2019 IEEE World Haptics Conference (WHC)
Tokyo, Japan
9-12 July 2019

PerceptNet: Learning Perceptual Similarity of Haptic Textures in Presence of Unorderable Triplets

Priyadarshini K , Siddhartha Chaudhuri[2], and Subhasis Chaudhuri[3]

bstract— **In order to design haptic icons or build a haptic vocabulary, we require a set of easily distinguishable haptic signals to avoid perceptual ambiguity, which in turn requires a way to accurately estimate the perceptual dis)similarity of such signals. In this work, we present a novel method to learn such a perceptual metric based on data from human studies. Our method is based on a deep neural network that projects signals to an embedding space where the natural Euclidean distance accurately models the degree of dissimilarity between two signals. The network is trained only on non-numerical comparisons of triplets of signals, using a novel triplet loss that considers both types of triplets that are easy to order inequality constraints), as well as those that are unorderable/ambiguous equality constraints). Unlike prior MDS-based non-parametric approaches, our method can be trained on a partial set of comparisons and can embed new haptic signals without retraining the model from scratch. Extensive experimental evaluations show that our method is significantly more effective at modeling perceptual dissimilarity than alternatives.**

I. Introduction

Tactile sensory inputs can be valuable for user-facing applications requiring *symbolic or cross modal communi cation*. For instance, tactile/haptic feedback can notify us about events, provide information on sensory data that is out-of-spectrum or in an unavailable modality, and act as a communication medium for people with visual or auditory impairments [27], [7]. All such applications require a mapping between haptic signals and the domain vocabulary, e.g. different colors, or different events, or different letters and words. In order to build any such application, therefore, it is critical to have a set of haptic signals which are easily distinguishable to avoid perceptual ambiguity. In turn, this requires that we have a robust *perceptual* metric that models the human-perceived dissimilarity of different haptic signals.

Developing such a metric is difficult. *First*, humans are bad at estimating the degree of dis)similarity of different signals on a consistent scale [1]. Thus, we must rely on coarse-grained boolean comparisons that only indicate whether a person can distinguish between two signals or not, possibly with reference to a base signal. *Second*, complex nuances of human perception such as just-noticeable-difference JND) [4] require metric learning to extract useful information not just from easily distinguishable exemplars, but also from *indistinguishable* exemplars, again possibly with reference to a base signal. *Third*, gathering large amounts of haptic training data is tedious and expensive. For instance, prior non-parametric metric learning methods such as the MDS-based approach of Hollins *et al.* [11] require a complete set of comparisons of all possible pairs of signal types.

In this work, we propose a novel parametric approach for learning a perceptual metric for haptic signals. Our method utilizes a deep neural network that learns to project low-level features of input signals to an embedding space where the Euclidean distance accurately reflects the human-perceived dissimilarity between signals. The network is trained with triplets of signals: each triplet indicates a) if one signal is more similar to a base signal than another an easily-ordered or *high margin* triplet), *or* b) if no such ordering can be reliably deduced an unorderable or *low margin* triplet). Such comparisons are easily obtained via human studies [22], and are informative for metric learning *without* requiring numerical estimates of the distance between two signals [25]. Unlike prior triplet-based learning approaches which tend to ignore triplets of the second type as uninformative [21], [31], our approach treats both triplet types with, or without, a clear ordering) as informative: we show this improves the learned metric. Since our method is parametric, we can train it on as much, or as little, data as is available: it generalizes gracefully to situations where many combinations of signals lack comparisons, and applies even when some signal types are completely unseen during training.

This paper focuses on *haptic textures*: force feedback recordings of different surface materials which can be played back via a haptic renderer. We use the texture database of Strese *et al.* [28], and use a novel) spectral representation of each texture as the input to the neural network. However, our metric learning approach is generally applicable, and can be used for any collection of signals where basic input features and triplet comparisons are available.

Some prior works use neural networks for *semantic* separation, to cluster signals by class label [32], [10], [13]. Note that the underlying spirit of our work is significantly different from semantic separation tasks. In semantic separation, the relationship among input instances is binary: all signals from the same class are considered equally similar, and all signals from different classes are considered equally different. In contrast, human perception is more granular: two classes could be *more dissimilar* than two other classes. For instance, the surface textures of "Brass" and "Rubber" could be perceived as more distinct than two other textures "Brass" and "Copper". Hence, classical approaches of material classification are insufficient for finding a perceptual embedding of the signal space.

