

Deep Learning for Tactile Understanding From Visual and Haptic Data

Yang Gao¹, Lisa Anne Hendricks¹, Katherine J. Kuchenbecker² and Trevor Darrell¹

Abstract—Robots which interact with the physical world will benefit from a fine-grained tactile understanding of objects and surfaces. Additionally, for certain tasks, robots may need to know the haptic properties of an object before touching it. To enable better tactile understanding for robots, we propose a method of classifying surfaces with haptic adjectives (e.g., compressible or smooth) from both visual and physical interaction data. Humans typically combine visual predictions and feedback from physical interactions to accurately predict haptic properties and interact with the world. Inspired by this cognitive pattern, we propose and explore a purely visual haptic prediction model. Purely visual models enable a robot to “feel” without physical interaction. Furthermore, we demonstrate that using both visual and physical interaction signals together yields more accurate haptic classification. Our models take advantage of recent advances in deep neural networks by employing a unified approach to learning features for physical interaction and visual observations. Even though we employ little domain specific knowledge, our model still achieves better results than methods based on hand-designed features.

I. INTRODUCTION

Tactile understanding is important for a wide variety of tasks. For example, humans constantly adjust their movements based on haptic feedback during object manipulation [21]. Similarly, robot performance is likely to improve on a diverse set of tasks if robots can understand the haptic properties of surfaces and objects. A robot might adjust its grip when manipulating a fragile object, avoid surfaces it perceives to be wet or slippery, or describe the tactile qualities of an unfamiliar object to a human. In this work, we explore methods to classify haptic properties of surfaces. Using our proposed methods for haptic classification, we believe haptic information can be more effectively harnessed for a wide array of robot tasks.

Humans rely on multiple senses to make judgments about the world. A well known example of this multisensory integration is the McGurk effect [27], in which humans perceive different phonemes based on the interaction of visual and auditory cues. Similarly, humans rely on both tactile and visual understanding to interact with their environment. Humans use both haptic and vision to correctly identify objects [36], and fMRI data demonstrates that haptic and visual signals are processed in a multi-sensory fashion

*This material is based upon work supported by DARPA, Berkeley Vision and Learning Center, as well as the U.S. National Science Foundation (NSF) under grants 1426787 and 1427425 as part of the National Robotics Initiative. Lisa Anne Hendricks is supported by the NDSEG.

¹ Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, California, CA 94701, USA {yg, lisa_anne, trevor}@eecs.berkeley.edu

² Haptics Group, GRASP Laboratory, Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. kuchenbe@seas.upenn.edu

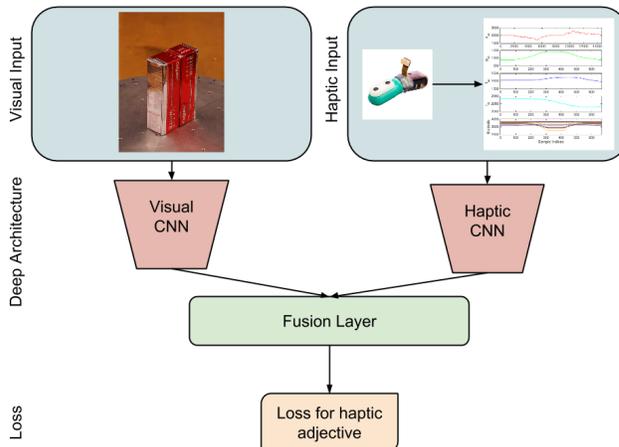


Fig. 1: We propose deep models as a unified approach to learning features for haptic classification of objects. We use both visual and haptic data. In comparison to training models on only haptic data or only visual data, we find that combining learned features across modalities, as pictured above, leads to superior performance. ¹

during object recognition [3]. Motivated by the cross-modal processing inherent in the human brain, we build a model that processes both haptic and visual input and demonstrate that this combination achieves higher performance than using either the haptic or visual input alone.

In order to effectively learn from haptic and visual data, we train deep neural networks, which have led to dramatic improvements in disparate learning tasks such as object recognition [23] and automatic speech recognition [19]. As opposed to hand-designing features for particular modalities, neural networks provide a unified framework to learn features directly from data. With our neural network model, we learn haptic features that outperform previous proposed features with little haptic domain knowledge. Additionally, when training models on visual features, we transfer learned models on the related task of material classification to haptic classification. Consequently, we can train a large visual model with less than 1,000 training instances.

Our contributions are as follows. First, we demonstrate that neural networks serve as a unifying framework for signal classification, allowing us to learn rich features on both visual and haptic data with little domain knowledge. We believe

¹To collect haptic data, we use the BioTac sensor [35]. The image of the sensor in the above figure is courtesy of the product webpage <http://www.syntouchllc.com/Products/BioTac/>.

similar methods can be used to learn models for other signals that robots need to understand. Furthermore, we demonstrate that visual data from a different yet related task, material classification, easily transfers to haptic classification. Finally, we show that haptic and visual signals are complementary, and combining modalities further improves classification.

