tissue box	100.0	15.8	48.2

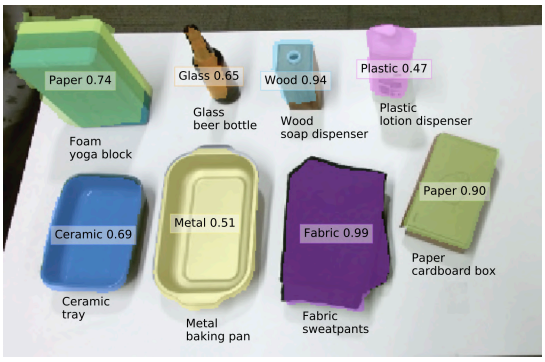

Fig. 10. Demonstration of object material recognition. (Left) Table scene of objects. (Right) Pixel-level material classification with accompanying prediction probabilities using our multimodal approach.

TABLE IV

8 MATERIAL LEAVE-ONE-OBJECT-OUT ACCURACY (MODEL B), RESIZING OR CENTER CROPPING IMAGE INPUT FOR DENSENET-201 FEATURES.

Image Preprocessing	Accuracy
(320×240) resize	**70.5**
(320×240) crop	61.6
(640×480) resize	69.0
(640×480) crop	65.7
(1280×960) resize	66.2

TABLE V

LEAVE-ONE-OBJECT-OUT ACCURACY (MODEL B), COMPARING FEATURES FROM IMAGENET MODELS. (320×240) RESIZED IMAGES.

Network	Accuracy
VGG19 [71]	63.7
ResNet-50 [72]	66.2
ResNet-101 [72]	67.4
ResNet-152 [72]	66.0
DenseNet-201 [65]	**70.5**
ResNeXt-101 [73]	68.7
[3]EfficientNet-B5 [74]	69.4

using a model trained on either spectral or image modalities independently performs better than a model trained on both modalities. This phenomenon occurs with both the plastic coffee-mate container and paper tissue box objects during leave-one-object-out cross-validation. For the plastic coffee-mate, texture imaging alone achieved 81.5% accuracy over 10 random seeds. Yet, our multimodal model only recognized this object as plastic 50.5% of the time. Similarly, our spectral model recognized the paper tissue box with 100.0% accuracy, yet when combined with image data, our multimodal model only classified this as paper 48.2% of the time. Upon inspection, we observed that the image model incorrectly classified 83.3% of paper tissue box images as plastic. We see these tissue box images often contain text similar to labels on commercial plastic objects in the dataset. These inaccuracies may be due in part to our dataset size, which may not fully capture the variation of materials across real-world objects. Increasing the number and variety of objects in the training set may further improve performance when generalizing material classification to new objects.

C. Texture Image Features

In Section III, we presented a neural network architecture for material classification with visual texture data, dependent on preprocessing images with a DenseNet-201 model trained on ImageNet. Prior research has indicated that models trained on ImageNet have a high prior for recognizing texture features within an image [60]. In this section, we evaluate different texture image preprocessing techniques and compare ImageNet-trained models for texture representation.

The original resolution of the captured close-range texture images is 1600×1200. CNN models are usually trained on ImageNet using center crops of 224×224. We test different resolutions for our texture images by center crop or resize, using DenseNet-201 as a feature extractor. Results are shown in Table IV for different raw image preprocessing techniques prior to computing the DenseNet-201 features, evaluated on 8 material leave-one-object-out material classification with the texture image model (model B). We find that resizing performs better than center cropping, suggesting that textural features are better captured with more visual surface area and context of the object, rather than a small but dense visual sample. Additionally, we observe that resizing to 320×240, which is near the image resolution that the CNN was trained at, performs better than resizing to higher resolutions.

We generate low-dimensional visual features from common ImageNet-trained CNNs using texture images that were resized[3] to a resolution of 320×240. Table V compares several ImageNet models for computing texture representations, which are then used to train the texture image model (model B) during leave-one-object-out cross-validation on all 8 materials. We observe that performance on ImageNet is loosely correlated with material classification accuracy with our image model. Further advances on CNN models benchmarked by ImageNet may continue to improve texture representation. Due to the architecture of our multimodal network, which learns separate and combined representations, advances in texture representation should lead to improvements on material classification with texture images.

