Learning Adverbs with Spectral Mixture Kernels

Tomoe Taniguchi† Daichi Mochihashi†† Ichiro Kobayashi†

† Ochanomizu University ††The Institute of Statistical Mathematics

1. Introduction

Background

- Technological advancements are making household robots that assist in daily tasks a reality
- Effective human-robot collaboration requires sharing and understanding experiences through language

Overview

Objective :
We focus on human actions **to understand the meanings of adverbs through motion features**

Dimensionality Reduction:
We use **Gaussian processes** to compress human motion data and extract frequency information

Joint Topic Model:
We propose a joint topic model which learns the relationship between human motions and adverbs to understand the meanings of adverbs related to human actions

2. Human Motion Representation

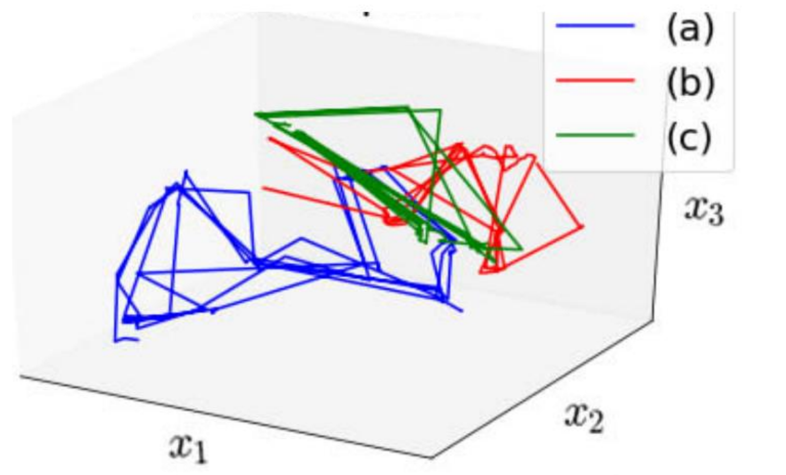

Figure 1 : Motion data compressed by GPLVM

- Human motion can be represented as smooth trajectories
- We use **Gaussian Process Latent Variable Model(GPLVM)** [Lawrence, 2003] to describe the human motions

・Three walking trajectories processed through GPLVM visualized in the three-dimensional latent space
・Cyclicity of the representations reflects the periodicity of human movements

3. Frequency components in a motion

SM kernel enables automatic learning of a mixed kernel from data by considering a combined Gaussian distribution in the Fourier domain

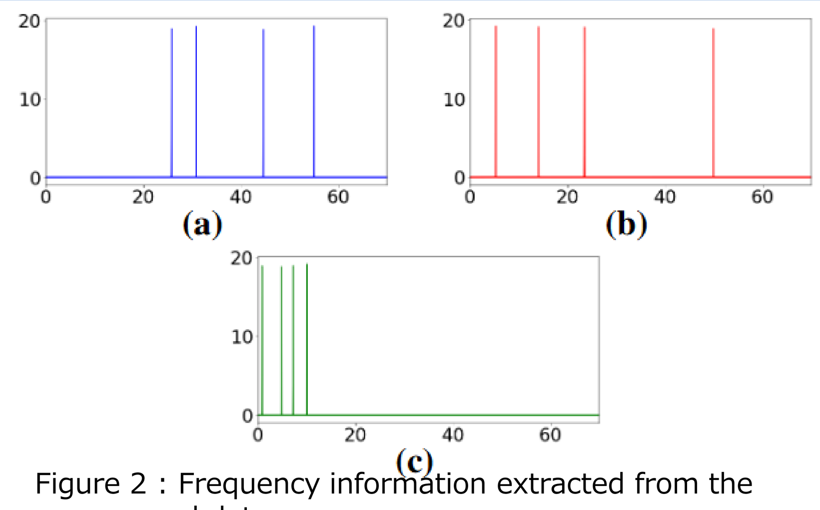

Figure 2 : Frequency information extracted from the compressed data

basis function

$$k(\tau)=\sum_{m=1}^{M} w_m \cos\left(2\pi\tau^{\mathrm{T}}\mu_m\right)\prod_{q=1}^{Q}\exp\left(-2\pi^2\tau_q^2 v_m^q\right)$$

- Human motion is cyclical
- We use **Spectral Mixture kernel (SM kernel)**[Wilson and Adams, 2013] to extract frequency components from human motions

・We analyzed the motions depicted in Figure 1 using the Spectral Mixture kernel
・The vertical and horizontal axes respectively represent the probability density and mean of the estimated four Gaussian distributions

