TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/10 pp645–653
DOI: 10.26599/TST.2018.9010095
Volume 24, Number 6, December 2019

Fabric Recognition Using Zero-Shot Learning

Feng Wang, Huaping Liu*, Fuchun Sun, and Haihong Pan

Abstract: In this work, we use a deep learning method to tackle the Zero-Shot Learning (ZSL) problem in tactile material recognition by incorporating the advanced semantic information into a training model. Our main technical contribution is our proposal of an end-to-end deep learning framework for solving the tactile ZSL problem. In this framework, we use a Convolutional Neural Network (CNN) to extract the spatial features and Long Short-Term Memory (LSTM) to extract the temporal features in dynamic tactile sequences, and develop a loss function suitable for the ZSL setting. We present the results of experimental evaluations on publicly available datasets, which show the effectiveness of the proposed method.

Key words: Zero-Shot-Learning (ZSL); fabric recognition; tactile recognition; deep learning

supervised + unsupervised

1 Introduction

Surface material properties such as texture, frictional coefficients, roughness, and compliance are among the most important aspects affecting interactions with the environment. Surface material categorization, therefore, plays a vital role in many fields, including environmental exploration[1], humanitarian demining[2], robotic manipulation and grasp[3], milling[4], and biomedical applications[5]. This topic has attracted increasing interest in recent years[6].

A popular sensor for identifying the surface materials is the camera[7]. The authors of Ref. [8] used acquired images and estimated 3D points to identify material categories. Recently, the authors of Refs. [9, 10] developed a deep learning methodology for material recognition and image segmentation in the wild. In Ref. [11], the authors tackled the problem of visually predicting surface friction in environments with diverse surfaces, and then integrated this knowledge into biped robot locomotion planning. Although image plays an important role in surface material recognition, it has some intrinsic limitations due to the diversity in the appearance of surface materials. Because instances from a single material category can span a range of object categories, shapes, colors, textures, and lighting and imaging conditions, imaged-based material recognition systems are sensitive to a number of image transformations such as view point changes, occlusion, and lighting variation. There are also many cases for which image cue alone is not sufficient to distinguish a surface material. For example, a pure cotton shirt might have the same appearance as one made of chemical fiber, making them difficult to distinguish visually. However, if they are touched, the tactile feedback makes it possible to tell the difference. This has motivated researchers and engineers to adopt tactile sensors to obtain complementary cues[12].

There has been much research dealing with tactile material recognition[13]. The authors in Ref. [14] developed a multi-functional tactile sensor for detecting material hardness, and those in Ref. [15] proposed a haptic exploration strategy for recognizing unknown object surface materials using a specially designed finger for contact sensing. In Ref. [16], the authors investigated material recognition based on heat transfer, given varying initial conditions and short-duration contact. In Ref. [2], an intelligent prodder was

• Feng Wang, Huaping Liu, and Fuchun Sun are with Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: w-f17@tsinghua.edu.cn; hpliu@tsinghua.edu.cn.
• Haihong Pan is with College of Mechanical Engineering, Guangxi University, Nanning 530003, China.
* To whom correspondence should be addressed.
Manuscript received: 2018-02-09; revised: 2018-04-18; accepted: 2018-04-27

developed to stimulate a material surface and read the response. The measurement procedure identified a set of suitable parameters from the viscoelastic response and used these parameters for material recognition and classifications. In Ref. [17], the authors developed an acoustics-based terrain classification system for legged robots. In Ref. [18], the authors predicted failure in lubricated surfaces using acoustic signals. In Ref. [19], the authors developed a robot-assisted acoustic infrastructure inspection system. In Ref. [20], a softness measurement technique was developed, whereby a forceps-type tactile sensor responds to acoustic reflection. Applying deep learning, the authors of Ref. [21] proposed a robust material classification scheme using a tactile skin. Recently, the authors of Refs. [22, 23] developed shape-independent hardness estimation methods using GelSight tactile sensors.

Fabric materials typically exhibit a range of attributes, including hardness, density, and so on. The authors of Ref. [24] investigated the affective and perceptual dimensions and connection between touched materials. The authors of Ref. [25] recognized social touch gestures using tactile signals. The authors of Ref. [26] developed a collection of tactile classification datasets and classified objects using binary tactile adjectives. Their work relied on hand-crafted features for tactile classification. The authors of Ref. [27] proposed a deep learning method for classifying objects using tactile adjectives, based on both visual and physical-interaction data. The work in Ref. [28] addressed the multi-label tactile property analysis problem.

Despite significant progress in tactile object recognition[29], most learning schemes require a sufficient number of labeled samples for effective classifier design. Unlike the visual modality, data collection for tactile modality is more difficult due to the need for expensive tactile sensors and a complicated exploration procedure. In practice, we may encounter situations for which no training sample is available for some categories, which makes it very difficult to establish a classifier design for the unseen category.

Zero-Shot Learning (ZSL), which learns models from datasets with no labeled data for novel classes, is being increasingly recognized as a way to deal with these difficulties. In the ZSL framework, there is no labeled data for the target class, but recognition models are generally built with supervision. To address this critical problem, a semantic space that is shared among classes is utilized for transferring knowledge from seen classes that are sufficiently labeled[30]. These semantic embedding vectors might be obtained from human-labeled object attributes. Figure 1 shows the basic formulation of the tactile ZSL concept.