II. RELATED WORK

A. Previous Approaches to Tactile Understanding

The tactile properties of consumer products have been explored to better understand the quality of items such as fabric [31] and skin cream [25]. Progress in tactile understanding has been driven by both the creation of better mechanical measurement systems ([11], [35]) as well as learning better algorithms for haptic understanding ([11], [12]). A variety of work ([12], [29]) concentrates on classifying specific textures (e.g., *paper*). Additionally, [12] asserts that biomimicry is important for texture recognition. In addition to using a BioTac sensor [35], which is designed to replicate human tactile sensations, [12] reproduces human movement when touching new textures. Unlike these approaches, our system produces predictions of *haptic adjectives*, enabling a qualitative description of previously unexplored textures and objects.

Prior work which focuses on haptic adjective classification includes [17], [8], and [7]. [17] demonstrates that a rich and diverse haptic measurement system that measures temperature, compliance, roughness, and friction is key to accurately discerning between haptic adjectives such as *sticky* and *rough*. Our work most closely resembles Chu et al. ([8], [7]), which detail the collection of haptic classification datasets (PHAC-1 and PHAC-2) and concentrates on classifying objects with binary haptic adjectives. [8] and [7] rely on hand-designed features for haptic classification. Two types of features are proposed: static and dynamic. Static features consist of simple data statistics while dynamic features are learned from fitting haptic signals to HMMs.

The haptic classification problem is closely related to the material classification problem. For example, classifying a surface as *glass* implies the presence of some haptic properties, such as *hard*, *slippery*, *smooth* and *solid*. However, notable exceptions exist. For example, different plastic surfaces have vastly different roughness and hardness properties: a plastic bag is *smooth* and *soft* but a sawed plastic block is *rough* and *hard*. Consequently, haptic classification goes beyond simply identifying object materials. [5] details the collection of a large material classification dataset and then demonstrates that deep models can be used for the material classification problem.

B. Neural Networks

Neural networks have led to state-of-the-art results on many important problems in artificial intelligence ([23], [19]). Neural networks are compositional models which are formed by stacking “layers” such that the output of one layer is the input of the next layer. A layer consists of an affine transformation of layer inputs and learned weights followed

by a nonlinearity. Other operations such as max-pooling and normalization can also be included. Neural networks are trained iteratively using backpropagation. Backpropagation applies the chain rule in order to determine $\frac{dL}{dw_l}$ where w are the weights of layer l and L is the loss function being minimized. At each iteration, the gradients $\frac{dL}{dw_l}$ are used to update the weights in order to minimize loss.

Convolutional neural networks (CNNs) [24] use a convolution as the linear operation within the layer, as opposed to fully connected layers (also called inner product layers) which use inner products as the linear operation within the layer. CNNs have been proposed to decrease the total number of parameters of a model. The reduction in the number of parameters has proven crucial for training large-scale object recognition systems [23] and has also been used for learning dynamics in 1-D time series, such as in speech recognition [1].

For understanding 1-D time series, recurrent networks have also achieved great success on tasks such as language modeling [28] and handwriting recognition [16]. Recurrent networks include hidden units which are updated at each time step based on the values of the hidden units at the previous time step and the current input. Certain types of recurrent networks, such as LSTMs [20] and GRUs [4], include special gates which learn to remember and forget previous states. Exemplary results on tasks such as machine translation [4], automatic speech recognition [15], and parsing [34] demonstrate that networks which include such gates are able to learn complex temporal dynamics, making them a natural fit for modeling time-series data.

In addition to performing better than other methods, neural networks are also advantageous because they require little domain knowledge. This benefit is particularly appealing in the domain of haptics because high-dimensional haptic signals can be difficult for humans to understand. Furthermore, weights learned from one dataset can be transferred to similar datasets through fine-tuning. When fine-tuning, the model is initialized with weights learned on a previous task before training. For example, weights from models trained on the large ImageNet dataset [9] can be transferred to related tasks such as semantic segmentation [26], scene classification [37] and object detection [13]. Pre-trained networks can also be used as off-the-shelf feature extractors by using the outputs (activations) of one of the layers of the network as a feature. When a network has been pre-trained on a related task, using the network as a feature extractor leads to good results on tasks such as subcategory recognition and scene classification [10].

Multimodal networks have been proposed in both unsupervised ([30]) and supervised ([18], [22]) settings. Both [18] and [22] first train deep models on individual data modalities then use activations from these models to train a multimodal classifier.

III. DEEP HAPTIC CLASSIFICATION MODELS

Deep learning provides a unified framework for learning models for classification. We describe models that are able

to achieve high performance on the haptic classification task using haptic data, visual data, and in a multimodal setting.

A. Haptic Signal

We explore both CNN and LSTM models for haptic classification.