In summary, we propose the first deep metric learning

IIT Bombay, India. priyadarshini k@iitb ac in,
[2]Adobe Research and IIT Bombay. sidch@cse iitb ac in,
[3] IIT Bombay, India. sc@ee iitb ac in

978-1-5386-9461-9/19/$31.00 © 2019 IEEE

approach for modeling the perceptual similarity of haptic signals. The model is trained only on non-numerical, relative comparisons of signals, with a novel loss function accommodating both triplets that can and cannot be ordered. The parametric framework can be trained on relatively little data, and generalizes to unseen signals. Extensive experiments show that our method improves upon prior alternatives.

II. Related Work

We overview the related work on semantic and perceptual embedding of signals of different modalities:

a) Semantic Embedding of Tactile Signals: Semantic embedding has been well explored in the vision and speech domains, where objects are classified into categorical labels using machine learning methods [12], [8]. However, classification of materials using tactile sensory input has only started recently with the emergence of haptic data recording tools [13], [32], [10], [29], [17]. Gao *et al.* [10] use both visual and haptic data to classify surface materials with haptic adjectives such as hard, metallic, fuzzy etc. Liu *et al.* [18] proposed a dictionary learning model for material categorization using audio and acceleration data. Another interesting work by Aujeszky *et al.* [2] shows improvement in classification results by considering physical properties of the ambient environment along with object properties. In contrast to our approach, all these methods are class-dependent and cannot model human-perceived similarity.

b) Perceptual Embedding of Tactile Signals: Our work is inspired by Enriquez *et al.* [6], [20], who develop a set of perceptually well-separated haptic icons using MDS multi-dimensional-scaling). Other relevant MDS-based work includes [11], [23], [9]. Although this approach has proved useful for perceptual embedding, it has some limitations. *First*, it requires ground-truth comparisons of all possible pairs of signals: the quadratic scaling is poor for larger datasets. *Second*, it requires numerical estimates of pairwise distances, which are hard for individual humans to provide and can be estimated only by statistical aggregation over large user studies. *Third*, MDS does not support uncertainty in relative orderings. *Fourth*, it is a non-parametric technique which does not apply to novel signals. In contrast, our method addresses all four limitations with a parametric solution, trained only on non-numerical triplet orderings, which can be directly applied to embed novel signals. Our framework can work with partial training data and incorporates uncertainty in relative ordering into the modeling process.

c) Perceptual Embedding of Visual and Audio Signals: Several prior works in vision and speech processing proposed parametric models for perceptual metrics, using deep learning or classical kernel methods [31], [19], [16], [21]. Our work has similar goals, albeit in the haptic domain. However, none of these prior works leverages unorderable/ambiguous triplets. Pei *et al.* [24] incorporate such constraints for image clustering. Their method assumes triplet constraints are based on an underlying set of latent class labels: each triplet has two samples from one class and a third from another class. We make no such assumption, since we require a perceptual and not a semantic embedding: our triplets typically compare signals from three different classes.

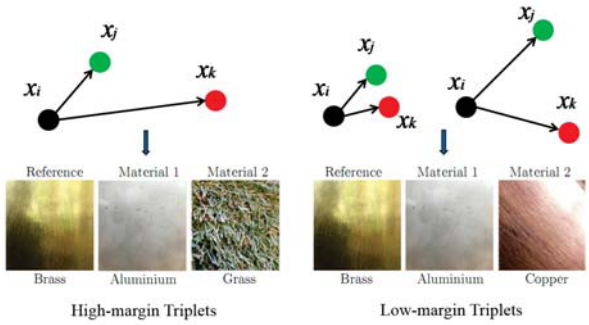


Fig. 1. Examples of high- and low-margin triplets of haptic textures [28]. The first material in each triplet is the reference. The high-margin triplet *(left)* shows brass is perceptually closer in terms of haptic feedback) to aluminum than grass; the low-margin triplet *(right)* shows brass is perceptually as similar to aluminum as is copper.

