Learning Adverbs with Spectral Mixture Kernels

Anonymous ACL submission

Abstract

For humans and robots to collaborate more in the real world, robots need to understand human intentions from the different manner of their behaviors. In our study, we focus on the meaning of adverbs which describe human motions. We propose a topic model, Hierarchical Dirichlet Process-Spectral Mixture Latent Dirichlet Allocation, which concurrently learns the relationship between those human motions and those adverbs by capturing the frequency kernels that represent motion characteristics and the shared topics of adverbs that depict such motions. We trained the model on datasets we made from movies about "walking" and "dancing", and found that our model outperforms representative neural network models in terms of perplexity score. We also demonstrate our model's ability to determine the adverbs for a given motion and confirmed that the model predicts more appropriate adverbs.

1 Introduction

With technological innovations in artificial intelligence, the widespread use of household robots that collaborate with humans to assist them in their daily lives is becoming a reality. In order to collaborate with humans, it is important for robots to share and understand their experiences through language, because language is the most convenient communication tool capable of conveying human experience and knowledge. With this background, research on language use by robots in the real world has been actively studied (Taniguchi et al., 2019; Tellex et al., 2020; Kalinowska et al., 2023; Karamcheti et al., 2023). Significantly, within this domain, Large-scale language models (LLMs) such as OpenAI's ChatGPT[1] and Google's PaLM (Chowdhery et al., 2022) are also used to control robots. ChatGPT is used to execute various types of robotics tasks (Vemprala et al., 2023), and PaLM-SayCan (Ahn et al., 2022) and PALM-E (Driess et al., 2023) have been developed based on PaLM (Chowdhery et al., 2022). Singh et al. (2023) and Huang et al. (2022) have proposed methodologies for generating task plans for robots that employ LLM. Their approach conveys robot's motion plans through a chain-of-thought framework (Wei et al., 2022). Though it is good at describing in language the general plan of action of a robot in accomplishing a specific task, the language description does not capture the precise correspondence between nuanced expressions and the actual robot behaviors in the real world. Furthermore, the focus of their studies is not on the verbal representation of the behaviors of the observed object by a robot, but on the robot's action plan. On the other hand, research is being conducted to elucidate the relationship between motions and the natural language that describes them. Bidirectional conversion models from natural language descriptions to motions, or vice versa, using sequence-to-sequence (Seq2seq) (Sutskever et al., 2014) learning have been proposed by Yamada et al. (2018); Plappert et al. (2018); Ito et al. (2022). Though these models can achieve bidirectional conversion between language and motion sequences, the relation between motions and language is learned as sequence patterns and lacks in learning the correspondence between the manner of motions and the language that represent them. Furthermore, in the conventional research the focus has predominantly revolved around finite motions, such as "take" and "put", which were preconceived by humans, thereby neglecting the pursuit of methodologies that facilitate the adaptable modulation of multiple motions contingent upon contextual cues. For the advancement of robotics, it becomes imperative to comprehensively and statistically grasp the repertoire of "motions" that humans genuinely exhibit, as well as discern the variations in individual characteristics and contextual nuances associated with those "motions". These insights should be

[1] https://chat.openai.com.

aptly assimilated within the robotic systems. Building upon the aforementioned, we shall address this challenge by casting our focus on adverbs, while establishing correspondence between motions and adverbs that represent them.

Limited research has been conducted thus far to delve into the semantic comprehension of adverbs. Notable instances within this domain include the Three-Stream Hybrid Model (Pang et al., 2018), which employs Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and inceptionV3 (Szegedy et al., 2016) to acquire knowledge related to adverbs. Additionally, Action Modifiers (Doughty et al., 2020), which employ an I3D network (Carreira and Zisserman, 2017) and scaled dot-product attention (Vaswani et al., 2017) to discern the impact of adverbs on motion sequences. These models employ image features derived from videos, such as RGB and optical flow (Simonyan and Zisserman, 2014), as representations of motions. However, these representations fail to capture the intrinsic essence of the motions themselves; these models models are capable of classify videos annotated adverbs by learning RGB or optical flow, but they are unable to discern the component of motions denoted by the adverb. Therefore, unlike conventional research approaches, in this study, we focus on the frequency components that make up human motion and attempt to express the motion by those components. By doing so, we aim to enable the robot to understand the meaning of adverbs related to motions such as "cut *roughly*", etc.

