2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)
Beijing, China, November 6-9, 2018

Superhuman Performance in Tactile Material Classification and Differentiation with a Flexible Pressure-Sensitive Skin

Andreea Tulbure Berthold Bäuml

***Abstract*— In this paper, we show that a robot equipped with a flexible and commercially available tactile skin can exceed human performance in the challenging tasks of material classification, i.e., uniquely identifying a given material by touch alone, and of material differentiation, i.e., deciding if the materials in a given pair of materials are the same or different. For processing the high dimensional spatio-temporal tactile signal, we use a new tactile deep learning network architecture TactNet-II which is based on TactNet [1] and is significantly extended with recently described architectural enhancements and training methods. TactNet-II reaches an accuracy for the material classification task as high as 95.0%. For the material differentiation a new Siamese network based architecture is presented which reaches an accuracy as high as 95.4%. All the results have been achieved on a new challenging dataset of 36 everyday household materials.**

In a thorough human performance experiment with 15 subjects, we show that the human performance is significantly lower than the robot's performance for both tactile tasks.

I. Introduction

For autonomous robots to be able to robustly and dextrously act in physical contact with their environment, the sense of touch is indispensable. A challenging example is the dextrous manipulation with multi-fingered hands where for the dynamical contact situation a feedback signal with high force resolution in combination with a high spatial and temporal resolution is needed. In fact, it is widely accepted that a key prerequisite for closing the large gap in manipulation performance between humans and robots is to come closer to humanlike performance in robotic tactile sensing [2] [3].

A task which can clearly demonstrate the capabilities of a tactile sensor with respect to its force and temporal resolution and, to a lesser extent, to its spatial resolution is the identification of an object's material by only gently touching or sweeping over its surface.

In this paper, we show that we can exceed human performance in tactile material classification and material differentiation using only the spatio-temporal force signal of a flexible tactile skin mounted on the hand of a humanoid robot. The sensor used is a commercially available and geometrically configurable tactile foil sensor from Tekscan [4] which could be easily mounted (e.g., glued) on the surface of any robotic system. This is an important advantage compared to other tactile sensors, e.g., the bulky sensor like the BioTac [5] [6] for which parts of the robot hand's structure have to be replaced or stiff sensors, e.g., [7], which can not be mounted on curved or elastic surfaces. Another example of a flexible sensor mounted on a soft surface is iCub's finger tip sensor [8].

The authors are with the DLR Institute of Robotics and Mechatronics, Münchnerstr. 20, 82234 Wessling, Germany.
berthold.baeuml@dlr.de

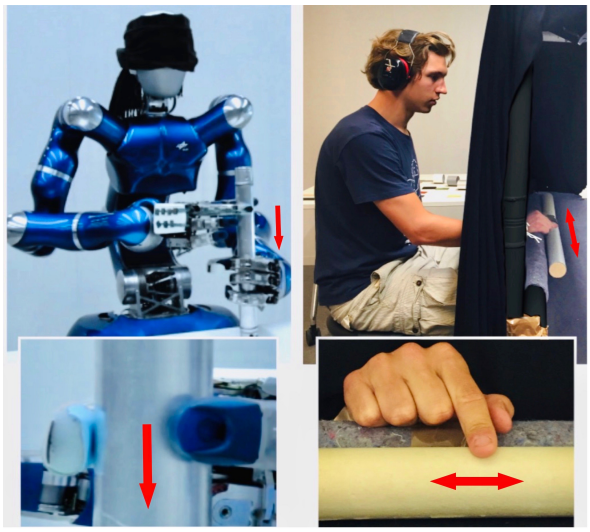

Fig. 1. Robot vs. human: DLR's Agile Justin [9] (left) and a human subject (right) performing a sweeping motion to identify the material on the surface of a given tube. Only the sense of touch can be used: the visual cue is removed for the human by presenting the tubes behind a curtain and the robot is blindfolded symbolically; the auditorial cue is removed by an ear protector for the human (Agile Justin does not have a microphone). Agile Justin is equipped with two DLR Hand-II [10] and to the soft finger tip of the index finger of the right hand a flexible tactile skin with a 4×4 taxel array is attached (see Fig. 2 for details). The procedure for exploring a given tube is performed autonomously by the robot: grasp the tube with the left hand to stabilize it; grasp the tube with the thumb and the index finger of the right hand and slide down along the tube with the right hand at a constant velocity of 3 cm/s for 2 s. The force is held roughly at 1 N (precision about 20%) using the hand's joint torque sensors. In the human performance experiment, the subject performs a similar sliding motion with the index finger of its dominant hand, only horizontally on a lying tube.

To reach superhuman performance, we use a modern deep learning architecture and training methods for processing the complex spatio-temporal tactile signal.

For comparing the tactile performance of humans and the robot, we conduct human performance experiments on a new representative set of 36 materials typically found in everyday household environments. One task is tactile classification where the unique identity of the touched material has to be reported. This task measures the more high level or cognitive performance. The second task is material differentiation in which pairs of materials are presented and the human subject has to report whether both materials are the same or different. This task measures the raw sensor and low level processing

978-1-5386-7282-2/18/$31.00 ©2018 IEEE

performance as no long-term memory is involved.

A. Related Work

The seminal work in tactile material classification is Fishel and Loeb [11] using the multi-modal BioTac [5] sensor (including static and dynamic pressure, temperature and heat flow). They report to classify $C = 117$ different materials based on $n = 15$ training samples per material with an accuracy of 95.4% (but needing 5 trial motions on average) using Gaussian classification on hand-designed features. However, the sample data is acquired using a precisely controlled test bench setup and the performance degrades dramatically when transferred to a real robotic setup [12].