III. Method

We are given a set of haptic signals $x\}^m \in {}^n$, each described by n features, and a set of constraints $C \quad H \cup L$ encoding relative comparisons between the signals. Each constraint is a triplet of signals (x, x_j, x_k) belonging to one of two types: *high margin* H) or *low margin* L). A **high-margin triplet** indicates a case where humans are able to clearly identify signal x as being more similar to x_j than to x_k. In other words, it captures the inequality relation $d(x, x_k) - d(x, x_j) \geq \xi$, where $d(\cdot,\cdot)$ is the ground-truth perceptual distance dissimilarity) between two signals, and $\xi > 0$ is some minimum margin representing the threshold of human discrimination. Conversely, a **low-margin triplet** indicates that humans are uncertain whether x_j or x_k is more similar to x, i.e. it encodes the approximate) equality relation $|d(x, x_k) - d(x, x_j)| < \xi$. Triplet-based orderings are easy to crowdsource Section IV describes our acquisition process), more reliable than numerical similarity judgements, and effective for training metric learning models [25], [22]. The consideration of low-margin triplets in addition to high-margin triplets, accounting for human uncertainty in perceptual embedding, is a contribution of our method.

Our goal is to learn an embedding kernel $\phi: {}^n \to {}^m$, such that the Euclidean distance $d_\phi(x, y) \quad \|\phi(x) - \phi(y)\|$ satisfies the triplet constraints as well as possible for a learned margin ξ_ϕ. In other words, $\forall c \quad (x, x_j, x_k) \in C$,

$$\begin{aligned} d_\phi(x, x_k) - d_\phi(x, x_j) &\geq \xi_\phi \quad \text{if } c \in H \\ |d_\phi(x, x_k) - d_\phi(x, x_j)| &< \xi_\phi \quad \text{if } c \in L \end{aligned} \tag{1}$$

Once learned, the metric d_ϕ can be used to estimate the perceptual dissimilarity of any pair of seen or unseen signals.

A. Learning the Perceptual Distance Metric

We use a deep neural network DNN) to learn the embedding ϕ that defines our distance metric d_ϕ. DNNs have achieved great success in various high-dimensional, nonlinear regression problems arising in computer vision,

speech recognition, natural language processing, etc. While traditional statistical learning uses relatively simple models built on handcrafted input features, deep networks learn powerful task-specific features from raw, low-level input. We develop a network that maps signal features x to transformed features $\phi(x)$ that accurately capture perceptual distances.

Our model, dubbed PerceptNet, takes as input a spectral representation x of the acceleration trace of a haptic texture details below). The network consists of a series of convolutional layers, interspersed with max-pooling layers for downsampling, and ending with a fully-connected layer with *linear* activation and zero bias that outputs a 128-dimensional feature vector $\phi(x)$ architectural details in Figure 2). Hence, the network can be thought of as a *fully convolutional* portion $\psi(x)$, followed by multiplication with a matrix W the linear fully-connected layer): $\phi(x)$ $W^T\psi(x)$. The perceptual distance between signals x and y is then:

$$d_\phi(x,y) \quad \|\phi(x)-\phi(y)\| \quad \|W^T\psi(x)-W^T\psi(y)\| \quad \sqrt{(\psi(x)-\psi(y))^T WW^T(\psi(x)-\psi(y))} \quad \sqrt{(\psi(x)-\psi(y))^T M(\psi(x)-\psi(y))}, \tag{2}$$

where M is symmetric and positive semi-definite. In other words, the network effectively learns a Mahalanobis distance on the fully-convolutional embedding, which is a useful alternative visualization motivating the network design.

PerceptNet is trained with a novel loss that tries to satisfy both high- and low-margin constraints. Learning an optimal model in the presence of two kinds of triplets with opposed objectives requires carefully selecting the loss function. We must maximize the distance margin $d_\phi(x\ ,x_k)-d_\phi(x\ ,x_j)$ for a high-margin triplet $(x\ ,x_j,x_k)\in H$, while simultaneously minimizing the margin $|d_\phi(x\ ,x_k)-d_\phi(x\ ,x_j)|$ for a low-margin triplet $\in L$. Following standard formulations of the triplet loss [25], [30], [26], we modify the margins slightly to express them as differences of squared distances. This effectively weights long baseline triplets a bit more to correct gross errors in the learned manifold although the loss without squaring gave comparable results in our experiments). The overall loss is E E_H+E_L, where

$$E_H \quad \sum_{c\in H}\exp\left(-\rho(c)\right), \quad E_L \quad \sum_{c\in L}\left(1-\exp\left(-|\rho(c)|\right)\right)$$
$$\text{and } \rho\left((x\ ,x_j,x_k)\right) \quad d_\phi^2(x\ ,x_k)-d_\phi^2(x\ ,x_j) \tag{3}$$