In this work, we tackle the ZSL problem for tactile material recognition by incorporating advanced semantic information into the training model. This is achieved by applying a deep learning method. The main technical contribution of this work is our proposal of an end-to-end deep learning framework to solve the tactile ZSL problem. In this framework, we use a Convolutional Neural Network (CNN) to extract the spatial features and Long Short-Term Memory (LSTM) to extract the temporal features for dynamic tactile sequences[31]. We also developed a loss function that is suitable for the ZSL setting. Finally, we performed experimental evaluations on publicly available datasets, and the results demonstrate the effectiveness of the proposed method. To the best of our knowledge, to date, only the authors of Ref. [32] have investigated tactile ZSL using a direct-attribute prediction method. Compared to that of Ref. [32], our approach is totally different.

The rest of this paper is organized as follows: In Section 2, we briefly review related ZSL work. In Section 3, we formulate our problem, and we describe our methodology in Section 4. In Section 5, we present our experimental results, and we draw our conclusion in Section 6.

2 Related Work

Recently, ZSL has attracted an increasing amount of

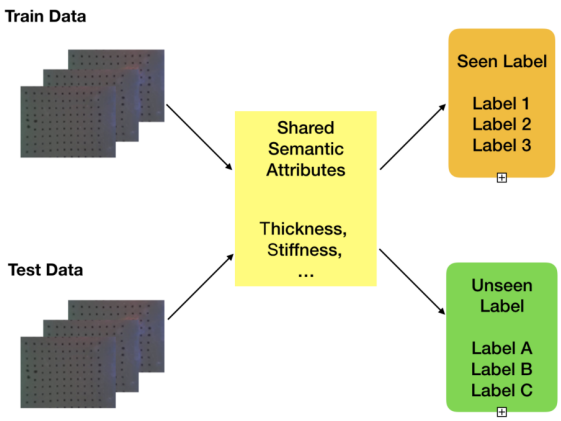


Fig. 1 Zero-shot tactile fabric recognition framework. The training and testing sets do not share a common label space, but do share a semantic attribute space.

attention. ZSL deals with the problem of learning to classify previously unseen class instances. It is particularly useful in large-scale classification, where labels are often missing for many instances or even entire categories. During ZSL training, source domain side information and target-domain data are provided that correspond only to a subset of classes, which we refer to as seen classes. During test time for the source domain, side information is provided for the unseen classes. A target-domain instance from an unknown unseen class is then presented. The goal during test time is to predict the class label for the unseen target-domain instance.

There has been much research effort devoted to the consideration of how to bridge the semantic gap between seen and unseen classes. The authors of Ref. [33] developed direct- and indirect-attribute prediction, which seeks an explicit mapping function based on feature descriptors of corresponding semantic representations. In Refs. [34, 35], the authors searched new spaces in which feature descriptors and semantic annotations in samples from the same class have maximum similarity. In Ref. [36], the authors exploited the inter-class relationship between seen and unseen classes in the semantic space. Some recent works have involved the development of various strategies to solve the ZSL problem[37–41]. However, all of these works focus on visual recognition. There has been very little work on ZSL for tactile recognition. Recently, the authors of Ref. [32] developed the first tactile ZSL system that enables a robot, using exploration alone, to recognize objects that it encounters for the first time. This system uses Direct-Attribute Prediction (DAP)[42] to train on the semantic representation of objects based on a list of tactile attributes. These attributes reflect physical properties such as shape, texture, and material, and constitute an intermediate layer of related objects, which is used for knowledge transfer.

3 Problem Formulation

ZSL learns knowledge from a labeled training set of seen classes about unseen classes that are semantically related to the seen classes. The goal is to predict a label for each instance in the unseen classes. For example, to learn a model from labeled samples of the classes *Cat*, *Bird*, and *Cow*, we must classify instance samples from the unseen classes *Dog*, *Pig*, and *Fish*.

Given the problem framework shown in Fig. 1, we know that ZSL deals with the problem of learning to classify previously unseen class instances, with the main challenge being the fact that the label set of the training samples is entirely different from that of the test samples. That is to say, we may have

$$\mathcal{Y} \bigcap \mathcal{Z} = \varnothing,$$

where $\mathcal{Y}$ and $\mathcal{Z}$ are the label sets of the training and testing samples, respectively. The task is to learn knowledge from object instances of $\mathcal{Y}$, and then use the obtained knowledge to recognize object instances in the $\mathcal{Z}$ set.

Tactile signals can exhibit different characteristics depending on the tactile sensors and exploration strategies adopted. The most representative tactile signals include low-resolution images, one-dimensional dynamic sequences, multi-dimensional dynamic sequences, and even tactile videos[22, 23]. In this work, we utilize the most complicated but more informative tactile video.