Fishel and Loeb [11] also conduct a small material differentiation experiment where five human subjects had to discriminate the materials in each of eight pairs of materials (which were informally selected out of the $C = 107$ materials by the authors as being the hardest to discriminate).

In our previous work [1], we show that robust material classification using a flexible tactile skin on a real robotic setup is feasible with deep learning on the raw 24000 dimensional signal, i.e., without any preprocessing. This can be regarded as a proof of concept where an accuracy of up to 97.3% is reached but for only $C = 6$ material classes and $n = 80$ samples per material.

In the same work [1], we also give an overview of related work on tactile material classification which we summarize here in Table I supplemented by more recent work shortly described below.

Gao et al. [13] present a deep learning convolutional neural network (HapticNet) to classify samples recorded from two BioTac sensors mounted on a robot gripper during five different exploration motions. Strese et al. [14] slide a self-made multi-modal pen by hand over materials recording 25 s long time series. Using only the haptic, i.e., acceleration, signal (and not, e.g., the also recorded images) they reach an accuracy of 39% using a Bayesian classifier on the signal's spectrogram [15]. Eguiluz et al. [16] use only the vibration channel of the BioTac sensor in a well controlled turntable setup where for each material in one 5 min sweep all tactile data is recorded. For continuous material classification they use a hidden Markov-Model based on features learned with principal component analysis from the the Fourier transformed sensor signal.

For the classification of $C = 14$ materials with a BioTac sensor in a test bench setup, Kerr et al. [17] reach an accuracy of 79% using surface texture and thermal properties. The authors apply principal component analysis (PCA) to extract important features from the sensor data and a simple neural network for learning them. They also conduct a material classification experiment with 12 human subjects in which the BioTac sensor based accuracy exceeds the human performance.

Multi-channel neural networks are used by Kerzel et al. [18] to classify a set of $C = 32$ materials based on the 3D force signals of an OptoForce sensor [19] and in a test bench setup. They reach an accuracy of 99% but the performance dramatically drops to $68\dot{\%}$ when artificial Gaussian noise is added to simulate a real world robotic setup.

TABLE I
RELATED WORK IN TACTILE MATERIAL CLASSIFICATION

paper	sensor	setup	C	n	a [%]
[11]	BioTac	test bench	127	15	95.4
[12]	BioTac	**ShadowHand**	10	15	99.0
[21]	BioTac	**PA10 arm**	49	50	97.0
[22]	stiff skin	**Nao**	5	30	100
[23]	BioTac	**ShadowHand**	20	10	83.5
[15]	accelerometer	**robot**	20	50	65.7
[24]	accelerometer	test bench	8	120	89.0
[13]	BioTac	**PR2**	53	10	83.2
[14]	pen	manual	69	188	39.0
[17]	BioTac	test bench	14	15	79.1
[16]	BioTac	test bench	34	eff. 300[1]	100
[18]	OptoForce	test bench	32	100	98.8
[25]	GelSight	manual	40	24	99.8
[1]	**flexible skin**	**Agile Justin**	6	80	97.3
this	**flexible skin**	**Agile Justin**	36	100	86.3

Erickson et al. [20] use semi-supervised learning to recognize the material classes of 72 household objects by touch (no sweeping, only moving until contact) based on the force, vibration and thermal flow sensor signal of a bulky sensor with no spatial resolution mounted on the PR2 robot's gripper. With 100 samples per object but for only $C = 6$ material classes they reach an accuracy of up to $96\dot{\%}$ using all $n = 1200$ samples per material in material classification.

In summary, most works on tactile material classification use test bench setups (see Table I) although those results do not transfer well to the noisy environment of a real robot system. In this paper, we use the humanoid robot Agile Justin which results in significant variation in the collected tactile samples (see Fig. 2).

Our previous work [1] and this paper are still the only ones which use solely the signal of a flexible tactile skin for material classification, hence, a sensor which has the aforementioned advantages with regard to providing a high resolution spatio-temporal signal for dextrous manipulation and its ease of mounting.

To our knowledge, there are no comparison studies of tactile material recognition performance between humans and real robotic systems. But also for the case of test bench setups, only the two works of Fishel and Loeb [11] and Kerr et al. [17] using the BioTac sensor conduct small experiments for a comparison with human performance.

B. Contributions

- We show superhuman performance for both, tactile material classification and material differentiation with a set of 36 everyday household materials and in a real robot setup using a commercially available flexible tactile skin.
- To our knowledge, we conduct for the first time an extensive human performance study for a combination of tactile classification (high level) and differentiation

[1]The 5 min recording time results in effective 300 samples assuming 1 s sweeping time as we use in our experiments.

(low level) performance and in comparison to the performance of a real robot setup.
- We present our extended deep learning architecture using modern training methods for substantially increasing the accuracy compared to the TactNet network [1].
- We present a network architecture for material differentiation based on a Siamese network.
- We provide a new tactile dataset with 3600 samples (100 samples per material) for 36 everyday household materials recorded with the tactile skin as a public benchmark dataset.

II. Sample Data Set and Experimental Setup

1) Robot setup: Fig. 1 and Fig. 2 summarize the experimental setup for recording the tactile data with a flexible tactile skin mounted on a robot's finger tip. The robot setup is similar to the one in [1] with the main difference that we improved the method for mounting the skin onto the finger tip allowing for long-term use ($> 10^4$ sweeps) of the skin without replacement. Hence, other than in [1], our focus here is not robustness when replacing and reattaching the sensor but robustness against drift in the robot setup (mainly the drift in the finger's torque sensors) during long-term experiments. For this, the dataset was recorded over a time period of four days and on each day for each material the same amount of samples were recorded.

