PerceptNet: Learning perceptual similarity of haptic textures in presence of unorderable triplets

Priyadarshini$^1$, Siddhartha Chaudhuri$^{1,2}$, and Subhasis Chaudhuri$^1$

$^1$IIT Bombay, $^2$Adobe Research

IEEE WHC 2019

July 10, 2019
**Goal** – To model perceptual dissimilarity between haptics textures

**Important aspects**

- Incorporate human notion of perceptual dissimilarity
- Model wide range of perceptual dissimilarity (highly similar to highly dissimilar)
- Embed new signals without retraining the model from scratch
Key Idea 1

Objective is to preserve human perceived relative dissimilarity between clusters
Low-margin triplets are informative in modeling perceptual dissimilarity in entirety.
Key Idea 3

Perceptual Embedding Methods

Parametric

- Allows out-of-sample extension [1, 2, 3]
- Can incorporate low-margin triplets
- Can be formulated in terms of relative as well as quantitative similarity [1, 2, 3]

Non-Parametric

- Does not work on new sample [4, 5, 6]
- Does not incorporate low-margin triplets [4, 5, 6]
- Typically formulated in terms of quantitative dissimilarity [4, 5]

1-Richard et al. CVPR2018, 2- Brian et al. JMLR2011, 3- Rui et al. ICASSP 2017, 4-Enriqz et al.ICMI 2006, 5-Sameer et al. AISTATS 2007, 6-Lauren et al. IWMLSP 2012
Perceptual Embedding of Haptic Texture
Related work

**Goal:** To design a set of well distinguishable haptic icons

**Input data:** 25 haptic stimuli

**Method:** MDS is used to select 9 most separable stimuli

**Limitations:**
- Requires users dissimilarity rating for all possible signal pairs
- Requires numerical estimates of pair-wise distance
- Non-parametric approach
- Fails to incorporate uncertainty in comparisons

Enriqz et. At ICMI - 2006
Our Method

\[ X = \{x_i\}_{1}^{m} \in R^n \]

Advantages

- Generalizes to unseen signals
- Works even with partial training data
- Requires non-numerical relative comparisons of signals
- Accommodates both types of triplets
Our Method

To learn an embedding function \( \phi: \mathbb{R}^n \rightarrow \mathbb{R}^m \) such that the Euclidean distance \( d_\phi(x, y) = \| \phi(x) - \phi(y) \| \) satisfies:

\[
\begin{align*}
    & d_\phi(x_i, x_k) - d_\phi(x_i, x_j) \geq \xi_\phi \quad \text{if } (x_i, x_j, x_k) \in H \\
    & | d_\phi(x_i, x_k) - d_\phi(x_i, x_j) | < \xi_\phi \quad \text{if } (x_i, x_j, x_k) \in L
\end{align*}
\]

We use a deep neural network (DNN) to learn \( \phi \)

\( \xi_\phi \): Hyper-parameter
Our Method

\[ \phi(x) = W^T \psi(x) \]

\[ d_{\phi}(x, y) = \| \phi(x) - \phi(y) \| = \| W^T (\psi(x) - \psi(y)) \| \]

\[ \sqrt{(\psi(x) - \psi(y))^T W W^T (\psi(x) - \psi(y))} \]

\[ \sqrt{(\psi(x) - \psi(y))^T M (\psi(x) - \psi(y))} \]
Our Method

Based on the type of triplet, distance margin is penalized by following loss function

$$\min_{\phi} \sum_{c \in H} \exp(-\rho(c)) + \sum_{c \in L} 1 - \exp(-|\rho(c)|)$$

$$\rho((x_i, x_j, x_k)) = d_{\phi}^2(x_i, x_k) - d_{\phi}^2(x_i, x_k)$$

Network is trained iteratively using standard backpropagation technique
Experiments

Input – CQFB features of acceleration signals recorded from 108 classes (metal, grass, etc) with 10 samples each and GT perceptual distance $d^*(x, y)$ of each pair of classes

Ground-truth

$d^*(x, y)$ - Fraction of subjects (out of 30) could distinguish between corresponding classes

$\xi^* - 10\%$ of the maximum margin over all possible triplets of signal

Triplets generation

$H = \{(x_i, x_j, x_k) \mid d^*(x_i, x_k) - d^*(x_i, x_j) \geq \xi^* \}$

$L = \{(x_i, x_j, x_k) \mid d^*(x_i, x_k) - d^*(x_i, x_j) < \xi^* \}$

Stresse et al. TOH 2017
Experiments

Evaluation

Triplet generalization accuracy (TGA) - Fraction of satisfied triplet constraints in a test set

\[ d_\phi(x_i, x_k) - d_\phi(x_i, x_j) \geq \xi_\phi \quad \text{if} \quad (x_i, x_j, x_k) \in H_{test} \]
\[ |d_\phi(x_i, x_k) - d_\phi(x_i, x_j)| < \xi_\phi \quad \text{if} \quad (x_i, x_j, x_k) \in L_{test} \]

