Deep Visuo-Tactile Learning: Estimation of Tactile Properties from Images (Extended Abstract)*

Kuniyuki Takahashi and **Jethro Tan**
Preferred Networks, Inc.
{takahashi, jettan}@preferred.jp

Abstract

Estimation of tactile properties from vision, such as slipperiness or roughness, is important to effectively interact with the environment. These tactile properties help humans as well as robots decide which actions they should choose and how to perform them. We, therefore, propose a model to estimate the degree of tactile properties from visual perception alone (e.g., the level of slipperiness or roughness). Our method extends an encoder-decoder network, in which the latent variables are visual and tactile features. In contrast to previous works, our method does not require manual labeling, but only RGB images and the corresponding tactile sensor data. All our data is collected with a webcam and tactile sensor mounted on the end-effector of a robot, which strokes the material surfaces. We show that our model generalizes to materials not included in the training data. [1] [2]

1 Introduction

Humans are able to perceive tactile properties, such as slipperiness and roughness, through haptics [Bergmann-Tiest, 2010]. After adequate visual-tactile experience, they are also capable of associating such properties from only visual perception [Tanaka and Horiuchi, 2015; Yanagisawa and Takatsuji, 2015]. More specifically, humans can roughly judge the *degree* of a certain tactile property (e.g., the *level* of slipperiness or roughness) [Fleming, 2014]. As an example, Figure 1 shows several materials with different degrees of softness and roughness judged by ourselves, although this may be subjective to our own judgment. Information on tactile properties can help us decide how we interact with our environment in advance, e.g., driving slower if we see that we have bad traction or grasp tighter if an item looks slippery. Like with humans, this ability to gauge the level of tactile properties can enable robots to deal with various objects and environments more effectively in both industrial settings and our daily lives.

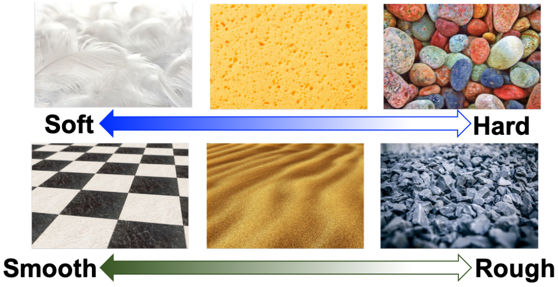


Figure 1: Example of material surfaces and their perceived material properties through visual information.

In the field of robotics and machine learning, a straightforward way to correlate vision with tactile properties is to design discrete classes per material type and to classify the images according to them. However, the performance of discrete classification methods highly depends on how well the designer chooses the number and types of class labels. Because of the wide variety of materials, which all have different tactile properties, discrete classes can not offer a sufficient resolution to judge the properties of the material well. Hence, we use an unsupervised method to represent tactile properties without using manually specified labels. We propose a method that we call *deep visuo-tactile learning* which extends a traditional encoder-decoder network with latent variables, where visual and tactile properties are embedded in a latent space. We emphasize that this is a *continuous* space, rather than a discrete one. This method is capable of generalizing to unknown materials when estimating their tactile properties, based on known tactile properties. Additionally, we only require the tactile sensor during the data collection phase and obtain a trained network model that can be used even in simulations or offline estimation, which allows for research without purchasing or damaging tactile sensors during runtime.

2 Related Work

2.1 Types of Tactile Sensors

Many researchers have developed tactile sensors [Dahiya *et al.*, 2013], some of which have been integrated to a robotic hand to enhance manipulation. The majority of these sensors falls in either of the following three categories:

1. Multi-touch sensors with sensing capabilities limited to

*Original full paper is presented in ICRA2019 [Takahashi and Tan, 2019].