2) Sample dataset: The $C = 36$ everyday household materials of our new sample data set are depicted in Fig. 3. The materials are glued to tubes or the tube is made of the material[2]. For each material tube, we record $n = 100$ samples resulting in $N = 3600$ samples overall. The tube is randomly rotated after each sweep and after 10 sweeps a different material is chosen. We made the dataset publicly available at [26].

3) Cross validation: To evaluate the classification performance we always use a variation of stratified cross validation [27] such that, e.g., the training set has the exactly equal number of samples for each material. Throughout this paper, we use 5-fold cross-validation with 2 runs, i.e., the $n = 100$ samples per class are split in folds of 20 test samples and 80 samples for learning. When we perform hyperparameter search the learning dataset with the 80 samples is further split, again using 5-fold cross-validation, into 64 training samples and 16 validation samples. This scheme guarantees that the test samples of a given (outer) fold are never used for training or hyperparameter optimization.

Performing a r-run k-fold cross validation results in rk accuracy results from which we compute the mean accuracy $\bar{a}$ and its standard deviation σ. This σ represents the uncertainty of the mean accuracy which is a combination of the uncertainty σ_{test} due to having only a finite number N_{test} of test samples and the uncertainty σ_{over} due to potential overfitting problems. σ_{test} can be computed separately by identifying the mean accuracy as a sum of Bernoulli distributed random variables, one for each test sample. This results in $\sigma_{\text{test}} = \sqrt{1/N_{\text{test}}\bar{a}(1-\bar{a})}$.

[2]Only for (9, 10, 13, 14, 19, 20, 34) the tubes are made of the materials.

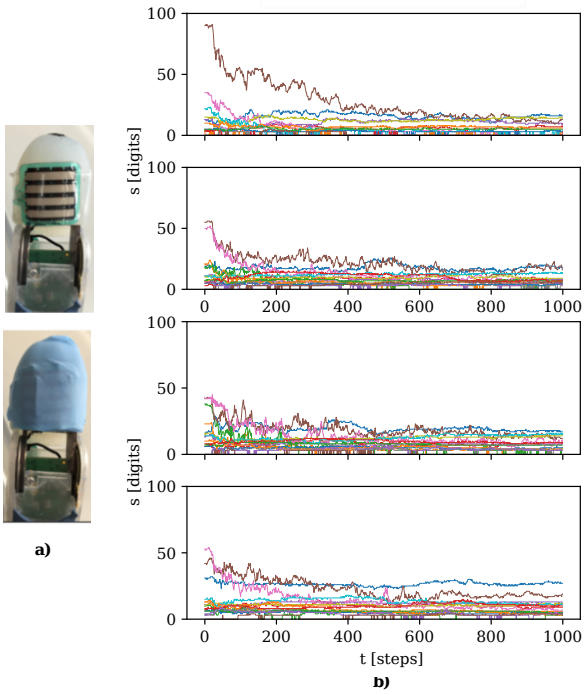

Fig. 2. **a)** The finger tip with the flexible tactile skin taped to it (upper) and the usage of a thin laboratory glove on top of the tactile skin to increase the grip (lower). The tactile skin is a 4×4 taxel array sensor, commercially available from Tekscan [4] (VersaTek® sensor 4256E), which provides a spatio-temporal pressure signal at 750 Hz sample rate. **b)** Four plots of the raw spatio-temporal signal of the tactile skin during the middle 1.33 s of the 2 s exploration motion along the tube. The 750 Hz sample rate of the sensor results in a $1000 \times 4 \times 4 = 16000$ dimensional sample, which is used without any pre-processing in our end-to-end deep learning method. To demonstrate how challenging material classification is in this real world robotic setup (in distinction to a test bench setups), the upper two plots depict two samples of the same material class (white metal) which our network correctly recognizes as "same" and the lower two plots show samples for two different materials (cotton fabric reverse and linen fabric smooth) which our network correctly classifies as "different". From looking at the plots it seems the other way around: the lower two samples look more similar than the upper two.

4) 1-sweep & 3-sweep exploration: The classification accuracy can be increased by sweeping multiple times over the given material. In this work, we evaluate the 3-sweeps case in comparison to the 1-sweep case. We simulate the three sweeps by randomly drawing three samples (x_1, x_2, x_3) of the same class t from the test dataset and the network computes the individual predictive probability distribution $p(t'|x_i)$ for each of the samples. Given $p(t'|x_i)$, we use two different schemes for computing the final reported class.

- Majority voting: compute the 1-sweep class $t_i = \arg\max_{t'} p(t'|x_i)$ for each sample and report the class t_i which occurred most often, in case of three different t_i, choose one randomly.
- Bayesian fusion: Because of $p(t'|x_1, x_2, x_3) \propto p(t'|x_1)p(t'|x_2)p(t'|x_3)$ in case of independent samples and same number of samples per class, report the label $t = \arg\max_{t'} p(t'|x_1)p(t'|x_2)p(t'|x_3)$.