4. HDP-Spectral Mixture LDA

Algorithm

1. Draw $G_0 \sim \mathrm{DP}(\gamma, H)$.
2. For $d = 1 \ldots D$,
   - Draw $\theta_d \sim \mathrm{DP}(\alpha, G_0)$.
3. For $n = 1 \ldots N_d$,
   - Draw $z_{dn} \sim \theta_d$
   - Draw $w_{dn} \sim \phi_{z_{dn}}$.
4. For $m = 1 \ldots M_d$,
   - Draw $y_{dm} \sim \theta_d$
   - Draw $x_{dm} \sim \mathcal{N}(\mu_{y_{dm}}, \sigma^2_{y_{dm}})$

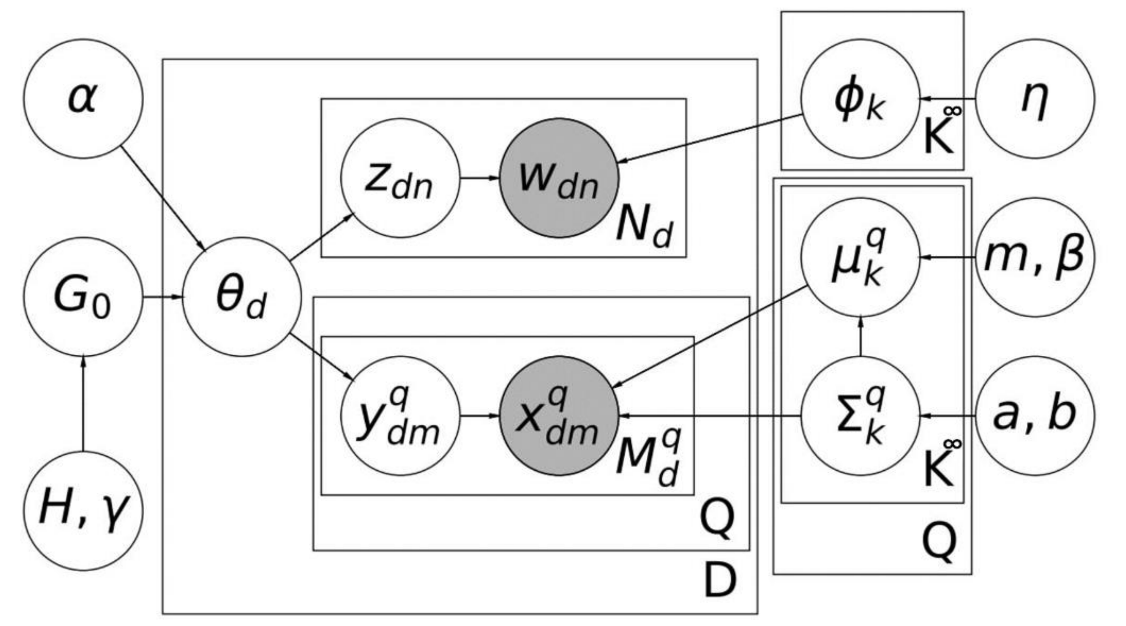

Figure 3 : Graphical model of HDP-SMLDA

- We employ collapsed Gibbs sampling [Griffiths and Steyvers, 2004] as the learning algorithm for estimating the topic distribution of adverbs and frequencies in the HDP-SMLDA
- We estimate the number of topics (K) from the data using the Chinese Restaurant Process

Here,
G_0 : base distribution
D : The number of videos
K : The number of topics
Q : Dimensionality of frequencies
N : The number of adverbs
M : The number of frequencies
Θ : Topic distribution
Z : The latent variables of adverbs
W : Adverbs
Φ : Word distribution
η : The parameter of φ
Y : The latent variables of frequencies
X : frequencies
μ : Mean of Gaussian distribution
Σ (= σ²) : Variance of Gaussian distribution

- η is iteratively updated using the Fixed-Point Iteration method

$$\eta' = \eta \cdot \frac{\sum_{k=1}^{K}\sum_{v=1}^{V}\Psi(N_{kv}+\eta) - KV\Psi(\eta)}{V\sum_{k=1}^{K}\Psi(N_k+\eta V) - KV\Psi(\eta V)}$$

- Σ is learned as a fixed value

$$\sigma^q = \frac{\max(\mathbf{X}^q) - \min(\mathbf{X}^q)}{6K^{+}}$$

- μ is sampled from the gaussian distribution (λ=1/σ²)

$$p(\mu|\mathbf{Y}) = \mathcal{N}(\mu|m, (\beta\lambda)^{-1})$$

$$\beta = M + \beta_0,\ m = \frac{1}{\beta}\left(\sum_{m=1}^{M} x_m + \beta_0 m_0\right)$$