The exponential allows training to focus on hard-to-satisfy triplets. The constant 1 in E_L is included only to visualize both losses per triplet) in $[0,1]$, and may be omitted. For each triplet, we pass its three signals through three copies of the network with shared weights) and suitably penalize the distance margin, depending on whether the triplet is high- or low-margin. The gradient of the loss, obtained by back-propagating the training error, is used to iteratively update the network weights and improve the model with the Adam optimizer [14] with a learning rate of 0.001. The model is implemented in PyTorch and trained for 1000 epochs with a batch size of 128. Given a trained model, we estimate the testing margin ξ_ϕ, which will be used to classify test triplets as high- or low-margin, by minimizing $|f_H-f_L|$, where f_H and f_L are the fractions of correctly classified high- and low-margin *training* triplets respectively.

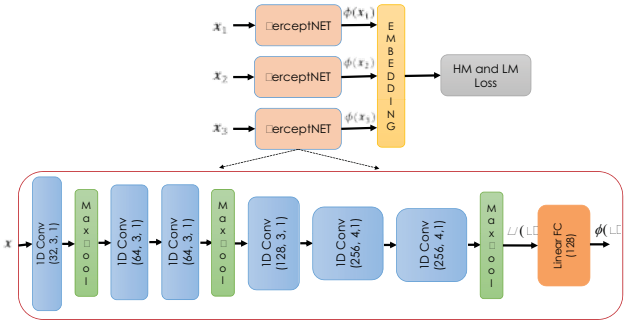


Fig. 2. PerceptNet architecture. The network has six 1D convolutional layers, three pooling layers, and a linear fully-connected layer.

B. CQFB Spectral Features

While deep networks largely eliminate the need to handcraft input features, they can still be helped by simple transformations of the input that may be hard for them to learn. In our case, we found that applying an initial spectral transform to the haptic signal, compactly summarizing periodic patterns in the data, led to better metric learning.

We use DFT321 [15] to find the spectrum magnitudes) of 3-axis acceleration data. We divide the spectrum into 32 bins, increasing geometrically in size by a factor of 1.8 [3], using a Gaussian filter of standard deviation 20 for some overlap between adjacent bins. We call this the constant Q-factor filter bank CQFB) feature vector. Note that since bins are ordered coherently by increasing frequency, convolution is a meaningful operation on this vector.

IV. Experiments

We evaluate the effectiveness of our model through several experiments. First, we study synthetic datasets to validate our algorithm in known linear and non-linear metric spaces. Then, we study a more challenging real-world haptic texture dataset [28]. The performance of each model is evaluated by the fraction of satisfied triplet constraints including, for high-margin triplets, predicting the correct ordering) in a held-out test set: the *triplet generalization accuracy* TGA).

A. Experiments on Synthetic Dataset

We first test our method with synthetic linear and nonlinear ground-truth metrics. For each metric, we draw 100 sample signals from an 8D standard normal distribution. We fix a random training threshold to distinguish between high- and low-margin triplets, according to the metric. We then randomly sample $10,000$ high-margin and $10,000$ low-margin training triplets. Each high-margin triplet is ordered appropriately. Similarly, we generate $20,000$ 10K + 10K) test triplets. The metrics we consider are:

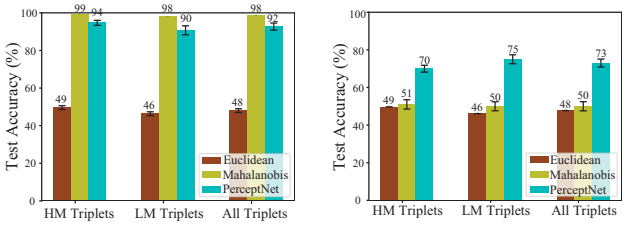


Fig. 3. Triplet generalization accuracy TGA) of PerceptNet compared to baselines, on synthetic datasets using linear Mahalanobis metric *(left)*; and non-linear elliptic Cayley-Klein metric *(right)*. Error bars represent standard deviation across 5 folds.