Fig. 3. The 36 everyday household materials. (0) *synthetic leather rough*, (1) *synthetic leather smooth reverse*, (2) *synthetic leather smooth*, (3) *metallic jersey*, (4) *cotton fabric reverse*, (5) *cotton fabric*, (6) *linen fabric smooth reverse*, (7) *velours paper reverse*, (8) *linen fabric smooth*, (9) *wood*, (10) *white metal*, (11) *reflecting fabric reverse*, (12) *reflecting fabric*, (13) *metal*, (14) *plastic smooth*, (15) *latex*, (16) *silicon*, (17) *jersey*, (18) *velours paper*, (19) *wallpaper*, (20) *plastic rough*, (21) *spun fleece*, (22) *synthetic leather rough reverse*, (23) *cork*, (24) *linen fabric rough*, (25) *gunny*, (26) *carton*, (27) *denim*, (28) *carpet rough*, (29) *carpet smooth*, (30) *metallic grid*, (31) *rubber rough vertical*, (32) *rubber smooth*, (33) *rubber rough horizontal*, (34) *foam*, (35) *neoprene*. Source: Materials 10 and 13 are from Alutruss (www.alutruss.com); 9, 14, 19, 20 from a standard h/w store; remaining from Modulor (www.modulor.de), a shop for designers.

III. Deep Learning Methods for Material Classification and Differentiation

A. Material Classification

In previous work [1], it has been shown that classification of the high dimensional spatio-temporal tactile signal based on deep learning in an end-to-end learning setting results in superior performance compared to more classical two step approaches. In those classical learning approaches, first a separate feature extraction step is performed resulting in low dimensional features which are then fed into a classifier, e.g., support vector machines or k-nearest neighbor.

The neural network architecture from [1], which we call TactNet, is only used for a rather small set of 6 different materials. Here, we take this original deep learning architecture as a starting point for our significantly more challenging classification task with 36 materials and extend it with recently reported architectural enhancements and training methods. Finally, we perform an extensive hyperparameter and architecture optimization via random grid search and using the procedure described in Sec. II-.3 with careful distinction between training, validation and test sets.

1) Base network: The 3D spatio-temporal signal is first converted into a 2D input signal by flattening the spatial dimensions into one dimension as tests have shown that using the full 3D signal results in no better performance

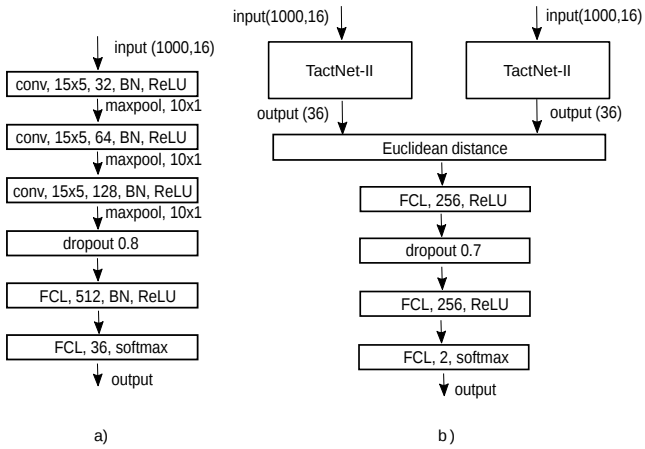

Fig. 4. **a)** Architecture of the TactNet-II network as used for tactile material classification. **b)** Adapted Siamese network for material differentiation. It is built from two pre-trained TactNet-II networks for mapping the inputs into an abstract feature space where the Euclidean distance is computed.

but is computationally less efficient in the here used deep learning software framework TensorFlow [28]. The signal is then fed into a stack of convolutional and max-pooling layers which implicitly perform feature extraction. Then the signal is fed into a fully connected layer followed by a softmax layer for the classification. Other than in [1], in each convolutional layer batch normalization [29] is performed before the application of the activation function. This allows for a more robust training irrespective of the variation in the samples' statistics. The final base network after the architectural optimization is depicted in Fig. 4. In the architectural search the number of convolutional layers, the kernel the size, and size of the fully connected layers were optimized.

The network was trained by minimizing a standard cross entropy loss function with a L2 regularization using an Adam optimizer [30] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. A standard learning rate scheduler is used which decreases the learning rate after 170 epochs by a factor of 10 from $\lambda = 10^{-4}$ to $\lambda = 10^{-5}$.

2) Adversarial training: Adversarial training [31] was originally developed to make a classifier more robust against adversarial attacks. But it can also be seen as a smart technique for efficient data augmentation with random noise. In random noise data augmentation, for each training sample x a number of perturbed samples $\tilde{x} = x + \epsilon\eta$ are added to the training set where the same class label t as for the original sample x is assumed. $\eta \in [0, 1]$ is a uniformly distributed random variable and ϵ is the noise scaling factor (e.g., in the order of sensor noise). This data augmentation makes the classifier robust against typical real world perturbations and, hence, is a kind of regularization.

The trick in adversarial training is, instead of augmenting with many random samples (which is very inefficient in high dimensional input spaces), to only add the worst case perturbation $\tilde{r} \leq |\epsilon|$, i.e., the one which changes the per sample loss function $E(x, t, \theta)$ for training the classifier's parameters θ the most. The adapted loss function for adversarial training

reads then

$$\tilde{r}(x,t,\theta) = \epsilon \,\mathrm{sign}(\nabla_x E(x,t,\theta))$$
$$\tilde{E}(x,t,\theta) = \alpha E(x,t,\theta) + (1-\alpha)E(x+\tilde{r}(x,t,\theta),t,\theta) \quad ,$$

where α is a weighting factor for the contributions of the original and adversarial sample and set to $\alpha = 0.5$ for all our experiments.