D. Table Scene Recognition

We further evaluate our multimodal sensing approach by classifying materials of a scene of objects placed on a table,

[3]Texture images were resized to 608×456 for EfficientNet-B5, near its native ImageNet input resolution.

similar to what may be observed in a kitchen or home environment. We place one object from each material category from the heldout dataset on a table in front of the PR2. Using a 3D point cloud from its head-mounted Kinect, the PR2 segments objects from the table and defines pixel-level clusters in the 2D visual image for each distinct object found in the point cloud. The robot then classifies the material of each cluster using spectral and close-range texture image measurements from each object. To capture a measurement, the PR2 moves its left end effector to a position just in front of each object, matching a surface normal for the object computed from the point cloud. Fig. 10 shows a table setup with the PR2 and the pixel-level classification of each object using predictions from our multimodal material classification model. Our model correctly recognized the materials for seven of the eight heldout objects, missing only the foam yoga block. This demonstration can be seen in greater detail in the supplementary video.

V. Conclusion

This paper introduces a multimodal sensing technique that combines near-infrared spectroscopy and close-range high resolution texture imaging for enabling robots to accurately classify the materials of household objects. We present and evaluate a new dataset of spectral measurements and high-resolution texture images for 144 household objects from 8 material categories. Compared to prior work in material classification with spectroscopy, our multimodal approach achieved 9% higher accuracy when generalizing to new, unseen household objects. In addition, we demonstrate how this sensing technique enables a robot to recognize materials across a scene of objects on a table, without physical contact with the objects. Through this work, we have shown that near-infrared spectroscopy and texturing imaging offers a reliable and accurate multimodal sensing approach for robots to estimate the materials of objects.

References

[1] G. Buckingham, J. S. Cant, and M. A. Goodale, "Living in a material world: how visual cues to material properties affect the way that we lift objects and perceive their weight," *Journal of Neurophysiology*, vol. 102, no. 6, pp. 3111–3118, 2009.

[2] Z. Erickson, N. Luskey, S. Chernova, and C. Kemp, "Classification of household materials via spectroscopy," *IEEE Robotics and Automation Letters*, vol. PP, pp. 1–1, 01 2019.

[3] T. Bhattacharjee, J. M. Rehg, and C. C. Kemp, "Haptic classification and recognition of objects using a tactile sensing forearm," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 4090–4097.

[4] S. Decherchi, P. Gastaldo, R. Dahiya, M. Valle, and R. Zunino, "Tactile-data classification of contact materials using computational intelligence," *IEEE Transactions on Robotics*, 2011.

[5] E. Kerr, T. M. McGinnity, and S. Coleman, "Material classification based on thermal properties—a robot and human evaluation," in *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2013, pp. 1048–1053.

[6] T. Bhattacharjee, J. Wade, and C. Kemp, "Material recognition from heat transfer given varying initial conditions and short-duration contact," in *Proceedings of Robotics: Science and Systems*, 2015.

[7] Y. Cho, N. Bianchi-Berthouze, N. Marquardt, and S. J. Julier, "Deep thermal imaging: proximate material type recognition in the wild through deep learning of spatial surface temperature patterns," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[8] H. Alagi, A. Heiligl, S. E. Navarro, T. Kroegerl, and B. Hein, "Material recognition using a capacitive proximity sensor with flexible spatial resolution," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 6284–6290.

[9] 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.

[10] C. Fang, D. Wang, D. Song, and J. Zou, "Toward fingertip non-contact material recognition and near-distance ranging for robotic grasping," in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 4967–4974.

[11] H.-S. Yeo, G. Flamich, P. Schrempf, D. Harris-Birtill, and A. Quigley, "Radarcat: Radar categorization for input & interaction," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016, pp. 833–841.