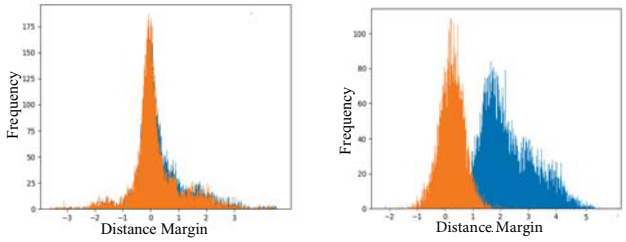


Fig. 4. Distribution of learned high-margin blue) and low-margin orange) triplet margins generated from texture data. *(Left)* In Mahalanobis space. *(Right)* In PerceptNet space.

1) A linear Mahalanobis metric defined by a random PSD matrix M, and
2) A nonlinear elliptic Cayley-Klein distance [5] defined by a random invertible symmetric matrix Ψ.

We repeat each experiment 5-fold for statistical significance, comparing the performance of PerceptNet with the natural Euclidean distance in CQFB space, and a learned Mahalanobis distance. The results are shown in Figure 3. With a ground-truth Mahalanobis metric, the learned Mahalanobis metric expectedly gives near-perfect accuracy, followed closely by PerceptNet. However, with nonlinear ground-truth, the limitations of linear modeling become clear and PerceptNet pulls significantly ahead of the other two.

B. Experiments on Haptic Texture Dataset

The TUM texture dataset [28] has acceleration signals recorded by freehand-tracing a sensing stylus over surface materials from 108 classes metals, papers, grass, etc), with 10 signals per class. The authors also measured the perceptual similarity of each pair of classes by asking 30 subjects to distinguish between signals drawn from them. Instead of directly crowdsourcing triplet comparisons in an explicit new study, we reuse the data from the TUM study to construct training and testing triplets. This allows us to efficiently sample high- and low-margin triplets, using a standard dataset that has been used in a wide range of prior work. We define the ground-truth perceptual distance $d\ (x, y)$ between two signals x, y as the fraction of subjects who could distinguish between the corresponding classes, normalized to $[0, 1]$. We also define the ground-truth perceptual margin ξ as 10% of the maximum margin over all possible triplets of signals.

Based on this dataset, we perform 3 experiments, ordered here from easiest to hardest. In each case, we repeat the experiment 5-fold, over 5 random train/test splits, and aggregate results over the folds for statistical robustness.

a) Held Out Triplets: In this experiment, we randomly sample $20,000$ triplets of signals for training, and $20,000$ different triplets for testing. We independently sample the complete collections of high- and low-margin triplets, determined by d and ξ above), to ensure each set is an even mix of $10,000$ high-margin and $10,000$ low-margin triplets for balanced training and testing. Each high-margin triplet is ordered appropriately. Any signal may appear in both training and testing triplets: thus, in this experiment, we only study how well our method generalizes to unseen comparisons, not to unseen signals. Given training and testing sets, we compare the performance of a trained PerceptNet to the natural Euclidean metric and a learned Mahalanobis metric. The accuracies of the three methods are compared in Figure 5 a) top): PerceptNet 84%) significantly outperforms the alternatives. To factor out discrepancies caused by inaccurate estimation of the testing margin ξ_ϕ for any metric, we also plot the performance of the three methods over all possible margins from 0 to 100% recall of low-margin triplets) in Figure 5 a) bottom). Again, PerceptNet is better than the alternatives over the entire range.