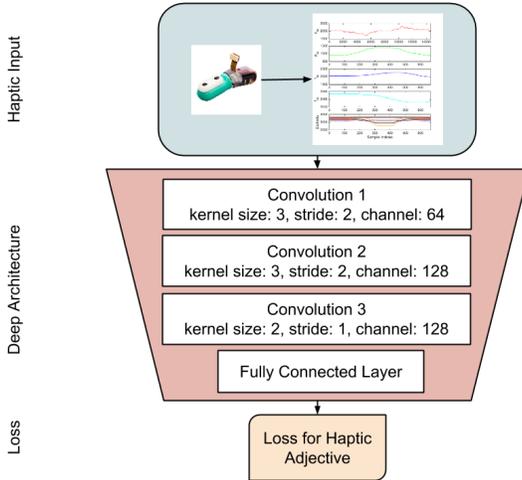


Fig. 2: Haptic CNN structure.

1) *Haptic CNN Model*: Our proposed CNN model performs temporal convolutions on an input haptic signal, similar to models which have previously shown good results on other tasks with 1-D signals, such as speech recognition [1]. CNNs decrease the total number of parameters in the model, making it easier to train with limited training data.

Fig. 2 shows the design of our CNN. Our haptic signal includes 32 haptic measurements (see IV-B for details). To form the input of our network, we concatenate signals along the *channel* axis as opposed to the *time* axis. Thus, concatenating two signals with dimension $T \times C$ where T is the number of time steps and C is the number of channels, results in a $T \times 2C$ signal. After each convolution, we use a rectified linear unit (ReLU) nonlinearity.

To further reduce the number of parameters, we use “grouping”. Grouping for a certain layer requires that groups of channels do not interact with other groups. We use groups of 32 in all our convolutional layers, which means interactions between different haptic signals are not learned until the fully connected layer. Generally, allowing cross-channel interactions earlier in the model is beneficial, and we anticipate better results without grouping when using a larger training dataset.

Our initial classification models are trained using logistic loss. However we find that after learning the weights for the convolutional layers, fine-tuning with hinge-loss obtains similar or slightly better results for all models. Unless otherwise mentioned, all reported results are from a network trained with a hinge loss.

2) *Haptic LSTM Model*: In addition to designing a CNN for haptic classification, we also explore LSTM models. Because LSTM models have a recurrent structure, they are

a natural fit for understanding haptic time-series signals. Our LSTM structure consists of 10 recurrent units, and is followed by a fully connected layer with 10 outputs and a ReLU nonlinearity. A final fully connected layer produces a binary prediction. We found that small changes in the LSTM structure (e.g. number of recurrent units) did not result in large changes in final classification scores. Though stacking LSTMs generally leads to better results [15], this led to performance degradation for haptic classification.

B. Visual CNN Model

In addition to the haptic signal, we also use visual cues for haptic classification. Since the haptic classification problem is closely related to material classification, we transfer weights from a CNN that is fine-tuned on the Materials in Context Database (MINC) [5]. MINC is a material recognition dataset, which consists of 23 classes (such as *brick* and *carpet*) and uses the GoogleNet architecture [33]. We refer to the fine-tuned MINC network as MINC-CNN.

To fine-tune a haptic classification model, we transfer weights for all layers below the “inception (5a)”² layer of the MINC-CNN. We find that placing an average-pooling layer and L2 normalization layer after the “inception (5a)” and before a loss layer yields best results. Fig. 3 summarizes the visual haptic adjective classification pipeline.

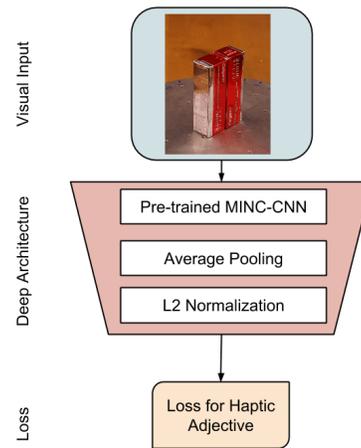


Fig. 3: Visual CNN structure.

C. Multimodal Learning

Figure 1 shows the structure of our multimodal model. The top left and top right denote the visual and haptic inputs respectively. The learned weights from visual and haptic CNN models are transferred to the multimodal network. Activations from the 3rd convolutional layer (conv3) of the haptic CNN and the activations after the L2-normalization of the inception (5a) layer of the visual CNN are concatenated. The final multimodal classification model is trained with a

²The GoogleNet architecture is built from “inception” modules which consist of convolutions and filter concatenations. Please see [33] for more details.

hinge loss. Note that when forming our multimodal network, we directly transfer weights from our previously trained haptic and visual networks, and only learn the classification layer.

D. Training method

We use standard backpropagation to train our networks. We use SGD with a constant learning rate of 0.01, momentum of 0.9 and initialize model weights with “xavier” initialization [14]. We train our network for 200 epochs with a batch size of 1000 (roughly $\frac{1}{5}$ of the entire training dataset).