\( \xi_\phi \) is estimated by minimizing \( |f_H - f_L| \) where

\( f_H \) - fraction of high-margin correctly classified training triplets
\( f_L \) - fraction of low-margin correctly classified training triplets
Experimental Results

Histogram of test triplet margins

Distribution of learned high-margin (blue) and low-margin (orange) triplet in Mahalanobis space (left) and in PerceptNet space (right)
Experimental Results

Three variants of experimental protocol:

• **Held-Out Triplets** – 50% of triplets are held-out for testing, however the samples and classes are common for training and testing

• **Held-Out Samples** - 20% samples from each class are held-out for testing

• **Held-Out Classes** - All samples from 20% class are held-out for testing
(1) Experimental Results (Held-Out Triplets)

Triplet Generalization Accuracy (TGA) of different metric at optimal threshold (left) and full range of thresholds (right).
The accuracies of PerceptNet reduces in this harder case (73%), but PerceptNet is still distinctly better.
The accuracies of PerceptNet further drops to (67%), but PerceptNet is still generalizes much better.
Experimental Results: Importance of low margin triplets

![Graph showing the comparison of test accuracy with different triplet generation methods.]

\[ H = \{(x_i, x_j, x_k) \mid d^*(x_i, x_k) - d^*(x_i, x_j) \geq \xi^*\} \]

\[ L = \{(x_i, x_j, x_k) \mid d^*(x_i, x_k) - d^*(x_i, x_j) < \xi^*\} \]
Experimental Results

Pairwise distinguishability of PerceptNet

Ground-truth generation - A pair is considered distinguishable if >50% subjects can distinguish

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

| Method    | AUC  |
|-----------|------|
| Perceptnet| 0.97 |
| Mahalanobis| 0.69 |
| Euclidean | 0.66 |
Experimental Results

Confusion matrix - White indicates low and black high similarity

PerceptNet is trained only with relative similarity, hence relative ordering is preserved not the numerical ground-truth confusion values
Experimental Results

Dependence of training set size

Accuracy increases proportionally with the size of the training set, but with decreasing benefits for larger sizes.
Perceptual Embedding of Olfactory Signals
Experiments

**Input**: Chemical features (X) and perceptual descriptor of 268 compounds (Octanol, Benzaldehyde, and Hexenel.)

**Chemical features**: hydrogen bond, molecular weight and heavy atom count, etc.

**Perceptual descriptors**: Human subjects rating against odor descriptors such as pungent, fruit, mint and smoke.

\[ d^* (x, y) \] - obtained using cosine similarity

**Triplets generation:**

\[
H = \{(x_i, x_j, x_k) \mid d^* (x_i, x_k) - d^* (x_i, x_j) \geq \xi^* \}
\]

\[
L = \{(x_i, x_j, x_k) \parallel d^* (x_i, x_k) - d^* (x_i, x_j) < \xi^* \}
\]

Kush et al. DSP 2016
Network Architecture
Unlike haptic dataset, in this case, model generalizes quite well even for compounds never seen before.
Perceptual Embedding of Image Data
Experiments

**Input:** 100 images and ground truth perceptual similarity matrix generated from crowd sourced perceptual grouping judgments

**Ground-truth:**

$d^*(x, y)$ - Fraction of subjects (out of 100) could distinguish between corresponding classes

$\xi^*$ - 10% of the maximum margin over all possible triplets of signal

**Triplets generation:**

\[
H = \{(x_i, x_j, x_k) \mid d^*(x_i, x_k) - d^*(x_i, x_j) \geq \xi^*\}
\]

\[
L = \{(x_i, x_j, x_k) \mid d^*(x_i, x_k) - d^*(x_i, x_j) < \xi^*\}
\]
Network Architecture

\[ x_1 \xrightarrow{\text{PerceptNET}} \phi(x_1) \]
\[ x_2 \xrightarrow{\text{PerceptNET}} \phi(x_2) \]
\[ x_3 \xrightarrow{\text{PerceptNET}} \phi(x_3) \]

Embedding

HM and LM Loss

\[ \psi(x) \]

\[ \phi(x) \]

Network Architecture:

1D Conv (32, 3, 1) → Max Pool
1D Conv (64, 3, 1) → Max Pool
1D Conv (128, 3, 1) → Max Pool
1D Conv (256, 4, 1)
1D Conv (256, 4, 1)
1D Conv (256, 4, 1)
Linear FC (128)
\[ \phi(x) \]
Experimental Results

- Performance of our model using different features
- Performance comparison of different metric using gist feature
Experimental Results

Pairwise distinguishability

Confusion matrix - White indicates low and black high similarity
Future Work

• Dealing with limited training data – Active learning
• Generating new sample from perceptual space by inverse mapping
• Better acquisition of data- finding trade-off between human effort and accuracy of model.