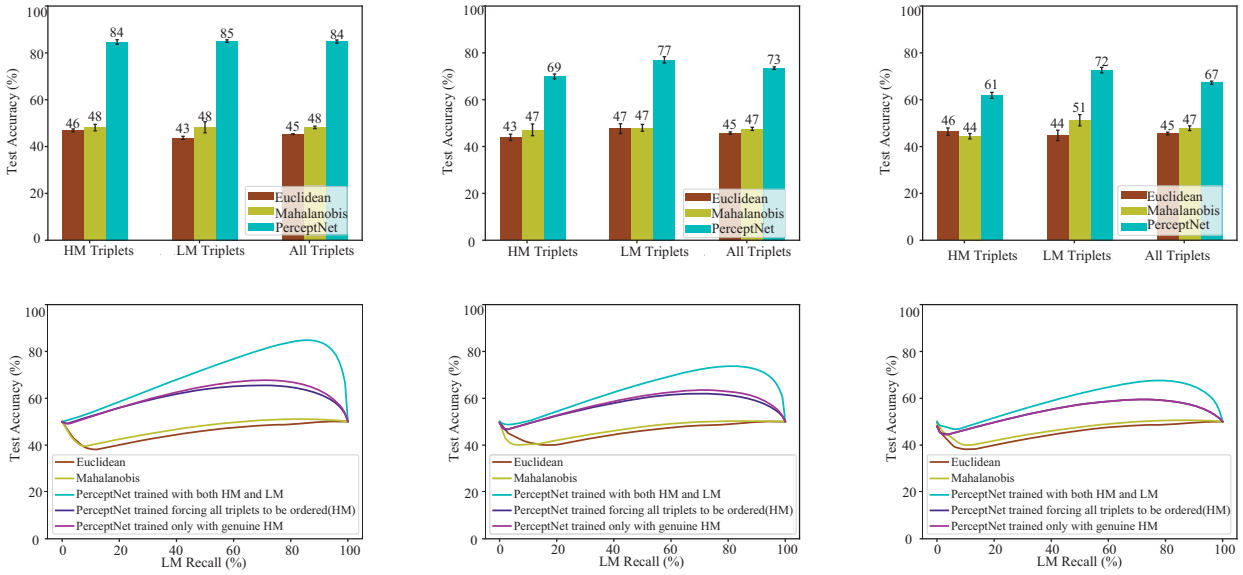
To gain insight into the embedding induced by PerceptNet, we plot a histogram of test triplet margins in Figure 4. In the learned embedding space, high- and low-margin triplets have well-separated distributions. In contrast, a learned Mahalanobis metric fails to effectively separate the two types of triplets, leading to poorer predictions.

b) Held Out Samples: Next, we randomly select 8 samples from each class for training, and hold out the remaining 2 samples for testing. We generate training and testing triplets as in the previous case, but with the additional constraint that training triplets must only use training samples, and testing triplets only use testing samples. This is a harder case since the method must generalize to unseen samples, albeit from known classes. The results are shown in Figure 5 b). The accuracy of PerceptNet 73%) reduces in this harder case, but it is still distinctly better than the baselines.

c) Held Out Classes: Finally, we hold out *all* samples from a random 20% of the 108 classes for testing. This is extremely challenging since the method must generalize to materials it has never seen before, and the training and testing distributions are quite different. As expected, performance drops further, but PerceptNet still generalizes much better 67%, Figure 5 c)).

We also observed that in all three experiments, training PerceptNet with CQFB features TGA 84, 73 and 67%) performs better than the original raw features 78, 64, 53%). This is a slightly unfair comparison since we simply widened the kernel sizes to adapt PerceptNet to the much higher-dimensional raw data. A more complex, hand-tuned network may be able to achieve comparable results with raw input also. However, CQFB enables a simpler, more efficient architecture with less manual effort.

We can make some high-level observations from the trends in these results. In all three experiments, the performance of



a) Held-out Triplets **b)** Held-out Samples **c)** Held-out Classes

Fig. 5. Triplet generalization accuracy TGA) for all three experiments of Section IV-B- a,b,c). We show both the optimal accuracy estimated using the learned threshold *(top)*, as well as the accuracy over the full range of thresholds *(bottom)*. To normalize different threshold ranges across different metric models, we map recall of low-margin triplets to the horizontal axis. In the bottom plots we also plot the accuracy of the same network architecture, trained without unorderable low-margin) triplets. Note that in the absence of low-margin training triplets, the method cannot learn an optimal HM-vs-LM threshold, and hence does not have entries in the bar graphs above: the performance gap at any given threshold can be judged from the bottom plots. Our model significantly outperforms the baselines in all three experiments.

Euclidean and Mahalanobis metrics remains ∼50%, showing that these linear models are insufficient to represent the complex nuances of perceptual similarity. Also, as problem complexity increases Figures 5 and 5 a) to c)), performance on low-margin triplets is a little better than on high-margin triplets. We believe this is so because correctly classifying a high-margin triplet requires both large separation *as well as* correct ordering of the signals, i.e. identifying which signal is nearer and which is further from the base signal. In contrast, a low-margin triplet requires only that the separation of the base signal from the other two signals be small.

In the rest of this section we present further experiments to validate important aspects of our method. Each is performed in the "held-out samples" scenario above: IV-B- b).