IV. EXPERIMENTAL SETUP

A. PHAC-2 Dataset

We demonstrate our approach on the Penn Haptic Adjective Corpus 2 (PHAC-2) dataset, which appears in [7]. The PHAC-2 dataset contains haptic signals and images of 53 household objects. Each object is explored by a pair of SynTouch biomimetic tactile sensors (BioTacs), which are mounted to the gripper of a Willow Garage Personal Robot 2 (PR2). Because the BioTac sensor mimics humans tactile capabilities, we believe its signals provide the rich haptic data necessary for fine-grained tactile understanding. Each object is felt with the following four exploratory procedures (EPs): *Squeeze*, *Hold*, *Slow Slide*, *Fast Slide*, which mimic how humans explore the tactile properties of objects. The BioTac sensor generates five types of signals: low-frequency fluid pressure (P_{DC}), high-frequency fluid vibrations (P_{AC}), core temperature (T_{DC}), core temperature change (T_{AC}), and 19 electrode impedance ($E_1 \dots E_{19}$) which are spatially distributed across the sensor (example signals in Fig. 4). The P_{AC} signal is sampled at 2200 Hz, and the other signals are all sampled at 100 Hz. Ten trials of each EP are performed per object. Although data on the joint positions and gripper velocity and acceleration are available, we concentrate on classifying the tactile signals.

The dataset also contains high resolution (3000×2000) images of each object from eight different viewpoints (see Fig 5). The objects are placed in the center of an aluminum plate. Although lighting conditions are not explicitly controlled, the variations are insignificant.

Each object is described with a set of 24 binary labels (Table I), corresponding to the existence or absence of each of the 24 haptic adjectives (e.g., *slippery* or *fuzzy*). For example, the toothpaste box is given the classification *smooth*. The binary label for a given object was determined by a majority vote of approximately twelve human annotators [7].

TABLE I: 24 haptic adjectives

absorbent	bumpy	compressible	cool	crinkly	fuzzy
hairy	hard	metallic	nice	porous	rough
scratchy	slippery	smooth	soft	solid	springy
squishy	sticky	textured	thick	thin	unpleasant

B. Data Preprocessing

1) *Haptic Signal: Normalization* For each signal $s \in \{P_{AC}, P_{DC}, T_{AC}, T_{DC}, E_1, \dots, E_{19}\}$, we calculate the mean \bar{s} and the standard deviation σ , and normalize the signal by $s' = \frac{s - \bar{s}}{\sigma}$.

Subsampling Because P_{AC} is sampled at a higher rate than other signals, we first downsample it to 100 Hz to match the sampling rate of other signals. The durations of signals for different objects are almost identical for the EPs *Hold*, *Slow Slide*, and *Fast Slide* but the signal length of *Squeeze* varies considerably among objects. To resolve the disparity in signal lengths, we downsample signals for each EP to a fixed length of 150 for simplicity. Though our subsampling scheme possibly discards important information, reducing data dimensionality is crucial for training deep models with such a small dataset.

PCA on Electrode Impedances There are 19 electrode impedances spatially distributed across the core of BioTac. We found that 4 principal components capture 95% of the variations of $E_1 \dots E_{19}$. Consequently, we transform the 19 dimensional electrode impedance signals to the first four coefficients of the principal components. PCA is done independently on each EP across all objects. Because we have four other signals ($P_{AC}, P_{DC}, T_{AC}, T_{DC}$), four electrode signals after PCA, and four EPs, we have a total of $(4 + 4) \times 4 = 32$ haptic signals which we use as input to our models.

Training Data Augmentation The total number of training instances is only 530 (53 objects with 10 trials each). We use data augmentation to avoid substantial overfitting during model training. To augment our data, we treat the two BioTac sensors on PR2 as two distinct instances. Furthermore, we sample five different signals by sub-sampling each signal at a different starting point. After data augmentation, we increase the number of total instances to 5,300.

2) *Visual Signal:* We follow standard image preprocessing steps and subtract the mean values from the RGB image and resize our image to the fixed input size of the MINC-CNN network (224×224). Instead of using the entire image, we extract a central crop which includes the object. The central crop is obtained by first detecting the circular aluminum plate by its color, then estimating the plate’s center and radius R . Our final crop is a rectangular region of size $2R \times R$ with the center of the crop above the center of the circular plate by R . More generally, the crop could be obtained by a general bounding box detection method, or simply by focusing the camera on the surface of interest. Fig.5a shows examples of images cropped with this method.

C. Combining Data Instances

Combining separate instances from a single sample is a common method to boost classification results. For example, in object recognition, it is common to augment data by using different image crops and mirroring images. Better test classification can be achieved by averaging model outputs across separate instances for a single image.