3) Monte Carlo dropout model: The Monte Carlo dropout model (MC dropout) [32] is a recent method for efficient Bayesian learning in deep neural networks. MC dropout reinterprets in a variational inference scheme the usual standard dropout as drawing samples of the network weights W from an approximate posterior distribution $p(W|X,T) \approx q_\phi(W)$ given the training samples and labels (X,T). The resulting loss function for optimizing the parameters ϕ of the approximator $q_\phi(W)$ is exactly the same as the loss function for standard learning with dropout. But during prediction, the dropout is kept switched on and for a given input x, multiple runs M through the network are performed (each with a new $W \sim q_\phi(W)$ via dropout). This results in a Monte Carlo approximation of the full Bayesian predictive distributions

$$p(t|x,X,T) = \int p(t|x,W)p(\theta,X,T)\mathrm{d}W$$
$$\approx \sum_m^M p(t|x,W^{(m)}), \text{ with } W^{(m)} \sim q_{\phi^*}(W^{(m)}).$$

The exact Bayesian predictive distribution would represent the correct prediction uncertainty, i.e., combined model uncertainty and noise, and would not suffer from overfitting. But the MC dropout approximation usually gives also better prediction than the point estimate of standard non-Bayesian deep learning. An important parameter is the dropout rate which we optimize in the hyperparameter search. The number of runs is set to $M = 100$ for all our experiments.

Table II summarizes all architectural and other hyperparameters that are optimized via random grid search.

TABLE II
OPTIMIZED HYPERPARAMETERS FOR TACTNET-II

FCL size	512
kernel size	15x5
max pool size	10x1
dropout (MC-dropout)	0.80 (0.84)
batch size	36
regularizer	L2, 10^{-3}
learning rate	10^{-4} (10^{-5})
ϵ (adversarial)	0.1

B. Classification Results

1) Network comparison: Table III reports the accuracies for the network architectures and training methods including the original TactNet [1] on our new dataset with 36 materials using the evaluation method as described in Sec. II-.3. For the 1-sweep as well as for the 3-sweeps case, our extended and by random grid search optimized base network performs significantly (about 10 %) better than the original TactNet. The advanced adversarial training and MC dropout training methods further increase the accuracy by 0.8 % up to 86.3 % in the 1-sweep case. This clearly proofs that tactile material classification with the flexible tactile skin is feasible on a large set of everyday materials and that the advanced network architecture and training methods are the key to this high performance. We name the new network architecture TactNet-II. For the 3-sweep case, we use the Bayeisan fusion scheme from Sec. II-.4 and the accuracy gets as high as 95.0 %.

2) Confusion matrix and grouping: The confusion matrix C in Fig. 5 and its diagonal values show that some materials are easier (e.g., the rubbers) and some are harder (e.g., the leathers) than average to classify.

For further analysis, we use spectral clustering on the confusion matrix C by setting the affinity matrix to $A = \frac{1}{2}((C-1)+(C-1))^{\mathrm{T}}$ to get 6 groups of materials which are "confused the most" with each other. The materials are then ordered such that for each group the materials in a group have consecutive material IDs. Actually, this order is used in all our figures including the Fig. 3 of the materials and the Fig. 5 of the confusion matrix. In the latter, one can clearly see that the hard to identify materials (especially the leathers) get only confused with one another.

3) No overfitting: As is described in Sec. II-.3, the standard deviations σ we report are computed via cross-validation and represent the combined uncertainty due to potential overfitting σ_{over} and due to the finite number of test samples per fold σ_{test}. According to Sec. II-.3, $\sigma_{\text{test}} = 1.4\,\%$ for TactNet-II and the 1-sweep case (with $36 \cdot 20 = 720$ test samples). Comparing this to Table III, it is clear that there is no additional uncertainty due to overfitting.

TABLE III
MATERIAL CLASSIFICATION PERFORMANCE

sweeps	network type	$\bar{\mathbf{a}}$ [%]	σ [%]
1	original TactNet	73.1	1.8
	base	85.5	1.5
	adversarial	86.1	1.1
	adversarial + MC dropout (**TactNet-II**)	**86.3**	1.2
3	original TactNet	86.2	2.2
	base	94.3	1.5
	adversarial	94.5	1.4
	adversarial + MC dropout (**TactNet-II**)	**95.0**	0.9

C. Material Differentiation

In material differentiation, we want to learn a function $f(x_a, x_b)$ which takes two input samples x_a and x_b and reports back 1 if both samples are from the same class, i.e., $t_a = t_b$, or 0 if they are from different classes, i.e., $t_a \neq t_b$.

For this, we adapt a Siamese network model [33] which was originally developed in the context of one-shot learning. The idea is, that both input samples are first independently mapped into an abstract feature space and then a distance in this feature space is computed between the input samples. Finally, the distance is mapped with an additional network to the decision probability for "same" or "different".

Fig. 4 shows our Siamese network model for material differentiation. The two sister TactNet-II networks are identical