We denote the training sample set as

$$\{\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_N\} \subset \mathcal{V},$$

where $\mathcal{V}$ is the set of tactile sequences and N is the number of the training samples. For each tactile sample $\boldsymbol{V}_i$, we have one corresponding label vector $\boldsymbol{y}_i \in \mathbf{R}^{|\mathcal{Y}|}$. Vector $\boldsymbol{y}_i$ is an elementary vector, where the element 1 indicates the class category. That is to say, the element of the label vector $\boldsymbol{y}_i$ is defined as

$$\boldsymbol{y}_i(c) = \begin{cases} +1, & \text{if } \boldsymbol{V}_i \text{ belongs to the } c\text{-th class;} \\ 0, & \text{othewise} \end{cases} \tag{1}$$

for $c = 1, 2, \ldots, |\mathcal{Y}|$.

Figure 2 shows examples of videos recorded by a GelSight sensor, in which we can see that the tactile images from thin and thick fabrics are quite different. These differences are visible in the videos recorded by the GelSight sensor[22].

The goal of this work was to develop a classifier that is able to recognize an input tactile sample $t \in \mathcal{V}$, that does not belong to one of the $|\mathcal{Y}|$ classes in $\mathcal{Y}$, but rather to one of the $|\mathcal{Z}|$ classes in $\mathcal{Z}$. Without additional information, this is obviously an impossible goal. Fortunately, many properties perceived by tactile sensors can be described using adjectives such as *Stiff*, *Stretchable*, *Flexible*, and so on.

To illustrate the power of adjective labels, we analyzed the GelSight tactile dataset[22] which we introduce later. The developer of this dataset labeled fabrics with estimations of their physical parameters, including *Thickness*, *Stiffness*, *Stretchiness*, and

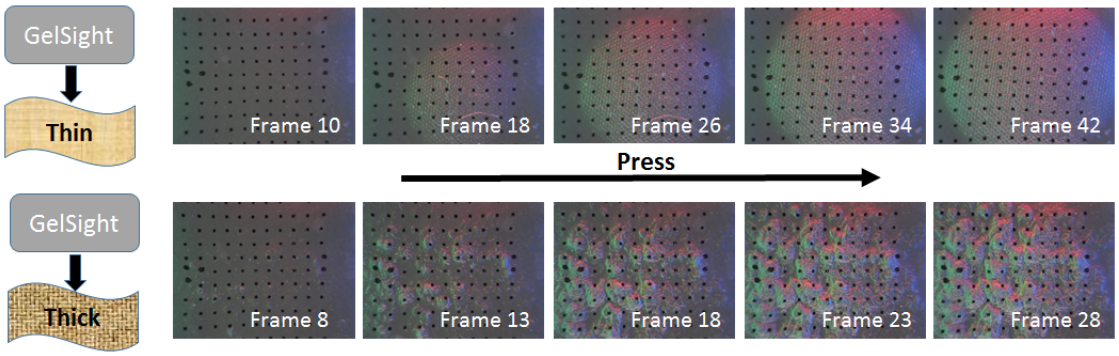


Fig. 2 Differences between GelSight tactile sequences for the thin (top) and thick (bottom) fabrics.

Density. Details of this protocol can be found in Ref. [22]. In addition, all of the samples from 100 instances are clustered into eight categories of human-estimated physical parameters using k-means clustering, as shown in Fig. 3. Intuitively, fabrics in the same category will be relatively similar in their physical properties. Figure 4 shows intuitive descriptions of each class, from which we find that each material object exhibits several adjective properties, and the adjective attribute is indeed a useful cue for differentiating object instances.

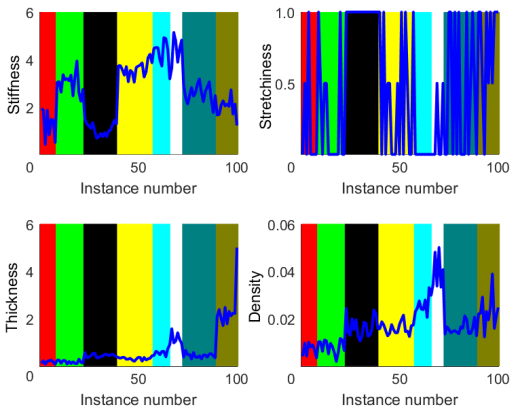


Fig. 3 Original attribute annotation information for 100 fabric instances in the GelSight datasets. Different color blocks represent different clusters.

Thus motivated, we can provide an attribute vector $\boldsymbol{l}_i \in \mathbf{R}^L$ for any tactile sample, where L is the number of relevant attributes. Given a set of L attributes, the element of the attribute vector $\boldsymbol{l}_i \in \mathbf{R}^L$ corresponding to the i-th sample is defined as follows:

$$\boldsymbol{l}_i(l) = \begin{cases} +1, & \text{if attribute } l \text{ is associated with sample } \boldsymbol{V}_i; \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

for $l = 1, 2, \ldots, L$. We note that it is permissible to have multiple elements in $\boldsymbol{l}_i$ that are non-zero, and we use $\mathcal{E}_L$ to denote the set of all feasible attribute vectors.

To construct a connection between the training set and test sample, we assign a semantic adjective vector to each of the test samples. This information is not difficult to obtain in practice since users can easily annotate the samples using tactile adjectives.

The task, then, is to develop an algorithm to identify the class label of a tactile sample $\boldsymbol{V} \in \mathcal{V}$, using the information provided by the training set $\{\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_N\}$ and the label set $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$, with the help of the adjective label set $\{\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_N\}$.