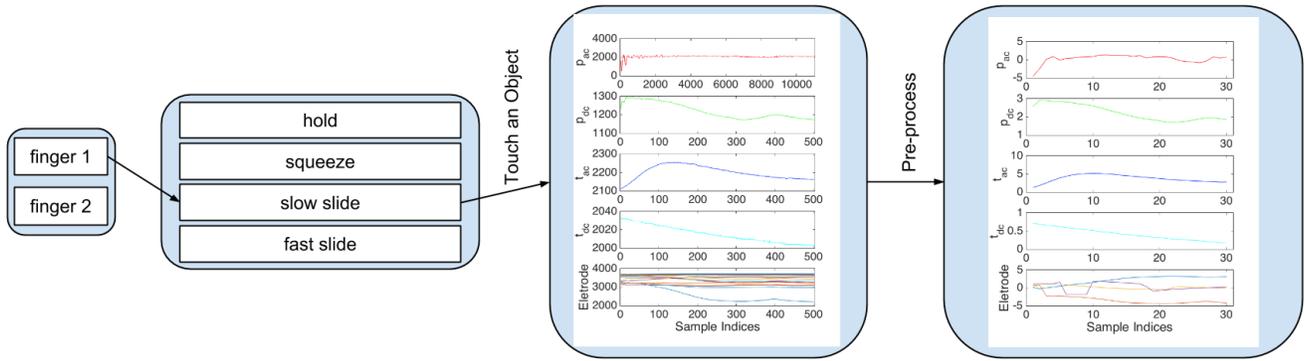
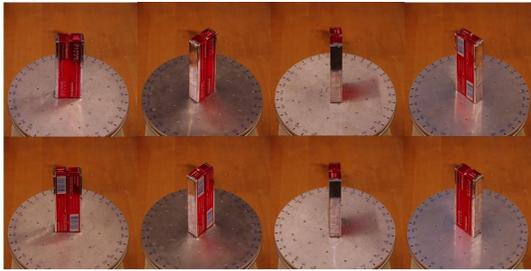


Fig. 4: A sample BioTac signal of the first finger in the *slow slide* exploratory procedure. Both raw signal and preprocessed signals are shown. The signals are p_{AC} , p_{DC} , t_{AC} , t_{DC} and *electrode impedance* respectively from top to bottom.



(a) Examples of images automatically cropped by our algorithm described in IV-B.2.



(b) Each object is photographed from 8 different viewpoints.

Fig. 5: Visual data.

Likewise, we explore combining the ten haptic explorations of each object and the eight visual viewpoints for each object. However, instead of averaging model outputs at test time, we concatenate activations (conv3 for haptic signals and inception (5a) for visual signals) for different instances and retrain the network loss layer.

D. Train/Test Splits

Following [7], [8], we partition the 53 objects into 24 90/10 train/test splits, one for each adjective. Creating 24 separate train/test splits is necessary because there exists no one split such that each adjective is present in both the train and test split. Note that although a single object has several corresponding measurements, we do not allow the same object to appear in both the train and test split.

E. Performance Metric

To measure the performance of the approaches proposed, we adopt the widely used Area Under Curve (AUC) met-

ric. AUC measures the area under the Receiver Operating Characteristic curve, which takes both the true positive rate and the false positive rate into consideration. The previous methods report F1 score [7], [8], which is biased towards true positives [32]. To provide a fair evaluation, we reproduce the baseline methods and report AUC scores.

V. EVALUATION

In Table II we compare the AUC scores from our proposed methods. We determine AUC scores for each adjective and report the average over AUC scores for all adjectives. We report models trained on the haptic, visual, and multimodal data sources as well as models trained by combining features from different trials on a single object (see section IV-C). To differentiate our models, we use the following notation: the prefix for each model denotes the data source on which it is trained, and the suffix (if present) denotes how many instances are combined for classification. For example, “Haptic-ConvNet-10” is a CNN model trained on haptic data and concatenating the conv3 activations from 10 haptic trials before classification.

Models A-F are SVMs trained with the shallow features from [7], [8]³. Models G-J and K-L are our deep haptic and deep visual models. Models M-N are our multimodal models, and model O is the result for a model that randomly guesses.

When comparing the features from previous work, we find that the static features (methods A, B) achieve a slightly higher AUC than dynamic features (methods C, D). Combining static and dynamic features (method F) increases the AUC by 0.9.

Though the hand-designed features perform well, our best deep CNN trained on haptic data (model I) improves the AUC score by 5.0. In model H, we augment our data by treating trials from the two BioTac as separate training instances and by subsampling each signal starting a different time step and using each sample as a separate training instance. This leads to a tenfold increase in the size of the training set (2

³Code for our reproduction of this result and further evaluation details, including comparison of reproduced F1 numbers, are available at <http://www.cs.berkeley.edu/~yg/icra2016/>

signals from each BioTac sensor and 5 subsampled signals). However, even though there is more available training data, we achieve better results by concatenating features from both BioTac recordings and each subsampled signal (model D). One explanation is that examples from each BioTac sensor and different subsampling schemes are extremely similar, causing the model to overfit. However, combining haptic signals for all trials for a given object (model J) results in slightly lower AUC compared to model I. This could be because haptic signals from various trials have informative differences that enable the model to generalize better to signals on new objects. We see the same general trend when using the “shallow” static and dynamic baseline features. We also classify haptic signals by using our LSTM model (model G). Though LSTMs generally are quite good at learning complex temporal dynamics, in the haptics domain they perform considerably worse than the haptic CNN model. Variations of LSTM, such as changing the inputs to the frequency domain, could potentially improve the results. We leave further investigations to future work.