Sampling topics of adverbs

$$p(t_{dn} = \ell|\mathbf{W}, \mathbf{T}_{\backslash dn}, \mathbf{Z}, \mathbf{Y}, \alpha, \gamma, \eta)$$

$$\propto \begin{cases} p(t_{dn} = \ell_{used})|\mathbf{W}, \mathbf{T}_{\backslash dn}, \mathbf{Z}, \mathbf{Y}, \alpha, \gamma, \eta) \\ p(t_{dn} = \ell_{new})|\mathbf{W}, \mathbf{T}_{\backslash dn}, \mathbf{Z}, \mathbf{Y}, \alpha, \gamma, \eta) \end{cases}$$

$$\propto \begin{cases} (N_{dl\backslash dn} + \sum_{q=1}^{Q} M^q_{dl})\dfrac{N_{kw_{dn}\backslash dn}+\eta}{N_{k\backslash dn}+\eta V} \\ \sum_{k=1}^{K}\frac{\alpha L_k}{L+\gamma}\dfrac{N_{kw_{dn}\backslash dn}+\eta}{N_{k\backslash dn}+\eta V} + \dfrac{\alpha\gamma}{L+\gamma}\dfrac{1}{V}. \end{cases}$$

$$p(z_{dl} = k|\mathbf{W}_{\backslash dn}, \mathbf{T}, \mathbf{Z}_{\backslash dl}, \alpha, \gamma, \beta)$$

$$\propto \begin{cases} p(z_{dl} = k_{used}|\mathbf{W}_{\backslash dn}, \mathbf{T}, \mathbf{Z}_{\backslash dl}, \alpha, \gamma, \beta) \\ p(z_{dl} = k_{new}|\mathbf{W}_{\backslash dn}, \mathbf{T}, \mathbf{Z}_{\backslash dl}, \alpha, \gamma, \beta) \end{cases}$$

$$\propto \begin{cases} L_k \cdot \dfrac{N_{kw_{dn}}+\eta}{N_{k\backslash dn}+\eta V} \\ \gamma \cdot \dfrac{1}{V} \end{cases} .$$

Sampling topics of frequencies

$$p(t_{dm} = \ell|\mathbf{W}, \mathbf{T}_{\backslash dm}, \mathbf{Z}, \mathbf{Y}, \alpha, \gamma, \eta)$$

$$\propto \begin{cases} p(t_{dm} = \ell_{used}|\mathbf{W}, \mathbf{T}_{\backslash dm}, \mathbf{Z}, \mathbf{Y}, \alpha, \gamma, \eta) \\ p(t_{dm} = \ell_{new}|\mathbf{W}, \mathbf{T}_{\backslash dm}, \mathbf{Z}, \mathbf{Y}, \alpha, \gamma, \eta) \end{cases}$$

$$\propto \begin{cases} (N_{dl} + \sum_{q=1}^{Q} M^q_{dl\backslash dm})\mathcal{N}(x|\mu_k, \sigma_k^2) \\ \sum_{k=1}^{K}\dfrac{\alpha L_k}{L+\gamma}\mathcal{N}(x|\mu_k, \sigma_k^2) + \\ \qquad \dfrac{\alpha\gamma}{L+\gamma}\mathcal{N}(x|\mu_{k_{new}}, \sigma^2_{k_{new}}), \end{cases}$$

$$p(z_{dl} = k|\mathbf{X}_{\backslash dm}, \mathbf{T}, \mathbf{Y}_{\backslash dl}, \alpha, \gamma, \beta)$$

$$\propto \begin{cases} p(z_{dl} = k_{used}|\mathbf{X}_{\backslash dm}, \mathbf{T}, \mathbf{Y}_{\backslash dl}, \alpha, \gamma, \beta) \\ p(z_{dl} = k_{new}|\mathbf{X}_{\backslash dm}, \mathbf{T}, \mathbf{Y}_{\backslash dl}, \alpha, \gamma, \beta) \end{cases}$$

$$\propto \begin{cases} L_k \cdot \mathcal{N}(x|\mu_k, \sigma_k^2) \\ \gamma \cdot \mathcal{N}(x|\mu_{k_{new}}, \sigma^2_{k_{new}}) \end{cases} .$$

5. Dataset

100 Walks (Walk data)
- Walk video in 2D format on Youtube
- Required 3D pose information for the experiment
- Divided video into 100 segments at motion breaks
- Applied four methods for 3D pose estimation

AIST++ (Dance data)
- Curated dance videos with copyright-cleared music
- Created and maintained by AIST
- Annotations in COCO format for 16 joint points [Li et al. ,2021]

Preprocessing of Videos

	Videos	Adverbs	average adverbs
Walk	100	264	12.93
Dance	1199	1767	16.18

Table 1 : Details of input data

- We requested Japanese adverb annotations for each video using the crowdsourcing site Lancers
- We utilized the direction vectors connecting each joint as input data
- To account for individual differences such as arm length, we compute unit vectors