For each class in the test set, we can provide an

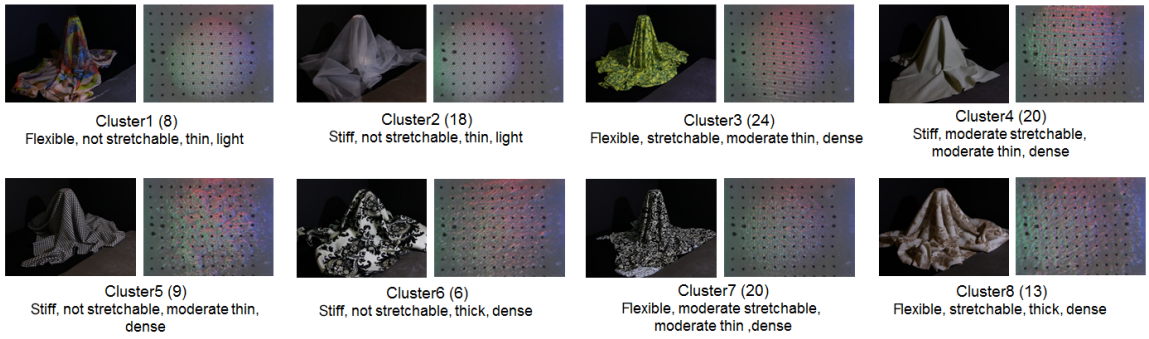


Fig. 4 Clustering of fabrics based on human labels. Numbers in brackets denote the fabric number in the cluster. Note that for each cluster, we show a color image and a tactile image, but we did not use the color images in this work. In addition, we refined the attribute annotation based on the work of Ref. [22].

adjective vector $\boldsymbol{a}_g \in \mathcal{E}_L$ for $g \in \mathcal{Z}$ to indicate its tactile adjective attributes. We set the l-th element of the attribute vector $\boldsymbol{a}_g \in \mathbf{R}^L$ corresponding to the g-th class to 1 if the adjective l is associated with the g-th unseen class, and 0 otherwise.

4 Methodology

4.1 Pre-processing

To deal with the tactile signals collected by GelSight sensors, which exhibit complicated and dynamic video characteristics, we developed a deep learning architecture, as shown in Fig. 5, to extract feature vectors, which can be mapped to attributes or labels.

There is a difficulty in the tactile signal processing in that the exploration activity strongly influences the tactile signal. For example, for one fabric sample, if two operators press upon it with significantly different force, the obtained tactile signals will differ. Some complicated methods have been developed to deal with this problem, such as dynamic time warping. In this work, we use the simple method proposed in Ref. [23] to constrain the video sequence so that it begins and ends at times that are consistent for all manipulation conditions.

Without loss of generality, we assume that video $\boldsymbol{V}_i$ contains T frames, denoted as $\boldsymbol{V}_i(1), \boldsymbol{V}_i(2), \ldots, \boldsymbol{V}_i(T)$. After analyzing the whole dataset, we find that the first frame $\boldsymbol{V}_i(1)$ is always obtained without pressing in all videos. Therefore, we take the mean intensity value $I(\boldsymbol{V}_i(1))$ of this frame as the basis. Then, we calculate the difference value for each frame as

$$\mathrm{Diff}_i(t) = |I(\boldsymbol{V}_i(t)) - I(\boldsymbol{V}_i(1))|,$$

for $t = 2, 3, \ldots, T$. Next, as the starting frame, we take the first frame for which the difference value is larger than a prescribed threshold, which is denoted as $\boldsymbol{V}_i^{(1)}$, and take the frame that exhibits the maximum difference value as the ending frame, which is denoted as $\boldsymbol{V}_i^{(5)}$. We choose the other three frames $\boldsymbol{V}_i^{(2)}$, $\boldsymbol{V}_i^{(3)}$, and $\boldsymbol{V}_i^{(4)}$ so that they are evenly distributed.

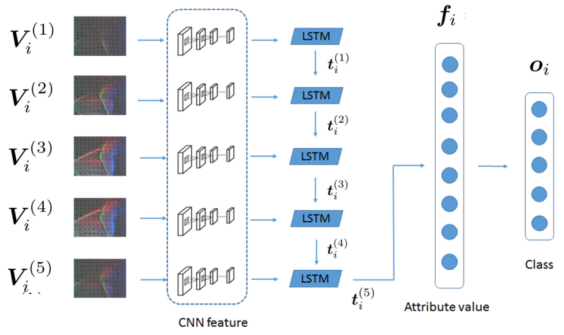


Fig. 5 **Architecture for tactile zero-shot learning.**

4.2 Deep learning

According to the above setting, we decompose each tactile video sample $\boldsymbol{V}_i \in \mathcal{V}$ into five consecutive images, which we denote as $\boldsymbol{V}_i^{(1)}$, $\boldsymbol{V}_i^{(2)}$, $\boldsymbol{V}_i^{(3)}$, $\boldsymbol{V}_i^{(4)}$, and $\boldsymbol{V}_i^{(5)}$, respectively.