Our visual models also perform well, though performance is 6.0 points below our best haptic model. In contrast to our haptic models, we find that combining features across multiple viewpoints is essential for good haptic classification and improves AUC from 71.5 to 77.2. For robotic applications, this implies that robots need to view objects from varied viewpoints in order to gain an understanding of haptic properties.

We find that our multimodal models perform the best by well over 2.7 points. When concatenating activations from our best haptic model (I) and best visual model (L), we achieve an AUC of 84.7. However, by concatenating activations from model J and model L, we achieve an AUC of 85.9, even though the corresponding haptic only model performs slightly worse.

TABLE II: Comparison of Haptic Classification Methods. See section V for details.

	Methods	AUC (%)
A	Haptic-static-1 trial	77.3
B	Haptic-static-10 trials	70.8
C	Haptic-dynamic-1 trial	75.5
D	Haptic-dynamic-10 trials	74.3
E	Haptic (Combine model A + C)	77.5
F	Haptic (Combine model A + D)	78.2
G	Haptic-LSTM	72.1
H	Haptic-ConvNet	81.6
I	Haptic-ConvNet-1 trial	83.2
J	Haptic-ConvNet-10 trials	82.1
K	Image-GoogleNet-1 view	71.5
L	Image-GoogleNet-8 views	77.2
M	Multimodal (Combine model I + L)	84.7
N	Multimodal (Combine model J + L)	85.9
O	Random guess	50.0

We release our code in order to make our results easily replicable⁴. Additionally, by releasing our models, we hope

⁴Code and data are available at <http://www.cs.berkeley.edu/~yg/icra2016/>.

our haptic classifiers can be integrated into current robotic pipelines that could benefit from better tactile understanding of real objects and surfaces.

VI. DISCUSSION

In order to gain a more intuitive understanding of our model, we present analysis to illustrate what our models are capable of learning.

A. Adjective Prediction Using Different Models

In order to compare the types of predictions produced by our models, we randomly select three objects out of the 53 objects in the data set and examine haptic predictions for each object.

Table III details the predicted adjective for three objects: *shelf liner*, *placemat*, and *furry eraser* (see Fig 5a). We observe that the haptic classifier tends to have high recall, predicting many adjectives for each class. In contrast, the visual classifier is more conservative and appears to have higher precision. For the three examples shown, the visual model predicts no false positives.

The prediction of the multimodal classifier is not a simple union of the haptic and visual classifiers, but rather combines the two sources of information sensibly. For example, the prediction of shelf liner contains the correct *textured* label which does not occur in either the haptic or visual results. Consequently, the multimodal model has higher precision and recall than the haptic model and higher recall than the visual model. For some objects, such as *furry eraser*, the multimodal classifier performs worse than the haptic classifier. However, the multimodal model generally performs better than either the haptic only or visual only model leading to higher overall AUC.

B. Haptic and Visual Data are Complementary

Although it seems most natural to classify haptic adjectives purely by touch, we demonstrate that visual and haptic signals are complementary. Fig. 6 compares the accuracy of our visual model to our haptic model. Adjectives which describe object size such as *thin* and *thick* are better classified by visual data whereas adjectives such as *squishy* and *absorbent* are better classified by the haptic signal. Surprisingly, even though the BioTac is equipped with a temperature sensor, our visual model is better at determining whether items are *cool*. Though the adjective *metallic* might seem like something that would be understood using visual cues, we find that our haptic model is better at classifying objects as *metallic* or not. This could possibly be because material properties such as temperature and roughness are important in classifying *metallic* objects, but not all metal surfaces have comparable metallic lusters. For example, objects such as the toothpaste box in Fig 5b appear shiny but are not actually *metallic*.

C. Which Haptic Signals Matter?

Although designing a deep architecture that reliably classifies haptic signals requires little haptic domain knowledge,

TABLE III: Comparison of Haptic Classification Methods. The letter in parentheses refers to the configuration in Table II.