[12] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, T. Darrell, and K. J. Kuchenbecker, "Robotic learning of haptic adjectives through physical interaction," *Robotics and Autonomous Systems*, vol. 63, pp. 279–292, 2015.

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

[14] D. Xu, G. E. Loeb, and J. A. Fishel, "Tactile identification of objects using bayesian exploration," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 3056–3061.

[15] E. Kerr, T. M. McGinnity, and S. Coleman, "Material recognition using tactile sensing," *Expert Systems with Applications*, 2018.

[16] L. Chin, J. Lipton, M. C. Yuen, R. Kramer-Bottiglio, and D. Rus, "Automated recycling separation enabled by soft robotic material classification," in *2019 2nd IEEE International Conference on Soft Robotics (RoboSoft)*. IEEE, 2019, pp. 102–107.

[17] D. S. Chathuranga, V. A. Ho, and S. Hirai, "Investigation of a biomimetic fingertip's ability to discriminate fabrics based on surface textures," in *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. IEEE, 2013, pp. 1667–1674.

[18] J. Sinapov, C. Schenck, K. Staley, V. Sukhoy, and A. Stoytchev, "Grounding semantic categories in behavioral interactions: Experiments with 100 objects," *Robotics and Autonomous Systems*, 2014.

[19] Z. Erickson, S. Chernova, and C. C. Kemp, "Semi-supervised haptic material recognition for robots using generative adversarial networks," in *Conference on Robot Learning*, 2017, pp. 157–166.

[20] T. Bhattacharjee, H. M. Clever, J. Wade, and C. C. Kemp, "Multimodal tactile perception of objects in a real home," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2523–2530, 2018.

[21] K. Zhang, M. Sharma, M. Veloso, and O. Kroemer, "Leveraging multimodal haptic sensory data for robust cutting," in *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, 2019.

[22] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz, "Exploring features in a bayesian framework for material recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 239–246.

[23] D. Hu, L. Bo, and X. Ren, "Toward robust material recognition for everyday objects." in *BMVC*, vol. 2. Citeseer, 2011, p. 6.

[24] A. Dimitrov and M. Golparvar-Fard, "Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections," *Advanced Engineering Informatics*, vol. 28, no. 1, pp. 37–49, 2014.

[25] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3479–3487.

[26] G. Schwartz and K. Nishino, "Recognizing material properties from images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[27] S. Su, F. Heide, R. Swanson, J. Klein, C. Callenberg, M. Hullin, and W. Heidrich, "Material classification using raw time-of-flight measurements," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3503–3511.

[28] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4d light-field dataset and cnn architectures for material recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 121–138.

[29] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a gelsight tactile sensor," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 951–958.

[30] R. Li and E. H. Adelson, "Sensing and recognizing surface textures using a gelsight sensor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1241–1247.

[31] A. Yamaguchi and C. G. Atkeson, "Implementing tactile behaviors using fingervision," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, 2017, pp. 241–248.

[32] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[33] L. Cramphorn, J. Lloyd, and N. F. Lepora, "Voronoi features for tactile sensing: Direct inference of pressure, shear, and contact locations," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2752–2757.

[34] W. Chen, H. Khamis, I. Birznieks, N. F. Lepora, and S. J. Redmond, "Tactile sensors for friction estimation and incipient slip detection-toward dexterous robotic manipulation: A review," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9049–9064, 2018.

[35] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 1045–1051.

[36] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2722–2727.

[37] J.-T. Lee, D. Bollegala, and S. Luo, "touching to see and seeing to feel: Robotic cross-modal sensory data generation for visual-tactile perception," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4276–4282.

[38] C. Kampouris, S. Zafeiriou, A. Ghosh, and S. Malassiotis, "Fine-grained material classification using micro-geometry and reflectance," in *European Conference on Computer Vision*. Springer, 2016.

[39] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, 2017.

[40] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 536–543.

[41] C. Pasquini, "Near infrared spectroscopy: a mature analytical technique with new perspectives - a review," *Analytica Chimica Acta*, 2018.