[1] Dataset: https://github.com/pfnet-research/Deep_visuo-tactile_learning_ICRA2019

[2] Video: https://youtu.be/ys0QtKVVlOQ

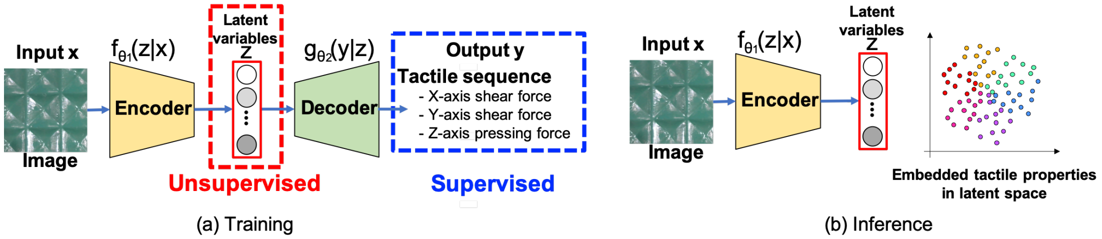


Figure 2: Proposed network architecture for deep visuo-tactile learning composed of encoder-decoder layers and latent variables. The input is a texture image of the material and, the output is the tactile data that contains measured forces by a tactile sensor in the x, y, and z axes. After training, latent variables would contain tactile properties of materials correlating images with tactile sense. Then, the network can infer tactile properties from only image input.

one axis per cell [Fishel and Loeb, 2012]

2. Single-touch sensors which can sense along three axes [Paulino *et al.*, 2017]

3. Multi-touch sensors which can sense along three axes. At the moment of writing, there are only three sensors: uSkin [Tomo *et al.*, 2016], Finger Vision [Yamaguchi and Atkeson, 2016], and GelSight [Dong *et al.*, 2017].

2.2 Recognition through Tactile Sensing

Research utilizing tactile sensors has grown recently as the availability and accessibility to tactile sensors has improved. Prior to the use of deep learning-based methods in these studies, data acquired from tactile sensors were often analyzed manually in order to define hand-crafted features [Yang *et al.*, 2016], or were only used as a trigger for certain actions [Yamaguchi and Atkeson, 2016]. Such methods may not scale well as technology for tactile sensing advances to provide e.g., higher resolution and larger amount of data, or whenever the task complexity grows. By utilizing learning methods, especially deep learning, tasks involving high-dimensional data such as image recognition [He *et al.*, 2016] and natural language processing [Conneau *et al.*, 2016] which were too difficult to process before can now be processed. Deep learning methods also found their way to applications where tactile sensing is involved [Schmitz *et al.*, 2014; Baishya and Bäuml, 2016; Yuan *et al.*, 2017a; Gao *et al.*, 2016]. Many of these studies, however, deal with the classification problem in order to e.g., recognize objects inside a robotic hand [Schmitz *et al.*, 2014], recognize materials [Baishya and Bäuml, 2016; Yuan *et al.*, 2017a] and properties [Gao *et al.*, 2016] from touch and image. Yuan *et al.*[Yuan *et al.*, 2017b] estimated object hardness as a continuous value using tactile sensor through supervised learning. We argue that these methods would be difficult to scale to different tactile properties due to the need to design each tactile property manually.

We note that our method differs from some other similar studies such as [Bell *et al.*, 2015] or [Schwartz and Nishino, 2019] in that we also make use of tactile information.

3 Deep Visuo-tactile Learning

We propose a method for deep visuo-tactile learning to estimate tactile properties from images by associating tactile information with images. Figure 2 shows our design of such a network. We aimed to design a network with a structure that is as simple as possible, but still sufficient for our purposes. We expect that increased complexity of the network architecture by e.g., using variational auto-encoder (VAE) and recurrent neural networks will mainly influence the accuracy and how tactile properties are represented as features, but that the results remain analogous. Complex models usually have the ability to learn more complex representations and larger datasets, but the effectiveness of our contribution can be shown using simpler models, hence our decision.

Our proposed network consists of 2D convolution layers for encoding, 3D deconvolution layers for decoding, and a multi layer perceptron (MLP) as hidden layers between the encoder f_θ and decoder g_θ. Our network outputs a time series sequence of tactile data consisting of applied forces and shear forces, while the input is an edge extracted image from the RGB image to prevent correlation to colors. The latent variables z are calculated from encoder $f_{\theta_1}(x)$ with training data $D = \{(x_1, y_1)..., (x_n, y_n)\}$, and the cost function L is calculated to minimize between expected output in training data y and inferred output y' from decoder $g_{\theta_2}(z)$ as follow:

$$\min_{\theta_1,\theta_2} \frac{1}{m} \sum_{i=1}^{m} L(y_i, g_{\theta_2}(f_{\theta_1}(x_i))), \qquad (1)$$

where θ is the parameters to be trained, $m(\leqq n)$ is the number of sequences for mini-batch training.