6. Experiment

Experimental Settings

Input:
Adverb data, Frequency data

Datasets:
AIST++ 1,063 videos

Optimize Parameters:
we optimize
  Concentration parameter α
  Word distribution parameter η
  Frequency distribution parameter m,β

Epochs: 1,000

Evaluation: Perplexity of words

$$perplexity(\mathbf{w}_{\mathbf{test}}) = \exp\left(-\frac{\sum_{d=1}^{D_{test}}\sum_{n=1}^{N_d}\log(p(w_{dn}))}{\sum_{d=1}^{D_{test}} N_d}\right)$$

Frequency sampling:
Using the weights from SM Kernel estimation, sample frequencies for each epoch

Results

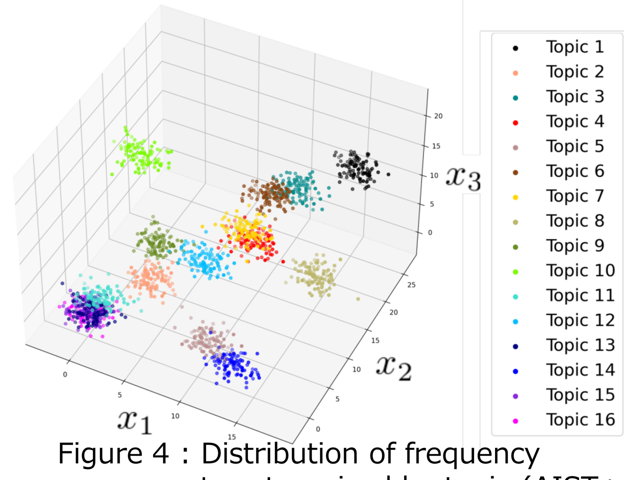

Figure 4 : Distribution of frequency components categorized by topic (AIST++)

	Unigram	LDA	HDP-SMLDA ($M_d = 4/10$)
Walk	156	99	**52** / 57
Dance	558	331	**218** / 249

Table 3 : Perplexity of topic models

- The learned Gaussian distribution means (μ) are scattered for each topic
- Topics with similar μ distances indicate similar actions
- Frequency information contributes to the classification of adverbial topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
intensely	joyfully	regularly	gracefully	powerfully	dancily	familiarly	rhythmically
powerfully	rhythmically	temporarily	elegantly	intensely	stepping	steadily	stylishly
clearly	lightly	dynamically	smoothly	intensely	joyfully	sinuously	comfortably
enthusiastically	bouncily	vividly	lightly	quickly	dynamically	briskly	smoothly
elegantly	energetically	boldly	circularly	boldly	uninterestedly	dynamically	stylishly

Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
leisurely	dynamically	springily	stylishly	dynamically	finely	carefully	lightly
smoothly	intensely	widely	stiffly	mechanically	circularly	comically	swaying
slowly	sinuously	tentatively	generously	comically	finely	carefully	wave-like
mechanically	largely	steadily	joyfully	firmly	circularly	cautiously	delicately
gently	sharply	calmly	mechanically	robotically	rhythmically	searchingly	robotically

Table 2 : Top 5 adverbs in each topic estimated by HDP-SMLDA (AIST++)

Generation of Adverbs from frequencies

- Using synonyms or adverbs labeled in other video data, appropriate adverbs can be inferred
- A Q value of 10 results in more accurate adverb predictions.

Figure 5 : A video for evaluation

Ground truth	HDP-SMLDA ($M_d = 4$)	HDP-SMLDA ($M_d = 10$)
passionately	powerfully	rhythmically
cheerfully	intensely	smoothly
rhythmically	intensely	stylishly
smoothly	boldly	flowing
flowing	confidently	cheerfully
strongly	briskly	sadly
boldly	dynamically	comfortably

Table 3 : Top 7 adverbs estimated by HDP-SMLDA

Comparative Results

- High scores, which does not necessarily indicate effective learning of adverbs
- Our model demonstrated the lowest scores on both datasets
- Our model showcases the ability to accurately estimate adverbs even with limited data

Models	Walk	Dance
Misra et al. (2017)	215	366
Nagarajan et al. (2018)	199	352
LSTM (3D/original)	210 / 402	1068 / 1794
MLP ($M_d = 4/10$)	253 / 284	994 / 1027
HDP-SMLDA ($M_d = 4/10$)	**89** / 117	**320** / 382

Table 4 : Perplexity of NN models

Experimental Settings
- Hidden layer size: 128
- Optimization function: SGD
- Loss function: Cross-entropy
- Number of epochs: 1000

7. Conclusion

- We have proposed HDP-SMLDA, which aims to comprehend the semantic nuances of sensory adverbs pertaining to human motions by learning co-occurrence relationships between motion features and adverbs.
- When compared to the other representative models, our model exhibits superior performance on classification of adverbs.