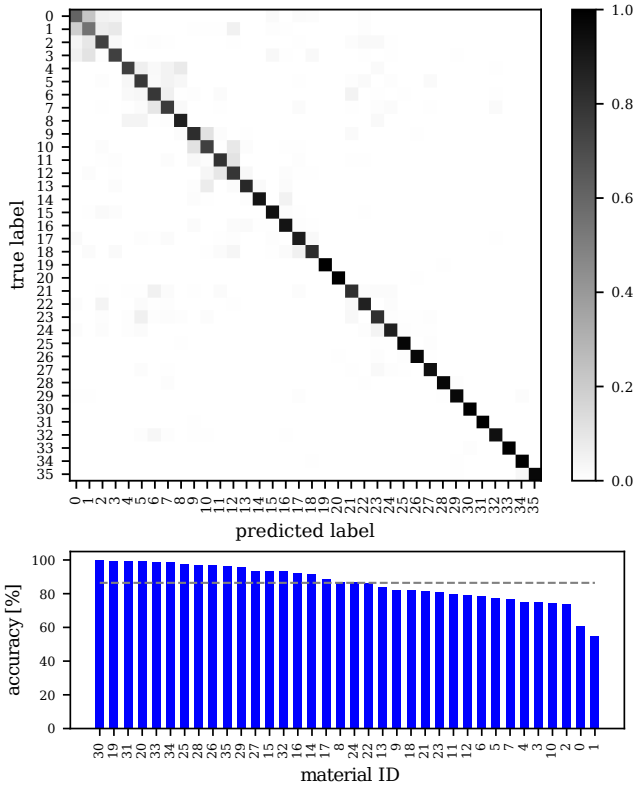

Fig. 5. Confusion matrix (upper) and its diagonal values (lower), hence, the per class accuracy. For the latter, the order is from high to low accuracies. All results are for TactNet-II and the 1-sweep case. The dashed line marks the average accuracy of 86.3% over all materials.

and pretrained from the classification task and their weights are fixed. Only the weights of the fully connected layers after the Euclidian distance layer are trained. We use the contrastive loss function [34] and an Adam optimizer with learning rate $\lambda = 0.001$ for 15 epochs (all other parameters as in Sec. III-A).

D. Differentiation Results

For the performance evaluation of our material differentiation network, we first perform the usual cross validation split into test and training samples as described in Sec. II-.3. Then, for each of the two datasets, we randomly generate sets of material pair samples with an equal number of "same" and "different" pair samples ($3.6 \cdot 10^4$ samples for training and $2.7 \cdot 10^3$ for testing).

Table IV reports the performance results. For the 1-sweep case, the accuracy in material differentiation is about 5 % larger than for material classification, showing that differentiation is a simpler task. For the 3-sweeps case, we had to use majority voting as the output of the Siamese based network is not a proper probability distributions in a Bayesian sense.

Again, we are not suffering from severe overfitting as $\sigma \approx \sigma_{\text{test}}$, the uncertainty due to the number of test samples.

IV. Human Performance Experiments

To have a benchmark for our robotic tactile sensor and processing, we performed human performance experiments for material classification and material differentiation. Fifteen

TABLE IV
Material differentiation performance

sweeps	$\bar{a}$ [%]	σ [%]
1	91.8	0.9
3	95.4	0.9

human subjects, eight males and seven females with age between 21 and 49 years, agreed to participate in the experiments. All participants were tested to have normal touch tactile sensitivity using the "Touch-Test Sensory Evaluation" from North Coast Medical (www.ncmedical.com).

As in the robotic case, also the human subjects should only use tactile information by sweeping a finger over the tubes but no other sensorial cue. For this, we used the experimental setup in Fig. 1: the tubes are presented to the human subject behind a curtain to remove the visual cue. In addition, the subjects had to wear ear protectors to remove the auditorial cue as it turned out that humans can hear for some materials the material identity while sweeping their finger over it.

For each subject, the experiments were conducted on two days, on the first day the material classification and on the second day the material differentiation experiment. All experiments were conducted by the same investigator.

A. Material Classification

The experiments consisted of a training and a testing phase. In the training phase, the subjects had 10 min to get familiar with the materials by touching and sweeping over material samples attached to small plates. During this phase, the materials had to be grouped in five to seven groups according to their subjective tactile similarity. This should help later when performing the tactile classification task. All subjects reported that the 10 min for this phase was more time than they needed. This might have been because our set of materials consists of everyday household materials the subjects were already familiar with. During the training phase, the visual cue could not be excluded as the subjects had to see where to find the material samples and had to reorder them.

The testing phase consisted of three directly successive sub-phases for each material:

1) Sweep once over the presented tube and decide for the material class by looking at the previously grouped material sample plates and telling the number written on them.
2) Sweep three times over the tube in both directions and decide again by looking at the sample plates for the material class.
3) Sweep ones more over the tube and decide by sweeping over the material samples for comparison by touch.

During testing, each material was presented to the subject once, but the subjects were not told about this, and the order of the materials was random. The overall time of an experiment was at maximum 45 min.

B. Classification Results

Fig. 6 reports the accuracies of the human experiment averaged over all subjects in comparison with the robot

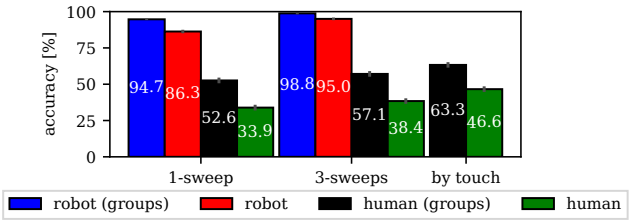

Fig. 6. Human vs. robot material classification accuracy. The depicted standard deviation for the human experiment for the case of the $N = 15 \cdot 36$ overall samples is computed in analogy to σ_{test} in Sec. II-.3.

performance. The human performance is dramatically worse than the robot performance. Even the 1-sweep robot accuracy is 40% higher than the "3-sweep and compare by touch" accuracy of the human. To make the task even simpler, we also report the accuracy for identifying at least the correct material group which the subjects formed individually in the training phase. But even this accuracy is still about 30% worse than the one of the robot for the way harder 1-sweep class identity task. For completeness, Fig. 6 also reports the robot's performance in identifying the correct group using the groups from Sec. III-B.2.