1) Importance of low margin triplets: A primary feature of our method is that we train with both inequality high-margin) and equality low-margin) constraints. We validate this choice by comparing it to two other models with the same network architecture. The first generates training triplets with ξ 0, forcing each triplet to be ordered i.e. high-margin) even when participants found it difficult to determine such an ordering. Normally, triplets would be treated as unorderable/equidistant i.e low-margin) to handle precisely this ambiguity. Hence, the data is noisy. The second model is trained with only high-margin triplets, completely ignoring the low-margin triplets for fairness, we pick double the usual number of HM triplets, i.e. 20,000, to match the standard training set size). Hence, it cannot leverage equality information from the latter. All models are tested on the same test set containing a mix of high- and low-margin triplets. The bottom plots of Figure 5 show that PerceptNet trained with both types of triplets more accurately models the underlying metric and outperforms the otherwise identical models trained with only high-margin triplets.

2) Pairwise distinguishability: The original perceptual similarity ground-truth [28] provides pairwise measurements, which we exploit to generate triplets. Applications like vocabulary design also use pairwise comparisons [20]. Hence, we verify if our learned metric can accurately predict distinguishable and indistinguishable signal pairs. We consider a pair distinguishable if $\geq 50\%$ subjects can tell them apart. We train PerceptNet with triplets as usual, but test on pairs of held-out signals, using a testing threshold to classify them as distinguishable or not based on their predicted separation. Figure 6 shows a precision/recall plot over the full range of thresholds. Our method is significantly more accurate AUC = 0.97) than alternatives AUC = 0.69, 0.66). We also show that the pairwise similarity matrix on test signals, where each

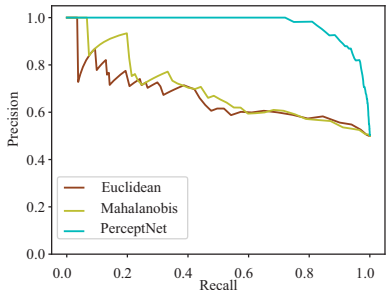


Fig. 6. Precision-recall plot for classifying distinguishable and indistinguishable pairs of signals.

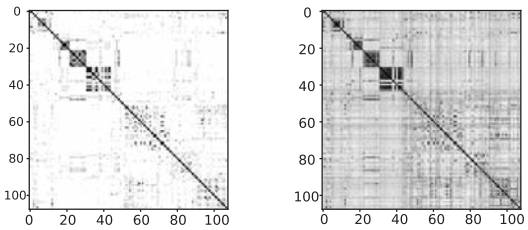


Fig. 7. Perceptual similarity matrices of 108 material classes: the ground-truth confusion matrix from [28] *(left)* and PerceptNet distances *(right)*. White indicates low and black high similarity. Since PerceptNet is trained only on non-numerical triplets and need not have a linear relationship with the numerical ground-truth confusion values, the two matrices do not have the same normalization and are not identically mapped to the greyscale range. However, the trends in the two matrices are near-identical, indicating the accuracy of the model in ranking perceptual similarity.

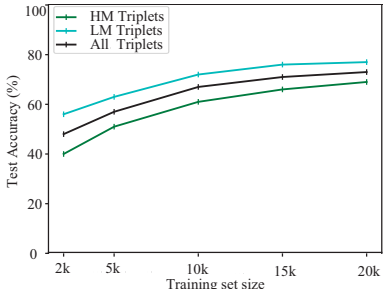


Fig. 8. Triplet generalization accuracy of PerceptNet for different training set sizes. Accuracy increases monotonically, but with decreasing benefits for larger sizes.

entry is the normalized average pairwise distance, captures very similar trends as the GT confusion matrix Figure 7).

3) Dependence on training set size: Our last experiment evaluates how our model generalizes with the number of training triplets. Given a "full" training set of 20K triplets, we randomly sample subsets of different sizes and train a model on each such subset. As Figure 8 shows, accuracy increases proportionally with the size of the training set.

V. Conclusion

In this work, we presented a deep metric learning approach for modeling the perceptual similarity of haptic signals. The model provides an end-to-end training framework for learning discriminative perceptual features from non-numerical triplet comparisons. Most significantly, we demonstrate the value of incorporating unorderable/equidistant signals into the training process, in order to better encode the nuances and limitations of human perception. Through extensive experiments, we show the our method's advantages over traditional metrics and its capacity to generalize to new data. We also find that the compact initial encoding of high-dimensional, highly-correlated, haptic acceleration traces as CQFB spectral features improves the accuracy of the learned model. In the future, we would like to model more complex nuances of human perception e.g. just-noticeable difference), generalize to wider ranges of novel signals, and incrementally train the model in an online fashion.