[42] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, and N. Jent, "A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies." *Journal of Pharmaceutical and Biomedical Analysis*, vol. 44 3, pp. 683–700, 2007.

[43] V. Bellon, J. L. Vigneau, and F. Svila, "Infrared and near-infrared technology for the food industry and agricultural uses: on-line applications," *Food Control*, vol. 5, no. 1, pp. 21 – 27, 1994.

[44] H. Masoumi, S. M. Safavi, and Z. Khani, "Identification and classification of plastic resins using near infrared reflectance spectroscopy," *International Journal of Mechanical and Industrial Engineering*, 2012.

[45] R. A. Crocombe, "Portable spectroscopy," *Applied Spectroscopy*, vol. 72, no. 12, pp. 1701–1751, 2018.

[46] G. Rateni, P. Dario, and F. Cavallo, "Smartphone-based food diagnostic technologies: A review," *Sensors*, vol. 17, p. 1453, 06 2017.

[47] H. Yan and H. W. Siesler, "Quantitative analysis of a pharmaceutical formulation: Performance comparison of different handheld near-infrared spectrometers," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 160, pp. 179 – 186, 2018.

[48] A. Das, A. Wahi, I. Kothari, and R. Raskar, "Ultra-portable, wireless smartphone spectrometer for rapid, non-destructive testing of fruit ripeness," *Scientific Reports*, vol. 6, p. 32504, 09 2016.

[49] S. Lee, T. G. Noh, J. H. Choi, J. Han, J. Y. Ha, J. Y. Lee, and Y. Park, "Nir spectroscopic sensing for point-of-need freshness assessment of meat, fish, vegetables and fruits," in *Sensing for Agriculture and Food Quality and Safety IX*, vol. 10217. International Society for Optics and Photonics, 2017, p. 1021708.

[50] A. Kartakoullis, J. Comaposada, . Cruz-Carrin, X. Serra, and P. Gou, "Feasibility study of smartphone-based near infrared spectroscopy (nirs) for salted meat composition diagnostics at different temperatures," *Food Chemistry*, vol. 278, 11 2018.

[51] W. Liu, A. Daruna, and S. Chernova, "Cage: Context-aware grasping engine," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

[52] L. Nair, N. Shrivatsav, Z. Erickson, and S. Chernova, "Autonomous tool construction using part shape and attachment prediction," *Proceedings of Robotics: Science and Systems*, 2019.

[53] N. Shrivatsav, L. Nair, and S. Chernova, "Tool substitution with shape and material reasoning using dual neural networks," *arXiv preprint arXiv:1911.04521*, 2019.

[54] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, "From bow to cnn: Two decades of texture representation for texture classification," *International Journal of Computer Vision*, 2019.

[55] P. Napoletano, "Hand-crafted vs learned descriptors for color texture classification," in *International Workshop on Computational Color Imaging*. Springer, 2017, pp. 259–271.

[56] G. Kalliatakis, G. Stamatiadis, S. Ehsan, A. Leonardis, J. Gall, A. Sticlaru, and K. D. McDonald-Maier, "Evaluating deep convolutional neural networks for material classification," *arXiv preprint arXiv:1703.04101*, 2017.

[57] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 262–270.

[58] T.-Y. Lin and S. Maji, "Visualizing and understanding deep texture representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2791–2799.

[59] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[60] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." in *International Conference on Learning Representations*, 2019.

[61] B. Long and T. Konkle, "The role of textural statistics vs. outer contours in deep cnn and neural responses to objects," in *Conference on Computational Cognitive Neuroscience*, 2018, p. 4.

[62] P. Ballester and R. M. Araujo, "On the performance of googlenet and alexnet applied to sketches," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[63] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[64] T. Strother, "Nir and raman: complementary techniques for raw material identification," *Thermo Fisher Scientific*, 2009.

[65] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[66] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[67] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, 2011, pp. 689–696.

[68] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 681–687.

[69] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," *arXiv preprint arXiv:1805.11730*, 2018.

[70] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[71] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[73] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[74] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.