2 Joint Topic Model of Motions and Adverbs

We propose a new topic model, Hierarchical Dirichlet Process-Spectral Mixture Latent Dirichlet Allocation (HDP-SMLDA) to capture the relationship between the frequency components of human motions and the adverbs that describe motions .

The model makes it statistically possible to establish a correspondence between adverbs and nuances associated with motions.

This enables the control of robot actions through verbal instructions, such as "handle *with more caution*" or "cut *roughly*", and it is also possible to make the robot understand human intentions due to slightly different manner of movement. On the contrary, from the perspective of natural language processing, it has been impossible to express the actual meaning behind words like "*freely*" or "*flexibly*". However, the integration with robotics makes it possible for the first time to represent their meaning, allowing not only the description of actions through language but also the generation of actions from language cues.

2.1 Human Motion Representation

Since human motion is represented as a smooth trajectory, we use a Gaussian process (GP) (Rasmussen and Williams, 2006), which is defined as a distribution over functions, to describe the motions. In a GP, the kernel function $k(x, x')$, which determines the similarity between two data points (x, x'), is applied to the data set to compute the covariance matrix and estimate the predictive distribution. The choice of kernel function is an important factor that affects the behavior and performance of the GP model. GP models are primarily used for regression and classification, fundamental techniques that are also widely used by the natural language processing community (Cohn et al., 2014).

2.2 Frequency components in a motion

Wilson et al. (Wilson and Adams, 2013) introduced a technique known as the Spectral Mixture kernel (SM kernel), which enables automatic learning of a mixed kernel from data by considering a combined Gaussian distribution in the Fourier domain. This approach surpasses the limitation of utilizing pre-existing bases or their combinations in Gaussian processes. As a fundamental component of the Gaussian process, we consider a radial basis function $k(\tau)$ that solely depends on $\tau = x - x'$. According to Bochner's theorem (Bochner et al., 1959; Stein, 1999), any $k(\tau)$ can be expressed in the following equation:

$$k(x, x') = k(\tau) = \int_{\mathbb{R}} e^{2\pi i s^{\mathrm{T}} \tau} \psi ds. \qquad (1)$$

As $k(\tau)$ is considered equivalent to probability density $\psi(s)$ in the frequency domain, we consider a mixture of Gaussian distributions for $\psi(s)$. Each component of the Gaussian distributions is equivalent to considering the following basis function in the original domain:

$$k(\tau|\sigma, \mu) = \exp(-2\pi^2\tau^2 v^2) \cos(2\pi\tau\mu). \quad (2)$$

Thus, we are considering a mixture of M basis functions as the basis. Here, μ_m^q and v_m^q represent the mean and variance, respectively, of the q-th dimension of the input $\boldsymbol{X}$ in the m-th basis:

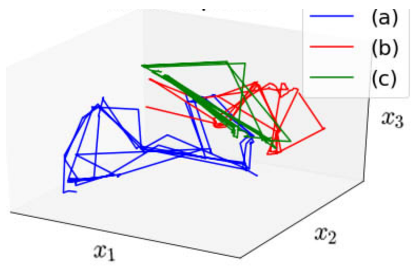


Figure 1: Nonlinear dimensionality reduction of motions achieved through GPLVM. The trajectories corresponding to three distinct walking motions (a)-(c) are portrayed in the latent space of three dimensions (thus, we set $Q = 3$ in Equation 3), denoted as $\boldsymbol{X}$.

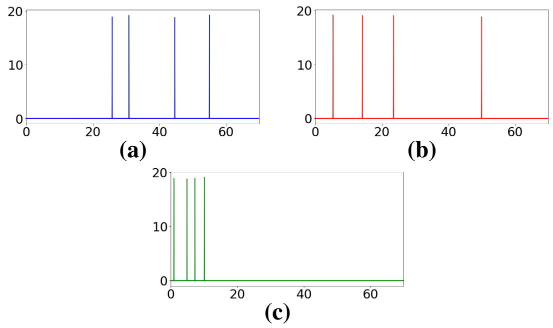


Figure 2: The motions depicted in Figure 1 were analyzed using the SM kernel. The vertical and horizontal axes respectively represent the probability density and mean of the estimated four Gaussian distributions (thus, we set $M = 4$ in Equation 3).

$$k(\tau) = \sum_{m=1}^{M} w_m \cos\left(2\pi\tau^{\mathrm{T}}\mu_m\right) \prod_{q=1}^{Q} \exp(-2\pi^2\tau_q^2 v_m^q). \tag{3}$$