	Shelf liner	Placemat	Furry eraser
Ground Truth	bumpy compressible squishy textured	smooth	fuzzy squishy textured
Haptic Outputs (I)	absorbent compressible soft springy squishy	compressible textured	fuzzy hairy soft squishy textured
Visual Outputs (L)	squishy	smooth	(empty)
Multimodal Outputs (M)	compressible scratchy squishy textured	compressible smooth thick	compressible rough squishy textured thick

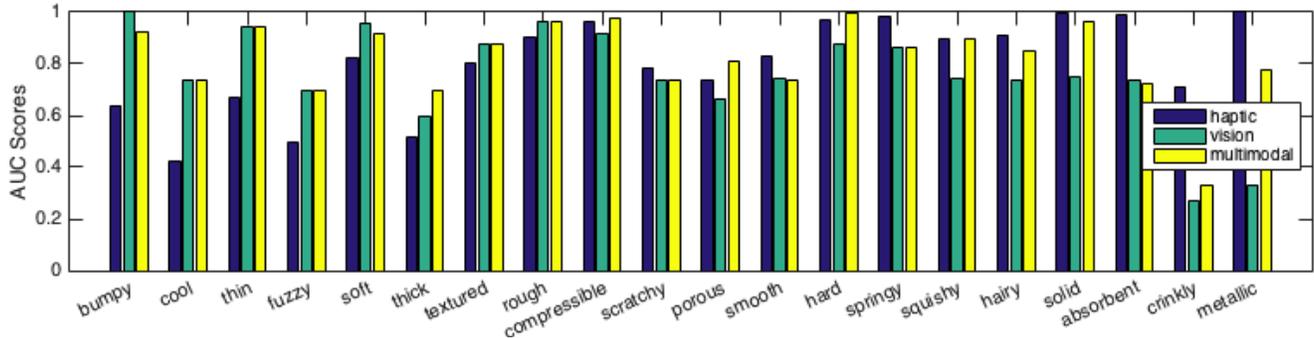


Fig. 6: Comparison of AUC scores for our best visual and best haptic classifier for each adjective. AUC scores are averaged over three train/test splits.

observing activations within the net can provide insight into what kinds of signals are important for classification. In our design of the haptic CNN, we use a “grouping” strategy, such that each signal, such as p_{AC} or p_{DC} , is processed independently before the fully connected layer. By looking at the activations of the final convolutional layer (conv3), we can observe which channels result in higher activations, and could thus be more important in haptic classification.

In Fig. 7, we plot conv3 activations for 2 adjective classifiers (*metallic* and *compressible*). The activations are obtained on the held-out positive testing instances. We show two *metallic* test instances which are classified correctly, and one *metallic* test instance which is classified incorrectly. Additionally, we show one test instance for *compressible* which is classified correctly. We observe that most of activations are zero, as many authors have observed [2] when analyzing networks trained on images.

For the adjective *metallic*, it appears activations for *Core Temperature Change* (t_{AC}) signal are important for classification, which makes intuitive sense. Because *metallic* objects are presumably made of thermally conductive material, temperature is likely a valuable signal for haptic classification. Furthermore, the activations of different trials look similar, as shown in Fig. 7 (a) and (b). This indicates the learnt haptic CNN is robust to signal variations across trials. Furthermore when the activations corresponding to t_{AC} are not strongly activated, as in Fig. 7 (c), the signal is incorrectly labeled as not *metallic*.

For the adjective *compressible* we note that electrode activations appear to be important. This finding implies that different signal channels are important for classification of different adjectives, suggesting that most of the signals recorded by the BioTac sensor are important for haptic classification. In future work, training models with data collected by less advanced sensors could verify the importance of the high dimensional BioTac sensor for fine-grained tactile understanding.

VII. CONCLUSION

We successfully design deep architectures for haptic classification using both haptic and visual signals, reinforcing that deep architectures can be used as a paradigm to learn features for a variety of signals. Furthermore, we demonstrate that haptic and visual signals are complementary, and combining data from both modalities improves performance. We believe that integrating our models into robotic pipelines can provide valuable haptic information which will help boost performance for myriad robot tasks.

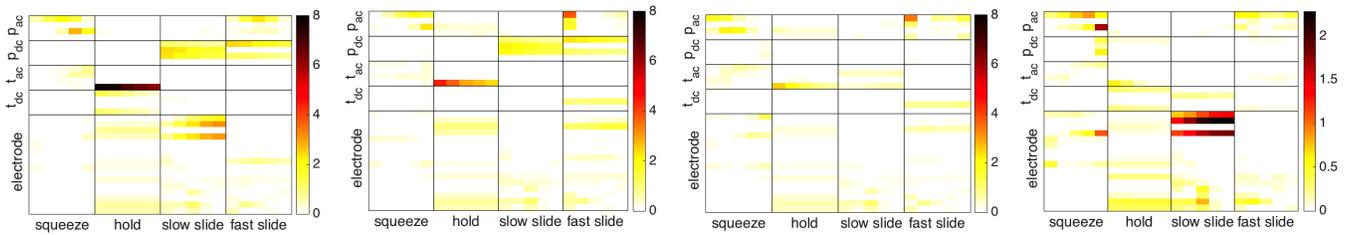
In the future, we believe that our model can be improved by a larger, more diverse dataset. Recently, [6] discussed efforts to collect a substantially larger dataset of paired visual and haptic data. A larger dataset could possibly improve our model in a variety of ways: it would allow to train a larger network, fine-tune through the entire multimodal network, and less aggressively down sample our haptic signal.