Fig. 7 shows the confusion matrix and the accuracy for each material class. Comparing this to the robot results in Fig. 5 shows that for every single class, the robot reaches at least human accuracy.

This surprisingly bad human performance is compatible with the statements of almost all human subjects: after the training phase (in which they could see the material samples while touching them), they expected the classification task to be way easier than they judged it after they had actually performed the experiment (but were not told about their performance). One explanation for this initial overrating of their tactile capabilities could be that humans almost always use additional visual cues to prime their tactile expectation.

C. Material Differentiation

Due to the high number of $\binom{36}{2} = 630$ possible material pairings, an exhaustive evaluation of the material differentiation performance in a study with human subjects is prohibitive. Therefore, to get at least a lower bound for the material differentiation performance we selected the eight hardest to differentiate material pairs. To have a fair set of the hardest pairs, we selected the four hardest pairs for the human and the four hardest pairs for the robot using the following two criteria based on the classification confusion matrices:

1) The hardest materials are the ones which have the smallest on-diagonal value.
2) The hardest materials are the ones which have the highest off-diagonal value.

For each criterion, we selected two materials from the robot and human confusion matrix. Because one material pair was the same for the robot and human case, only seven actual pairs were used in these experiments: (0, 1), (12, 7), (16, 10), (10, 14), (9, 10), (11, 12), (4, 8). These seven material pairs are made up from 11 different materials. In the experiment,

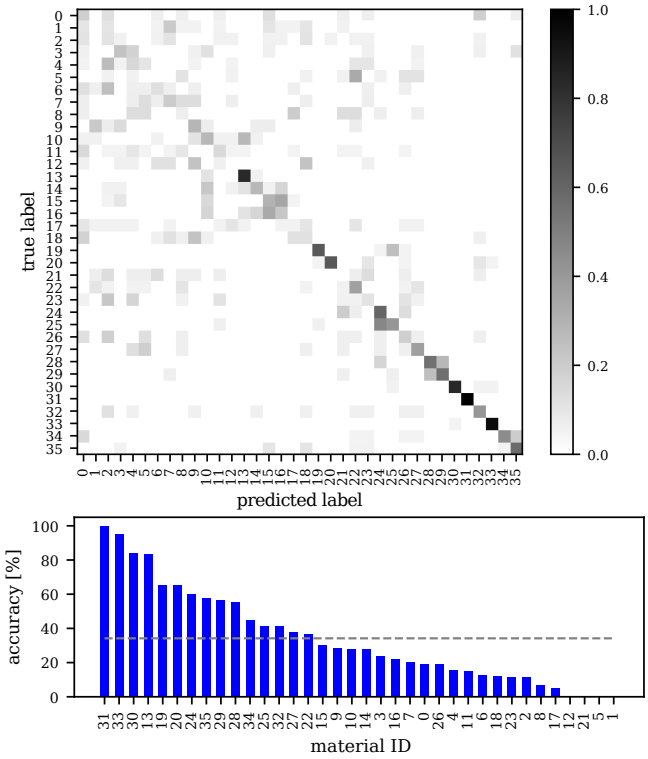

Fig. 7. Confusion matrix (upper) for the human performance experiment and its diagonal values (lower) ordered from high to low accuracies. The dashed line marks the average accuracy of 34.19% over all materials.

we presented the human subjects an equal number of those "different pairs" and of "same pairs" made up from the same 11 materials.

In this experiment, there was no training phase, but the subjects were allowed to make them familiar with all material samples again. The subjects were not told that only pairs from a subset of the materials will be presented to not bias their decision.

In the testing phase, a pair of tubes was presented to the subjects behind the curtain. Each presentation had two sub-phases:

1) Sweep once over each tube of the given pair and decide and say if the materials are the same or different.
2) Perform two additional sweeps in both directions over the first and then two sweeps over the second tube. Finally, it was allowed to sweep once more over the first tube before the decision had to be made.

Each of the "different pairs" were presented twice and an equal number of the "same pairs". The presentation order of the these pairs was random. The overall time of an experiment was at maximum 45 min.

D. Differentiation Results

Fig. 8 reports the human accuracy for the differentiation task in comparison to the robot performance for the selected hardest material pairs. To show that this accuracy is a lower bound for the accuracy computed over all pairs, also the robot accuracy for all pairs is depicted.

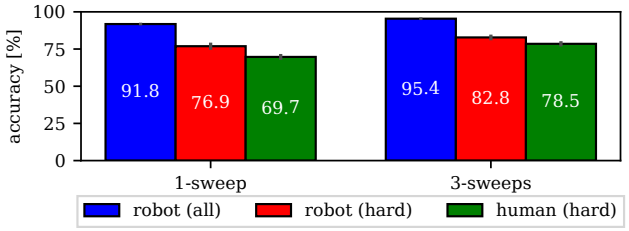

Fig. 8. Human vs. robot material differentiation performance. The "hard" results are for the selected seven hardest material pairs whereas the "all" result is for all possible material pairs.

The robot clearly outperforms the human in material differentiation, although by a smaller margin than for material classification. An interpretation of this finding could be that the human raw sensorial and low level processing tactile performance is good but that human tactile memory is not well trained. It would be interesting to repeat the experiments with a blind subject which is more dependent on its tactile performance.

V. Conclusions

In this paper, we have shown for the first time that for the higher level tactile material classification task as well as for the low level material differentiation task, a robot equipped with a flexible and tactile skin can exceed human performance, hence, reaches superhuman performance.