Acknowledgments. We thank Amit More for useful discussions which improved the paper.

References

[1] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, and S. Belongie. Generalized non-metric multidimensional scaling. In *AISTATS*, 2007.

[2] T. Aujeszky, G. Korres, and M. Eid. Material classification with laser thermography and machine learning. *QIRT*, 2018.

[3] S. Bensmaïa and M. Hollins. Pacinian representations of fine surface texture. *Perception & Psychophysics*, 67 5), 2005.

[4] A. Bhardwaj, S. Chaudhuri, and O. Dabeer. Deadzone analysis of 2D kinesthetic perception. In *Haptics Symposium*, 2014.

[5] Y. Bi, B. Fan, and F. Wu. Beyond mahalanobis metric: cayley-klein metric learning. In *CVPR*, 2015.

[6] M. Enriquez, K. MacLean, and C. Chita. Haptic phonemes: basic building blocks of haptic communication. In *ICMI*, 2006.

[7] S. Ertan, C. Lee, A. Willets, H. Tan, and A. Pentland. A wearable haptic navigation guidance system. In *ISWC*, 1998.

[8] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. *Trans. Multimedia*, 13 2), 2011.

[9] N. Gaißert, K. Ulrichs, and C. Wallraven. Visual and haptic perceptual spaces from parametrically-defined to natural objects. In *AAAI*, 2010.

[10] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell. Deep learning for tactile understanding from visual and haptic data. In *ICRA*, 2016.

[11] M. Hollins, R. Faldowski, S. Rao, and F. Young. Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis. *Perception & Psychophysics*, 54 6), 1993.

[12] P. Kamavisdar, S. Saluja, and S. Agrawal. A survey on image classification approaches and techniques. *IJARCCE*, 2 1), 2013.

[13] M. Kerzel, M. Ali, H. G. Ng, and S. Wermter. Haptic material classification with a multi-channel neural network. In *IJCNN*, 2017.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] N. Landin, J. M. Romano, W. McMahan, and K. J. Kuchenbecker. Dimensional reduction of high-frequency accelerations for haptic rendering. In *Eurohaptics*, 2010.

[16] B. Li, E. Chang, and Y. Wu. Discovery of a perceptual distance function for measuring image similarity. *Multimed. Sys.*, 8 6), 2003.

[17] H. Liu and F. Sun. Material identification using tactile perception: A semantics-regularized dictionary learning method. *Trans. Mechatron ics*, 23 3), 2018.

[18] H. Liu, F. Sun, B. Fang, and S. Lu. Multi-modal measurements fusion for surface material categorization. *Trans. Instrumentation and Measurement*, 67 2), 2018.

[19] R. Lu, K. Wu, Z. Duan, and C. Zhang. Deep ranking: Triplet MatchNet for music metric learning. In *ICASSP*, 2017.

[20] K. MacLean and M. Enriquez. Perceptual design of haptic icons. In *Eurohaptics*, 2003.

[21] B. McFee and G. Lanckriet. Learning multi-modal similarity. *JMLR*, 12, 2011.

[22] J. C. Nunnally, I. H. Bernstein, et al. *Psychometric Theory*. 1967.

[23] J. Pasquero, J. Luk, S. Little, and K. MacLean. Perceptual analysis of haptic icons: an investigation into the validity of cluster sorted mds. In *WHC*, 2006.

[24] Y. Pei, X. Z. Fern, R. Rosales, and T. V. Tjahja. Discriminative clustering with relative constraints. *arXiv preprint arXiv:1501.00037*, 2014.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[26] C. Shen, J. Kim, L. Wang, and A. van den Hengel. Positive semidefinite metric learning using boosting-like algorithms. *JMLR*, 13, 2012.

[27] C. Sjostrom. Designing haptic computer interfaces for blind people. In *ISSPA*, 2001.

[28] M. Strese, Y. Boeck, and E. Steinbach. Content-based surface material retrieval. *WHC*, 2017.

[29] M. Strese, C. Schuwerk, and E. Steinbach. Surface classification using acceleration signals recorded during human freehand movement. In *WHC*, 2015.

[30] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10, 2009.

[31] R. Zhang, P. Isola, A. A. Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[32] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Özer, and E. Steinbach. Deep learning for surface material classification using haptic and visual information. *Trans. Multimedia*, 18 12), 2016.