After training, z will hold visuo-tactile features that can be used to correlate the input images to the time series tactile data. We then map the embedded input to the latent space spanned by these variables; the coordinates of the embeddings in this space will represent the material's degree of the tactile property represented by the latent variable. Then, the network can infer tactile properties only from image input. However, we remind the reader that we do not focus on inferring the tactile time series data as output from the input images. Rather, we attempt to estimate the level of tactile properties, which can now be done by extracting the latent variables from the trained network. The reason for not directly using the values from the inferred time series data is that they are too sensitive to contact differences in e.g., the posture used to initiate the contact, the movement speed during contact, and the wear condition of the contact surface.

4 Experiment Setup

4.1 Hardware Setup

Tactile Sensor

The uSkin tactile sensor we use [Tomo *et al.*, 2016] consists of 16 taxels in a 4×4 square formation and is capable of measuring applied pressing forces and shear force in the x, y, and z axes (Figure 3(a) shows the coordinate system of the tactile sensor). To prevent damage to the tactile sensor, we have covered all surfaces of the sensor with lycra fabric. For our experiments, we only use the raw values of the pressure readings $x, y, z \in [0, 65535]$ on each of the taxels, which are configured to sample at 100 Hz.

Materials

For the materials, we have prepared 50x150 mm samples of 25 materials with different textures and rigidity that can be obtained off the shelf from a hardware store, see Figure 4. 15 of these materials are used for training, while the remaining 10 were used to evaluate our trained network as unknown materials. To normalize the experiments between each material and simplify the process of our data collection, we have glued each of the samples to their own PVC plate (See Figure 3(b)).

Sawyer

To conduct our experiments, we make use of a Sawyer 7-DOF robotic arm with a custom 3D-printed end-effector on which the uSkin tactile sensor and a Logitech C310 HD camera are mounted (See Figure 3 (a)).

4.2 Data Collection

For data collection, the following process is repeated ten times per material by the robot.

1. Move to a fixed initial position
2. Detect material surface: move down from a fixed initial height until force threshold Fz $5.0\dot{\text{N}}$ has been reached
3. Capture image: move up 1.6×10^{-2} m from detected material surface and take a picture
4. Move back to material surface and start capturing data from tactile sensor
5. Stroke material: move 3.0×10^{-2} m with constant velocity 2.0×10^{-3} m/sec in positive y-axis direction while tactile sensor makes contact with material surface

After data collection, we process all data to obtain our training data by doing the following. We normalize values to be between -1 and 1 and sample down each sequence of 900 time steps to 90 steps. Moreover, we perform rotations and croppings (from 640×480 pixels to pieces of 200×200 pixels) covering various areas to the obtained images. By doing this, we augment our data by 64 times per material and obtain a total of 960 samples of image-tactile pairs. Furthermore, we extract the edges from the RGB images of the materials with normalized pixel values between -1 and 1, because we reason that touch sense does not depend on material colors, and performing this preprocessing enables us to train our network with less data. For training, we use eight out of the ten collected images and tactile sequences. The remaining two image-tactile sequence pairs were split for validation and testing, respectively.

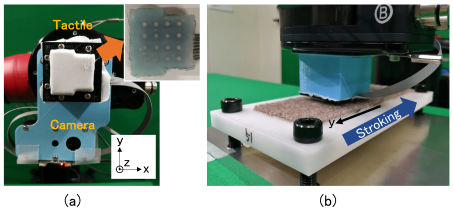


Figure 3: (a) end-effector with a tactile sensor and a web camera, (b) the Sawyer stroking a material to the minus y-axis direction.