To model temporal information, we represent each GelSight image $\boldsymbol{V}_i^{(k)}$ using convolutional network features $\mathrm{CNN}(\boldsymbol{V}_i^{(k)})$ and a recurrent neural network with LSTM units as

$$\boldsymbol{t}_i^{(k)} = \mathrm{LSTM}(\boldsymbol{t}_i^{(k-1)}, \mathrm{CNN}(\boldsymbol{V}_i^{(k)})).$$

for $k = 1, 2, \ldots, 5$, where the feature vector $\boldsymbol{t}_i^{(k)}$ represents the hidden state, which is updated by the LSTM network.

We obtain the estimated attribution vector $\boldsymbol{f}_i$ using $\boldsymbol{t}_i^{(5)}$ and a fully-connected layer,

$$\boldsymbol{f}_i = \sigma(W\boldsymbol{t}_i^{(5)} + b),$$

where $\sigma(\cdot)$ is the sigmoid function and W and b are the learnable parameters.

The final layer outputs a $|\mathcal{Y}|$-dimensional vector $\boldsymbol{o}_i$. The c-th element of $\boldsymbol{o}_i$ reflects the similarity between the attrubite vector of the c-th class and $\boldsymbol{f}_i$.

$$o_i(c) = -||\boldsymbol{f}_i - \bar{\boldsymbol{l}}_c||_2^2 \tag{3}$$

where $\bar{\boldsymbol{l}}_c$ is the shared attribute vector of all of the samples in the c-class.

Assume the label number of the i-th sample is c^*, i.e., we have $\boldsymbol{y}_i(c^*) = 1$. According to the k-nearest neighbor concept, we always hope that $o_i(c^*)$ is sufficiently large. Therefore, we can develop the softmax loss function, which is defined as

$$\mathcal{C} = \sum_{i=1}^{N} -\log \frac{\exp(o_i(c^*))}{\sum_{c=1}^{|\mathcal{Y}|} \exp(o_i(c))}.$$

The above loss concerns the classification performance when using the supervised information. However, we may encounter some situations in which some attributes are not distinguishable in the training samples. For example, all of the training samples may have the attribute value *Thick* = 0, whereas some of the testing samples may have the attribute value *Thick* = 1. Therefore, the above objective function cannot use this attribute for discriminating between training samples.

To address this problem, we added a regularization term that requires that the calculated attribute vector be

similar to the true attribute vector.

$$\mathcal{R} = \sum_{i=1}^{N} ||\boldsymbol{f}_i - \boldsymbol{l}_i||_2^2.$$

We then designed the final objective function as follows:

$$\mathcal{L} = \mathcal{C} + \lambda\mathcal{R} \tag{4}$$

where λ is a penalty parameter. Although both $\mathcal{C}$ and $\mathcal{R}$ terms are calculated using the Euclidean distance between estimated attribute vectors and true attribute vectors, they have different objectives. $\mathcal{C}$ enables more discriminative attribute estimation, whereas $\mathcal{R}$ concerns the absolute attribute estimation value, which brings the estimation close to the real value. By comparison, the work in Ref. [37] considered only the classification task.

4.3 Classification of unseen samples

For one unseen sample $\boldsymbol{V}$, we follow the same setting in Section 4.1 to get the five frames which are denoted as $\boldsymbol{V}^{(1)}$, $\boldsymbol{V}^{(2)}$, $\boldsymbol{V}^{(3)}$, $\boldsymbol{V}^{(4)}$, and $\boldsymbol{V}^{(5)}$, and follow the setting described in Section 4.2 to obtain the corresponding feature vectors $\boldsymbol{t}^{(1)}$, $\boldsymbol{t}^{(2)}$, $\boldsymbol{t}^{(3)}$, $\boldsymbol{t}^{(4)}$, and $\boldsymbol{t}^{(5)}$. The feature descriptor for this testing sample is given by

$$\boldsymbol{f} = \sigma(W^*\boldsymbol{t}^{(5)} + b^*),$$

where W^* and b^* are the learned parameters.

Since the goal is to determine the label for this sample, and the label is in the $\mathcal{Z}$ set, we compare the obtained feature descriptor $\boldsymbol{f}$ with all of the attribute vectors $\boldsymbol{a}_g$ for all of the classes $g \in \mathcal{Z}$. Then, we obtain the label:

$$g^* = \arg\min_{g\in\mathcal{Z}} ||\boldsymbol{f} - \boldsymbol{a}_g||_2.$$

5 Experimental Results

We performed an experimental validation on the recently released GelSight dataset[22], visual samples of which are shown in Fig. 4. According to the introduction in Ref. [22], during the data collection period for the press process, a human presses the GelSight tactile sensor onto the fabrics and obtains a sequence of GelSight tactile images, with a resolution of 960×720. Note that the original dataset also provides camera images of the sample fabrics. However, according to our experimental results, we found neither color nor depth images to help in the tactile ZSL process. We will investigate the fusion of these parameters in the future.

The GelSight dataset contains three forms of tactile data from 119 fabrics. The first form is *Flat*, when the GelSight sensor is pressed onto a single layer of flat fabric; the second is *Fold*, when the GelSight sensor is pressed onto a fold of the fabric; and the third is *Rand*, which indicates that the fabrics are randomly placed. For each fabric, we collected from 10–15 samples.