First, we introduced a new tactile dataset with 3600 overall samples from 36 everyday household materials. We made this dataset publicly available [26]. Then we presented our new TactNet-II deep learning network which is based on TactNet [1] but is extended with recent architectural enhancements and training methods. TactNet-II reaches an accuracy for the material classification task as high as 95.0 %. For material differentiation we used TactNet-II as a building block in a Siamese-like network architecture and reach an accuracy as high as 95.4%.

Finally, we performed a thorough human performance experiment with 15 subjects. In the material classification task, the humans performed poorly with an accuracy of at least 30% lower than the robot. For the low level material differentiation task the human performance was still significantly lower than the robot performance, but by a smaller margin.

In future work, we will concentrate on sample efficiency, i.e., using as little training samples as possible.

Acknowledgments

We thank Bernhard Weber for helping in the design of the human performance experiment and all who participated in it.

References

[1] S. Baishya and B. Bäuml, "Robust material classification with a tactile skin using deep learning," in *Proc. ICRA*, 2016.
[2] R. S. Dahiya *et al.*, "Directions toward effective utilization of tactile skin: A review," *IEEE Sensors Journal*, vol. 13, no. 11, 2013.
[3] Z. Kappassov, J. A. C. Ramon, and V. Perdereau, "Tactile sensing in dexterous robot hands - review," *Robotics and Autonomous Systems*, vol. 74, no. Part A, pp. 195–220, 2015.
[4] Tekscan. [Online]. Available: https://www.tekscan.com
[5] N. Wettels, V. Santos, R. Johansson, and G. Loeb, "Biomimetic tactile sensor array," *Advanced Robotics*, vol. 22, no. 8, pp. 829–849, 2008.
[6] SynTouch. [Online]. Available: http://www.syntouchllc.com
[7] P. Mittendorfer and G. Cheng, "Humanoid multimodal tactile-sensing modules," *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 401–410, 2011.
[8] A. Schmitz *et al.*, "A tactile sensor for the fingertips of the humanoid robot iCub," in *Proc. IROS*, 2010.
[9] B. Bäuml *et al.*, "Agile Justin: An upgraded member of DLR's family of lightweight and torque controlled humanoids," in *Proc. IEEE International Conference on Robotics and Automation*, 2014.
[10] J. Butterfaß, M. Grebenstein, H. Liu, and G. Hirzinger, "DLR-Hand II: Next generation of a dextrous robot hand," in *Proc. IEEE International Conference on Robotics and Automation*, 2001, pp. 109–114.
[11] J. Fishel and G. Loeb, "Bayesian exploration for intelligent identification of textures," *Frontiers in Neurorobotics*, vol. 6, no. 4, pp. 1–20, 2012.
[12] D. Xu, G. Loeb, and J. Fishel, "Tactile identification of objects using Bayesian exploration," in *Proc. ICRA*, 2013.
[13] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," in *Proc. IEEE International Conference on Robotics and Automation*, 2016.
[14] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, "Multimodal feature-based surface material classification," *IEEE Transactions on Haptics*, vol. PP, no. 99, 2016.
[15] J. Sinapov, V. Sukhoy, R. Sahai, and A. Stoytchev, "Vibrotactile recognition and categorization of surfaces by a humanoid robot," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 488–497, 2011.
[16] A. G. Eguiluz, I. Rano, S. Coleman, and T. McGinnity, "Continuous material identification through tactile sensing," in *Proc. Int. Joint Conference on Neural Networks*, 2016.
[17] E. Kerr, T. M. McGinnity, and S. Coleman, "Material classification based on thermal and surface texture properties evaluated against human performance," in *13th International Conference on Control Automation Robotics Vision (ICARCV)*, Dec 2014, pp. 444–449.
[18] M. Kerzel, M. Ali, H. G. Ng, and S. Wermter, "Haptic material classification with a multi-channel neural network," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017.
[19] OptoForce. [Online]. Available: http://optoforce.com/3dsensor/
[20] Z. Erickson, S. Chernova, and C. C. Kemp, "Semi-supervised haptic material recognition for robots using generative adversarial networks," in *Proc. Conference on Robot Learning*, 2017.
[21] J. Hoelscher, J. Peters, and T. Hermans, "Evaluation of tactile feature extraction for interactive object recognition," in *Proc. IEEE-RAS International Conference on Humanoid Robots*, 2015.
[22] M. Kaboli, P. Mittendorfer, V. Hugel, and G. Cheng, "Humanoids learn object properties from robust tactile feature descriptors via multi-modal artificial skin," in *Proc.Humanoids*, 2014.
[23] M. Kaboli, A. D. L. Rosa, R. Walker, and G. Cheng, "In-hand object recognition via texture properties with robotic hands, artificial skin, and novel tactile descriptors," in *Proc. IEEE/RAS International Conference on Humanoid Robots*, 2015.
[24] D. S. Chathuranga *et al.*, "Robust real time material classification algorithm using soft three axis tactile sensor: Evaluation of the algorithm," in *Proc. IROS*, 2015.
[25] R. Li and E. H. Adelson, "Sensing and recognizing surface textures using a gelsight sensor," in *Proc. CVPR*, 2013.
[26] DLR tactile dataset. [Online]. Available: dlr-alr.github.io/dlr-tactmat
[27] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conference on Artificial Intelligence*, 1995.
[28] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," Google Research, Tech. Rep., 2015.
[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.
[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
[31] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *ArXiv e-prints*, Dec. 2014.
[32] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conference on Machine Learning*, 2016.
[33] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learning workshop*, 2015.
[34] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. CVPR*, 2006.