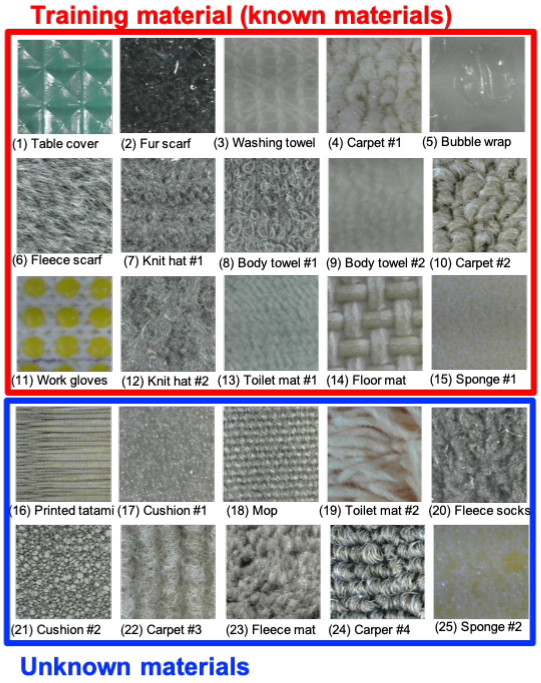


Figure 4: Trained materials (red) and unknown materials (blue) with their corresponding names included our dataset.

4.3 Network Hyper-parameters & Training

The architecture of our network model with four 2D and four 3D convolutional layers, and two full-connected MLPs to perform deep visuo-tactile learning is shown in Figure 2 as described in Section 3. More details on the network parameters are shown in Section 4.3. For all layers except last layer in the network, we make use of batch normalization. For training, we use mean squared error as the cost function, and a batch size of 15. All our network experiments were conducted on a machine equipped with 128 GB RAM, an Intel Xeon E5-2623v3 CPU, and a GeForce GTX Titan X with 12GB resulting in about 1.5 hours of training time.

	Layer	In	Out	Filter size	Stride	Padding	Activation function
Encoder	1st	1	32	(8,8)	(2,2)	(0,0)	ReLu
	2nd	32	32	(8,8)	(2,2)	(0,0)	ReLu
	3rd	32	32	(4,4)	(2,2)	(0,0)	ReLu
	4th	32	32	(4,4)	(2,2)	(0,0)	Tanh
Decoder	1st	1	32	(1,1,3)	(1,1,1)	(0,0,0)	ReLu
	2nd	32	32	(1,1,3)	(1,1,2)	(0,0,0)	ReLu
	3rd	32	32	(2,2,4)	(1,1,2)	(0,0,3)	ReLu
	4th	32	3	(2,2,4)	(1,1,2)	(1,1,2)	Tanh

Table 1: Network Design[3]

[3] For the hidden layer between encoder and decoder, we use two MPLs with 4 and 160 neurons with ReLu as activation function, respectively.

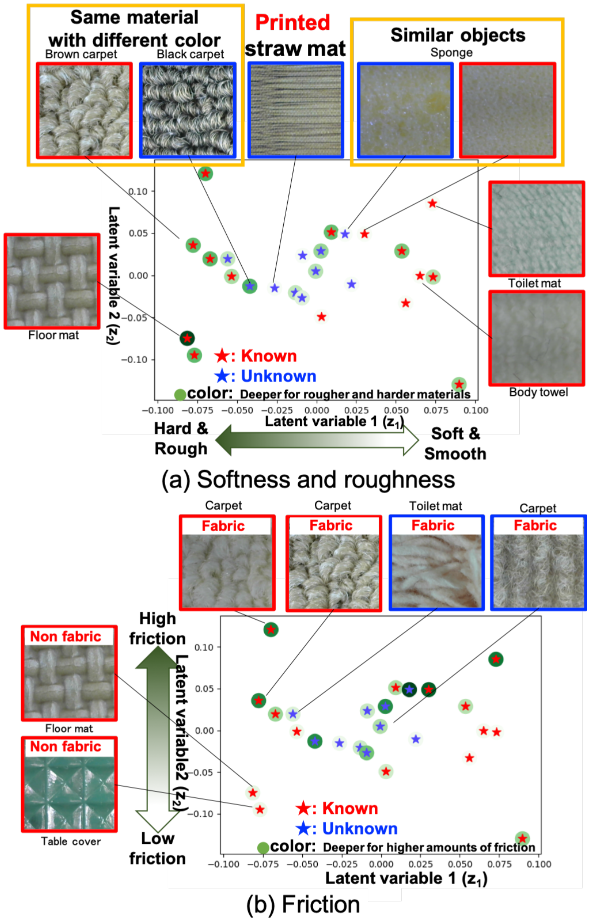
5 Results of Estimation of Tactile Properties