ACKNOWLEDGMENT

We would like to thank Jeff Donahue for advice and guidance during the initial stages of the experiments, as well as for useful discussions on deep models.

REFERENCES

- [1] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4277–4280.
- [2] P. Agrawal, R. Girshick, and J. Malik, “Analyzing the performance of multilayer neural networks for object recognition,” in *Computer Vision—ECCV 2014*. Springer, 2014, pp. 329–344.
- [3] A. Amedi, K. von Kriegstein, N. M. van Atteveldt, M. Beauchamp, and M. J. Naumer, “Functional imaging of human crossmodal identification and object recognition,” *Experimental Brain Research*, vol. 166, no. 3-4, pp. 559–571, 2005.
- [4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.
- [5] S. Bell, P. Upchurch, N. Snaveley, and K. Bala, “Material recognition in the wild with the materials in context database,” *arXiv preprint arXiv:1412.0623*, 2014.



(a) Adjective: metallic; Object: aluminum channel, Correct classification
 (b) Same as (a) but activations from another trial.
 (c) Adjective: metallic; Object: aluminum channel, Wrong classification
 (d) Adjective: compressible, Object: dish cloth, Correct classification

Fig. 7: Activations of conv3 in our haptic classification pipeline. We show the activations for two objects with the *metallic* and *compressible* property respectively. The haptic classifier correctly predicts the *metallic* adjective in (a) and (b), but not (c). The classifier correctly predicts *compressible* in (d). Darker colors represent higher activations.

- [6] A. Burka, S. Hu, S. Krishnan, K. J. Kuchenbecker, L. A. Hendricks, Y. Gao, and T. Darrell, "Toward a large-scale visuo-haptic dataset for robotic learning," 2015, proceedings of the *Workshop on the Future of Datasets in Vision at Computer Vision and Pattern Recognition (CVPR)*.
- [7] 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, no. 3, pp. 279–292, January 2015.
- [8] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, N. Fitter, J. Nappo, T. Darrell, and K. J. Kuchenbecker, "Using robotic exploratory procedures to learn the meaning of haptic adjectives," in *Proc. IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, May 2013, pp. 3048–3055.
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.
- [10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.
- [11] Z. Du, W. Yu, Y. Yu, L. Chen, Y. Ni, J. Luo, and X. Yu, "Fabric handle clusters based on fuzzy clustering algorithm," in *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on*, vol. 3. IEEE, 2008, pp. 546–550.
- [12] J. A. Fishel and G. E. Loeb, "Bayesian exploration for intelligent identification of textures," *Frontiers in neurorobotics*, vol. 6, 2012.
- [13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 580–587.
- [14] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010, pp. 249–256.
- [15] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1764–1772.
- [16] A. Graves, M. Liwicki, H. Bunke, J. Schmidhuber, and S. Fernández, "Unconstrained on-line handwriting recognition with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2008, pp. 577–584.
- [17] S. Griffith, J. Sinapov, V. Sukhoy, and A. Stoytchev, "A behavior-grounded approach to forming object categories: Separating containers from noncontainers," *Autonomous Mental Development, IEEE Transactions on*, vol. 4, no. 1, pp. 54–69, 2012.
- [18] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Computer Vision—ECCV 2014*. Springer, 2014, pp. 345–360.
- [19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.
- [20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [21] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 345–359, 2009.
- [22] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, A. Courville, P. Vincent, et al., "Emonets: Multimodal deep learning approaches for emotion recognition in video," *arXiv preprint arXiv:1503.01800*, 2015.
- [23] 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.
- [24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [25] M. Lodén, H. Olsson, L. Skare, and T. Axéll, "Instrumental and sensory evaluation of the frictional response of the skin following a single application of five moisturizing creams," *J Soc Cosmet Chem*, vol. 43, pp. 13–20, 1992.
- [26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv preprint arXiv:1411.4038*, 2014.
- [27] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," 1976.
- [28] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 1045–1048.
- [29] Y. Mukaibo, H. Shirado, M. Konyo, and T. Maeno, "Development of a texture sensor emulating the tissue structure and perceptual mechanism of human fingers," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*. IEEE, 2005, pp. 2565–2570.
- [30] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.
- [31] N. Pan, "Quantification and evaluation of human tactile sense towards fabrics," *International Journal of Design and Nature*, vol. 1, no. 1, pp. 48–60, 2007.
- [32] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.
- [33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.
- [34] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," *arXiv preprint arXiv:1412.7449*, 2014.
- [35] N. Wetts, V. J. Santos, R. S. Johansson, and G. E. Loeb, "Biomimetic tactile sensor array," *Advanced Robotics*, vol. 22, no. 8, pp. 829–849, 2008.
- [36] A. T. Woods and F. N. Newell, "Visual, haptic and cross-modal recognition of objects and scenes," *Journal of physiology-Paris*, vol. 98, no. 1, pp. 147–159, 2004.
- [37] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, 2014, pp. 487–495.