According to the manual annotations shown in Fig. 4, we constructed an attribute set comprising *Thin*, *Thick*, *Dense*, *Light*, *Stiff*, *Flexible*, *Not-stretchable*, and *Stretchable*. We note that all of the training samples had the attribute property *Thick* $= 0$, whereas the test samples had different values for this attribute.

To test the ZSL performance, we took samples belonging to classes 2, 3, 4, 5, and 7 as the training set, and the remainder as the test set. In this setting, the labels of test set are discontinuous with those of the training set.

In our work, we used the penultimate layer fc7 of the VGG network for CNN feature extraction and we set the dimension of the feature descriptor $\boldsymbol{f}_i$ to 200. We trained the model by fine-tuning a pre-trained VGG network using stochastic gradient descent with a momentum of 0.9 and a learning rate of 0.001. To implement the algorithm, we used Tensorflow software.

5.1 Results

Since there exists an attribute (thick), which is not distinguishable in the training samples, we could not use the classical DAP method[42] for comparison. Instead, we used the more recently developed Embarrasingly Simple Zero Shot Learning (ESZSL) method, which was proposed in Ref. [37]. This method uses a fixed attribute matrix together with training samples to learn the mapping from the feature space to the attribute space. In fact, ESZSL has been reported to always perform better than some baseline methods such as DAP in most scenarios[37].

Figure 6 shows the recognition accuracies of different methods on the three datasets. To analyze the influences of parameter λ, we select it from the set $\{0, 0.1, 1, 10, \ln f\}$, and Table 1 shows the results. Based on these results, we can make the following observations:

(1) In all of the situations, the proposed method performs better than the ESZSL by a significant margin.

(2) The performance of the proposed method is influenced by parameter λ, which indicates that both terms in the objective function (Eq. (4)) play a role. We note that when $\lambda = 0$, the objective function concerns only the classification tasks. In addition, we use $\lambda =$

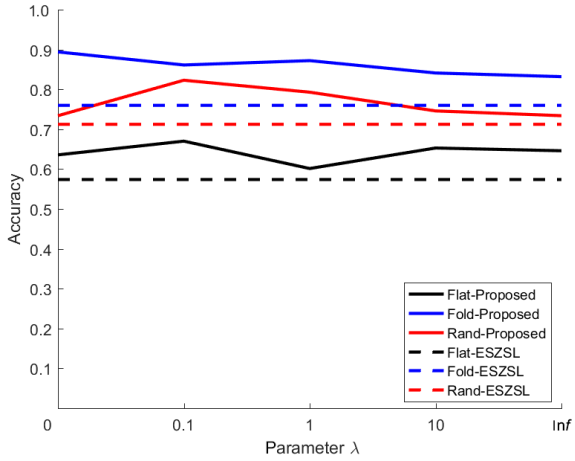


Fig. 6 Accuracy versus the parameter λ.

Table 1 Comparison of accuracies for different parameter λ values on three datasets: Flat, Fold, and Rand. The last column shows the accuracy using the ESZSL method on the three datasets. The last row shows the average accuracy calculated for each method.

Dataset	Our method with different λ					ESZSL
	0	0.1	1	10	ln f	
Flat	63.57	**67.01**	60.14	65.29	64.60	57.38
Fold	**89.41**	86.13	87.22	84.12	83.20	76.03
Rand	73.40	**82.30**	79.30	74.60	73.40	71.23
Average	75.45	**78.48**	75.55	74.67	73.73	68.21

ln f to represent the case in which the objective function only contains the term $\mathcal{C}$.

(3) For the three datasets, the performance on the *Fold* dataset is better than those for *Flat* and *Rand*. This result is consistent with that reported in Ref. [22].

5.2 Analysis

We found that the accuracy increases when λ changes from 0 to 0.1 in the Rand and Flat datasets. The Fold dataset differs slightly and achieves the best accuracy when we employ only classification loss. On average, it performed best when $\lambda = 0.1$. In the Rand dataset, which most often reflects real life, the accuracy is obviously increased when regression loss is incorporated. This result indicates that the best accuracy is achieved when the loss combines both kinds of losses, rather than just one or the other. We consider that in the Rand or Flat dataset, the best accuracy is achieved by a compromise between the two losses. Classification loss enables discrimination between data from different labels, and regression loss brings the predicted attribute values close to the actual values. The use of a single classification loss may not be able to fit the attribute values well and a single regression loss may result in overfitting. Using both losses, we can find the optimal combination.

6 Conclusion

In this work, we addressed the zero-shot learning problem for tactile material recognition by incorporating advanced semantic information into the training model. To do so, we used a deep learning method. Our main technical contribution is our proposal of an end-to-end deep learning framework to solve the tactile zero-shot-learning problem. In this framework, we use a CNN to extract the spatial feature and an LSTM to extract the temporal features to generate dynamic tactile sequences. We also developed a loss function that is suitable for the zero-shot-learning setting. The results of our experimental evaluations on publicly available datasets demonstrate the effectiveness of the proposed method.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Nos. 61673238, 61703284, and 61327809), in part by the Beijing Municipal Science and Technology Commission (No. D171100005017002), and in part by the Nanning Science Research and Technology Development Plan.

References

[1] A. Vicente, J. D. Liu, and G. Z. Yang, Surface classification based on vibration on omni-wheel mobile base, in *Proc. 2015 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Hamburg, Germany, 2015, pp. 916–921.