We present the results of estimated tactile properties in the latent space. After training with the 15 known materials shown in Figure 4, we let our network infer tactile properties with both known and 10 unknown materials. The tactile properties for all these materials are represented in four latent variables as z. Figure 5 shows the latent space of two of those latent variables. We have, to the best of our ability, analyzed the remaining two latent variables, but infer that the information they seem to represent are too diverse to analyze. Known materials used during training are represented by red-colored stars as seen in Figure 4, while unknown materials are represented by blue-colored stars.

To qualitatively evaluate the results of how tactile properties are represented in the latent space, we calculate the values for roughness, hardness, and friction for each material from tactile sequence data with forces in the x, y and z axes for all the 16 sensor taxels. We expect that the y-axis values contain information on friction between the end-effector and the material due to the applied shear forces while stroking. We also expect that the z-axis embeds information on roughness as well as softness of a material surface. The color of the circles in Figure 5 (a) is deeper for more rough and harder materials, and deeper colors in Figure 5 (b) represent higher friction of materials. This enables us to see whether the mapping of these tactile properties for each material in the latent space corresponds to the degree of roughness, hardness, and friction from our calculated values. We note that tactile properties are represented in the latent space according to what the tactile sensor perceived. Therefore, what we personally perceive as the degree of tactile properties might not correspond to our result.

Figure 5 (a) indicates that materials with relatively high degree of hardness and roughness tend to get mapped to regions with lower values of the latent variable z_1. For example, floor mats and brown carpets were recognized as hard and rough, while materials like body towels and toilet mats were recognized as soft and smooth. Moreover, we see that an unknown black carpet is relatively closer to the somewhat similarly textured, known brown carpet than to the other unknown materials in the center region, despite their difference in color. In the same manner, the values of latent variables of known and unknown sponges are close to each other. From this point of view, Figure 5 (a) suggests that the degree of softness and roughness of materials are embedded in latent variable z_1. An interesting case is a Japanese straw mat surface printed on paper and was estimated to have a high degree of roughness. However, tactile values corresponding to this degree of roughness could not be obtained by the sensor. This shows the limitation of our current model on how accurate tactile properties can be estimated from only two-dimensional images as input.

Furthermore, Figure 5 (b) indicates that materials with seemingly low friction tend to get mapped to regions with low values of z_2. For example, fabriclike materials have relatively high friction when stroked by the sensor due to contact with the lycra cover of the tactile sensor. On the other hand, non fabric materials like plastic slip more easily when stroked and have relatively low friction as a result. We can see that relatively glossy (thus seemingly slippery) materials (table cover and floor mat) are mapped to areas with the lowest z_2 values. Therefore, we believe that z_2 is connected to the amount of friction surfaces provide during stroking.



(a) Softness and roughness

(b) Friction

Figure 5: Visualization of tactile properties of (a) softness and roughness, and (b) friction from latent spaces of the hidden layer.

6 Conclusion

We proposed a method to estimate tactile properties from images, called deep visuo-tactile learning, for which we built an encoder-decoder network with latent variables. The network is trained with material texture images as input and time series sequences tactile acquired from a tactile sensor as output. After training, we obtained a continuous latent space representing tactile properties and their degrees for various materials. Our experiments showed that unlike conventional methods relying on classification, our network is able to deal with unknown material surfaces and adapted the latent variables accordingly without the need for manually designed class labels.

Acknowledgements

The authors would like to thank Kenta Yonekura for his assistance for our experiments, and Wilson Ko for proofreading this article.

References

[Baishya and Bäuml, 2016] Shiv S Baishya and Berthold Bäuml. Robust Material Classification With a Tactile Skin Using Deep Learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8–15, 2016.

[Bell *et al.*, 2015] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015.

[Bergmann-Tiest, 2010] Wouter M. Bergmann-Tiest. Tactual Perception of Material Properties. *Vision Research*, 50(24):2775–2782, 2010.

[Conneau *et al.*, 2016] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. Very Deep Convolutional Networks for Natural Language Processing. *arXiv preprint arXiv:1606.01781*, 2016.

[Dahiya *et al.*, 2013] Ravinder S Dahiya, Philipp Mittendorfer, Maurizio Valle, Gordon Cheng, and Vladimir J Lumelsky. Directions Toward Effective Utilization of Tactile Skin: A Review. *IEEE Sensors Journal*, 13(11):4121–4138, 2013.

[Dong *et al.*, 2017] Siyuan Dong, Wenzhen Yuan, and Edward Adelson. Improved GelSight Tactile Sensor for Measuring Geometry and Slip. *arXiv preprint arXiv:1708.00922*, 2017.

[Fishel and Loeb, 2012] Jeremy A Fishel and Gerald E Loeb. Sensing Tactile Microvibrations with the BioTac—Comparison with Human Sensitivity. In *IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 1122–1127, 2012.

[Fleming, 2014] Roland W. Fleming. Visual Perception of Materials and Their Properties. *Vision Research*, 94:62–75, 2014.

[Gao *et al.*, 2016] Yang Gao, Lisa Anne Hendricks, Katherine J Kuchenbecker, and Trevor Darrell. Deep learning for tactile understanding from visual and haptic data. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 536–543. IEEE, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Paulino *et al.*, 2017] Tiago Paulino, Pedro Ribeiro, Miguel Neto, Susana Cardoso, Alexander Schmitz, Jose Santos-Victor, Alexandre Bernardino, and Lorenzo Jamone. Low-cost 3-axis Soft Tactile Sensors for the Human-Friendly Robot Vizzy. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 966–971, 2017.

[Schmitz *et al.*, 2014] Alexander Schmitz, Yusuke Bansho, Kuniaki Noda, Hiroyasu Iwata, Tetsuya Ogata, and Shigeki Sugano. Tactile Object Recognition Using Deep Learning and Dropout. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 1044–1050, 2014.

[Schwartz and Nishino, 2019] Gabriel Schwartz and Ko Nishino. Recognizing material properties from images. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[Takahashi and Tan, 2019] Kuniyuki Takahashi and Jethro Tan. Deep visuo-tactile learning: Estimation of tactile properties from images. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8951–8957. IEEE, 2019.

[Tanaka and Horiuchi, 2015] Midori Tanaka and Takahiko Horiuchi. Investigating Perceptual Qualities of Static Surface Appearance Using Real Materials and Displayed Images. *Vision Research*, 115:246–258, 2015.

[Tomo *et al.*, 2016] Tito Pradhono Tomo, Wai Keat Wong, Alexander Schmitz, Harris Kristanto, Alexandre Sarazin, Lorenzo Jamone, Sophon Somlor, and Shigeki Sugano. A Modular, Distributed, Soft, 3-axis Sensor System for Robot Hands. In *IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 454–460, 2016.

[Yamaguchi and Atkeson, 2016] Akihiko Yamaguchi and Christopher G Atkeson. Combining Finger Vision and Optical Tactile Sensing: Reducing and Handling Errors While Cutting Vegetables. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 1045–1051, 2016.

[Yanagisawa and Takatsuji, 2015] H. Yanagisawa and K. Takatsuji. Effects of Visual Expectation on Perceived Tactile Perception: An Evaluation Method of Surface Texture with Expectation Effect. *International Journal of Design*, 9(1), 2015.

[Yang *et al.*, 2016] Haolin Yang, Fuchun Sun, Wenbing Huang, Lele Cao, and Bin Fang. Tactile Sequence Based Object Categorization: A Bag of Features Modeled by Linear Dynamic System with Symmetric Transition Matrix. In *International Joint Conference on Neural Networks (IJCNN)*, pages 5218–5225, 2016.

[Yuan *et al.*, 2017a] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting Look and Feel: Associating the Visual and Tactile Properties of Physical Materials. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR17)*, pages 21–26, 2017.

[Yuan *et al.*, 2017b] Wenzhen Yuan, Chenzhuo Zhu, Andrew Owens, Mandayam A Srinivasan, and Edward H Adelson. Shape-independent Hardness Estimation Using Deep Learning and a GelSight Tactile Sensor. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 951–958, 2017.