[2] S. Baglio, L. Cantelli, F. Giusa, and G. Muscato, Intelligent prodder: Implementation of measurement methodologies for material recognition and classification with humanitarian demining applications, *IEEE Trans. Instrum. Meas.*, vol. 64, no. 8, pp. 2217–2226, 2015.

[3] S. Chitta, J. Sturm, M. Piccoli, and W. Burgard, Tactile sensing for mobile manipulation, *IEEE Trans. Rob.*, vol. 27, no. 3, pp. 558–568, 2011.

[4] Y. F. Hou, D. H. Zhang, B. H. Wu, and M. Luo, Milling force modeling of worn tool and tool flank wear recognition in end milling, *IEEE/ASME Trans. Mechatron.*, vol. 20, no. 3, pp. 1024–1035, 2015.

[5] M. Khadem, C. Rossa, N. Usmani, R. S. Sloboda, and M. Tavakoli, A two-body rigid/flexible model of needle steering dynamics in soft tissue, *IEEE/ASME Trans. Mechatron.*, vol. 21, no. 5, pp. 2352–2364, 2016.

[6] H. P. Liu and F. C. Sun, Material identification using tactile perception: A semantics-regularized dictionary learning method, *IEEE/ASME Trans. Mechatron.*, vol. 23, no. 3, pp. 1050–1058, 2018.

[7] S. Shirmohammadi and A. Ferrero, Camera as the instrument: The rising trend of vision based measurement, *IEEE Instrum. Meas. Mag.*, vol. 17, no. 3, pp. 41–47, 2014.

[8] J. Degol, M. Golparvar-Fard, and D. Hoiem, Geometry-informed material recognition, in *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 1554–1562.

[9] S. Bell, P. Upchurch, N. Snavely, and K. Bala, Material recognition in the wild with the materials in context database, in *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3479–3487.

[10] Q. L. Wang, P. H. Li, W. M. Zuo, and L. Zhang, RAID-G: Robust estimation of approximate infinite dimensional Gaussian with application to material recognition, in *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4433–4441.

[11] M. Brandao, Y. M. Shiguematsu, K. Hashimoto, and A. Takanishi, Material recognition CNNs and hierarchical planning for biped robot locomotion on slippery terrain, in *Proc. 2016 IEEE-RAS 16th Int. Conf. on Humanoid Robots*, Cancun, Mexico, 2016, pp. 81–88.

[12] H. P. Liu, Y. L. Yu, F. C. Sun, and J. Gu, Visual–tactile fusion for object recognition, *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 2, pp. 996–1008, 2017.

[13] H. O. Liu, J. Qin, F. C. Sun, and D. Guo, Extreme kernel sparse learning for tactile object recognition, *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4509–4520, 2017.

[14] A. Kimoto and Y. Matsue, A new multifunctional tactile sensor for detection of material hardness, *IEEE Trans. Instrum. Meas.*, vol. 60, no. 4, pp. 1334–1339, 2011.

[15] H. B. Liu, X. J. Song, J. Bimbo, L. Seneviratne, and K. Althoefer, Surface material recognition through haptic exploration using an intelligent contact sensing finger, in *Proc. 2012 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Vilamoura, Portugal, 2012, pp. 52–57.

[16] T. Bhattacharjee, J. Wade, and C. Kemp, Material recognition from heat transfer given varying initial conditions and short-duration contact, in *Proc. Robotics: Science and Systems*, Rome, Italy, 2015.

[17] J. Christie and N. Kottege, Acoustics based terrain classification for legged robots, in *Proc. 2016 IEEE Int. Conf. on Robotics and Automation*, Stockholm, Sweden, 2016, pp. 3596–3603.

[18] S. A. Shevchik, F. Saeidi, B. Meylan, and K. Wasmer, Prediction of failure in lubricated surfaces using acoustic time-frequency features and random forest algorithm, *IEEE Trans. Ind. Inf.*, vol. 13, no. 4, pp. 1541–1553, 2017.

[19] A. Watanabe, J. Even, L. Y. Morales, and C. Ishi, Robot-assisted acoustic inspection of infrastructures-cooperative hammer sounding inspection, in *Proc. 2015 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Hamburg, Germany, 2015, pp. 5942–5947.

[20] T. Fukuda, Y. Tanaka, M. Fujiwara, and A. Sano, Softness measurement by forceps-type tactile sensor using acoustic reflection, in *Proc. 2015 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Hamburg, Germany, 2015, pp. 3791–3796.

[21] S. S. Baishya and B. Bäuml, Robust material classification with a tactile skin using deep learning, in *Proc. 2016 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Daejeon, Korea, 2016, pp. 8–15.

[22] W. Z. Yuan, S. X. Wang, S. Y. Dong, and E. Adelson, Connecting look and feel: Associating the visual and tactile properties of physical materials, arXiv preprint arXiv: 1704.03822, 2017.

[23] W. Z. Yuan, C. Z. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, Shape-independent hardness estimation using deep learning and a gelsight tactile sensor, arXiv preprint arXiv: 1704.03955, 2017.

[24] K. Drewing, C. Weyel, H. Celebi, and D. Kaya, Feeling and feelings: Affective and perceptual dimensions of touched materials and their connection, in *Proc. 2017 IEEE World Haptics Conf.*, Munich, Germany, 2017, pp. 25–30.

[25] D. Hughes, A. Krauthammer, and N. Correll, Recognizing social touch gestures using recurrent and convolutional neural networks, in *Proc. 2017 IEEE Int. Conf. on Robotics and Automation*, Singapore, 2017, pp. 2315–2321.

[26] 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, *Rob. Autonom. Syst.*, vol. 63, pp. 279–292, 2015.

[27] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, Deep learning for tactile understanding from visual and haptic data, in *Proc. 2016 IEEE Int. Conf. on Robotics and Automation*, Stockholm, Sweden, 2016.

[28] H. P. Liu, Y. P. Wu, F. C. Sun, D. Guo, and B. Fang, Multi-label tactile property analysis, in *Proc. 2017 IEEE Int. Conf. on Robotics and Automation*, Singapore, 2017, pp. 366–371.

[29] H. P. Liu, Y. P. Wu, F. C. Sun, B. Fang, and D. Guo, Weakly paired multimodal fusion for object recognition, *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 2, pp. 784–795, 2018.

[30] X. C. Feng, L. F. Huang, B. Qin, Y. Lin, H. Ji, and T. Liu, Multi-level cross-lingual attentive neural architecture for low resource name tagging, *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 633–645, 2017.

[31] H. X. Chen, S. Feng, X. Pei, Z. Zhang, and D. Y. Yao, Dangerous driving behavior recognition and prevention using an autoregressive time-series model, *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 682–690, 2017.

[32] Z. Abderrahmane, G. Ganesh, A. Cherubini, and A. Crosnier, Zero-shot object recognition based on haptic attributes, in *Proc. 2017 IEEE Int. Conf. on Robotics and Automation,* Singapore, 2017.

[33] C. H. Lampert, H. Nickisch, and S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in *Proc. 2009 IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 951–958.

[34] Z. M. Zhang and V. Saligrama, Zero-shot learning via joint latent similarity embedding, in *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 6034–6042.

[35] Y. W. Fu, T. M. Hospedales, T. Xiang, and S. G. Gong,

Transductive multi-view zero-shot learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2332–2345, 2015.

[36] S. Changpinyo, W. L. Chao, B. Q. Gong, and F. Sha, Synthesized classifiers for zero-shot learning, in *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 5327–5336.

[37] B. Romera-Paredes and P. H. S. Torr, An embarrassingly simple approach to zero-shot learning, in *Proc. 32nd Int. Conf. on Int. Conf. on Machine Learning*, Lille, France, 2015, pp. 2152–2161.

[38] E. Kodirov, T. Xiang, and S. G. Gong, Semantic autoencoder for zero-shot learning, in *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017.

[39] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, Zero-shot learning by convex combination of semantic embeddings, arXiv preprint arXiv: 1312.5650, 2013.

[40] R. Fakoor, M. Bansal, and M. R. Walter, Deep attributebased zero-shot learning with layer-specific regularizers, in *NIPS 2015 Workshop on Transfer and Multi-Task Learning*, Montreal, Canada, 2015.

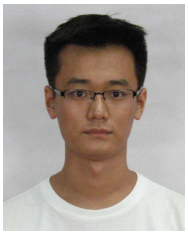
[41] P. Morgado and N. Vasconcelos, Semantically consistent regularization for zero-shot recognition, arXiv preprint arXiv: 1704.03039, 2017.

[42] C. H. Lampert, H. Nickisch, and S. Harmeling, Attribute-based classification for zero-shot visual object categorization, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, 2014.

Huaping Liu received the PhD degree in computer science from Tsinghua University in 2004. He is an associate professor with the Department of Computer Science and Technology, Tsinghua University. His current research interests include robotic perception and learning. He serves as an associate editor of some journals including *Cognitive Computation, Neurocomputing, IEEE Transactions on Industrial Informatics*, *IEEE Transactions on Automation Science and Engineering*, and *IEEE Robotics and Automation Letters*, and some conferences including ICRA and IROS. He also served as the program committee member of RSS, IJCAI, and AAAI.

Fuchun Sun received the PhD degree in computer science from Tsinghua University in 1997. He is a full professor with the Department of Computer Science and Technology, Tsinghua University, China. His current research interest includes robotic perception and cognition. He was a recipient of the National Science Fund for Distinguished Young Scholars. He serves as an associate editor of a series of international journals including *IEEE Transactions on Systems, Man and Cybernetics: Systems*, *IEEE Transactions on Fuzzy Systems, Mechatronics*, and *Robotics and Autonomous Systems*. He is a fellow of the IEEE.

Haihong Pan received the PhD degree in mechatronic engineering from Huazhong University of Science and Technology in 2007. He is currently a professor of mechanical engineering, Guangxi University. He is also a member of Guangxi Institute of Mechanical Engineering. He has published more than 60 technical papers. More than 50 patents and 10 software copyrights have been granted. His research interests include electromechanical control theory, robotics, artificial intelligence, machine vision, and PMSM servo drive motor control theory.



Feng Wang received the BS degree from Shandong University in 2017. He is currently pursuing the MS degree in Tsinghua University. His research interests include cross modal retrieval and transfer learning.