The weights parameter $\mathbf{w}$, mean $\boldsymbol{\mu}$, and variance $\boldsymbol{v}$ can be learned through hyperparameter optimization of Gaussian processes. We employ this method to extract M frequency components (represented by the mean $\boldsymbol{\mu}$) that are expected to be relevant to adverbs from the three-dimensional latent variable $\boldsymbol{X}$ obtained through GPLVM for each motion. These components are then used as observed values that capture the characteristics of the motions. It is worth noting that while the trajectory in $\boldsymbol{X}$ can be directly Fourier transformed, doing so would not allow us to distinguish between the function passing through particular points (the *phase* of the function) and the *features* of the function itself.

2.3 Hierarchical Dirichlet Process-Spectral Mixture LDA

The extracted frequency components from the motions are assumed to be associated with the adverbs assigned to those motions. By employing Gaussian-Multinomial LDA (GM-LDA) (Blei and Jordan, 2003), we can cluster the frequency components and adverbs simultaneously into topics, thereby identifying frequency components that are likely to co-occur with a given adverb. It is important to note that GM-LDA requires the number of topics K to be known in advance. However, the number of topics is typically unknown, and assuming prior knowledge of this parameter is a significant limitation. To address this issue, we propose the Hierarchical Dirichlet Process Spectral Mixture LDA (HDP-SMLDA), which automatically estimates the number of topics from the data by incorporating a hierarchical Dirichlet process into GM-LDA. The graphical model, as depicted in Figure 3, considers Q as the number of dimensions of the frequency components. In our study, we set $Q = 3$ because the data processed by GPLVM is three-dimensional. The number of kernel mixtures M in the Spectral Mixture (SM) kernel discussed in the previous section is denoted as M_d in this model. Adverbs are sampled from a categorical distribution, while the frequency component is treated as continuous data, assuming a Gaussian distribution as the prior distribution. Let us assume the existence of a potential topic distribution θ_d for each motion d. The dimensionality of the topics, denoted as K, is variable, allowing for flexibility. The generation process of the adverb w_{dn} $(n = 1, \ldots, N_d)$ and the frequency component x_{dm} $(d = 1, \ldots, D;\ m = 1, \ldots, M_d)$ associated with the motions is outlined as follows:

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 $x_{dm} \sim \mathcal{N}(\mu_{y_{dm}}, \sigma^2_{y_{dm}})$.

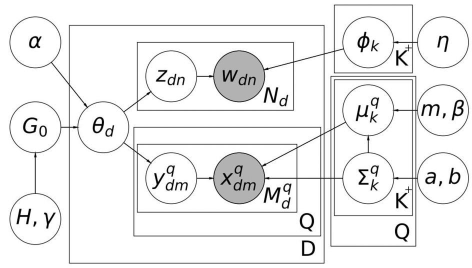


Figure 3: The graphical model of HDP-SMLDA. K^+ represents the variable number of topics. At each iteration of training, the hyperparameter α is estimated based on the size of the dataset, ensuring flexibility in the model.

In the generative process, ϕ_k represents the categorical distribution of the adverb corresponding to the k-th topic, while $\mathcal{N}(\mu_k, \sigma_k^2)$ denotes the Gaussian distribution of the frequency component associated with the same topic. The topic distribution θ is calculated based on the information from both the adverbs and frequency components. This topic distribution is then utilized to assign topics to each adverb and frequency component iteratively for each motion d.

Sampling Topics of Adverbs and Frequencies

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.

Sampling topics of adverbs Let T represents the set of table assignments and ℓ denotes the table number. According to the Chinese restaurant process (Teh et al., 2006), the topic z_{dn} assigned to the adverb w_{dn} is determined by sampling the occupied table T_{dn} using the following formula. Here, ℓ_{used} and ℓ_{new} correspond to existing and new tables, L_k and L represent the number of tables assigned to topic k and the total number of tables, respectively, and V signifies the number of vocabularies:

$$
\begin{aligned}
&p(t_{dn} = \ell | \boldsymbol{W}, \boldsymbol{T}_{\backslash \boldsymbol{dn}}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta) \\
&\propto \begin{cases} p(t_{dn} = \ell_{used}) | \boldsymbol{W}, \boldsymbol{T}_{\backslash \boldsymbol{dn}}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta) \\ p(t_{dn} = \ell_{new}) | \boldsymbol{W}, \boldsymbol{T}_{\backslash \boldsymbol{dn}}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta) \end{cases} \\
&\propto \begin{cases} (N_{dl \backslash dn} + \sum_{q=1}^{Q} M_{dl}^{q}) \dfrac{N_{kw_{dn} \backslash dn} + \eta}{N_{k \backslash dn} + \eta V} \\ \sum_{k=1}^{K} \frac{\alpha L_k}{L+\gamma} \dfrac{N_{kw_{dn} \backslash dn} + \eta}{N_{k \backslash dn} + \eta V} + \dfrac{\alpha\gamma}{L+\gamma} \dfrac{1}{V} . \end{cases}
\end{aligned} \tag{4}
$$

The following formula is employed to sample the topics assigned to the new table. Here, k_{used} refers to existing topics, while k_{new} represents new topics:

$$
\begin{aligned}
&p(z_{dl} = k | \mathbf{W}_{\backslash dn}, \mathbf{T}, \mathbf{Z}_{\backslash dl}, \alpha, \gamma, \beta) \\
&\propto \begin{cases} p(z_{dl} = k_{used} | \mathbf{W}_{\backslash dn}, \mathbf{T}, \mathbf{Z}_{\backslash dl}, \alpha, \gamma, \beta) \\ p(z_{dl} = k_{new} | \mathbf{W}_{\backslash dn}, \mathbf{T}, \mathbf{Z}_{\backslash dl}, \alpha, \gamma, \beta) \end{cases} \\
&\propto \begin{cases} L_k \dfrac{N_{kw_{dn}} + \eta}{N_{k \backslash dn} + \eta V} \\ \gamma \dfrac{1}{V} \end{cases} .
\end{aligned} \tag{5}
$$

The hyperparameter η is iteratively updated using the Fixed-Point Iteration method(Minka, 2003) based on the following equation:

$$
\eta' = \eta \frac{\sum_{k=1}^{K} \sum_{v=1}^{V} \Psi(N_{kv} + \eta) - KV\Psi(\eta)}{V \sum_{k=1}^{K} \Psi(N_k + \eta V) - KV\Psi(\eta V)} . \tag{6}
$$

Sampling topics of frequencies The topic y_{dm} assigned to the frequency component x_{dm} is sampled using the following equation:

$$
\begin{aligned}
&p(t_{dm} = \ell | \boldsymbol{W}, \boldsymbol{T}_{\backslash dm}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta) \\
&\propto \begin{cases} p(t_{dm} = \ell_{used} | \boldsymbol{W}, \boldsymbol{T}_{\backslash dm}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta) \\ p(t_{dm} = \ell_{new} | \boldsymbol{W}, \boldsymbol{T}_{\backslash dm}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta) \end{cases} \\
&\propto \begin{cases} (N_{dl} + \sum_{q=1}^{Q} M_{dl \backslash dm}^{q}) f(x | \mu_k, \sigma_k^2) \\ \sum_{k=1}^{K} \dfrac{\alpha L_k}{L+\gamma} f(x | \mu_k, \sigma_k^2) \\ \qquad + \dfrac{\alpha\gamma}{L+\gamma} f(x | \mu_{k_{new}}, \sigma_{k_{new}}^2), \end{cases}
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
&p(z_{dl} = k | \mathbf{X}_{\backslash dm}, \mathbf{T}, \mathbf{Y}_{\backslash dl}, \alpha, \gamma, \beta) \\
&\propto \begin{cases} p(z_{dl} = k_{used} | \mathbf{X}_{\backslash dm}, \mathbf{T}, \mathbf{Y}_{\backslash dl}, \alpha, \gamma, \beta) \\ p(z_{dl} = k_{new} | \mathbf{X}_{\backslash dm}, \mathbf{T}, \mathbf{Y}_{\backslash dl}, \alpha, \gamma, \beta) \end{cases} \\
&\propto \begin{cases} L_k f(x | \mu_k, \sigma_k^2) \\ \gamma f(x | \mu_{k_{new}}, \sigma_{k_{new}}^2) . \end{cases}
\end{aligned} \tag{8}
$$

The variance parameter σ^2 of the Gaussian distribution is learned as a fixed value. To ensure that the Gaussian distribution is evenly distributed over the data range, we calculate σ using the following equation. This is done because the data typically fall within the range of approximately -3σ to 3σ when the mean is set to 0. Here, K^+ represents the number of topics at the current iteration:

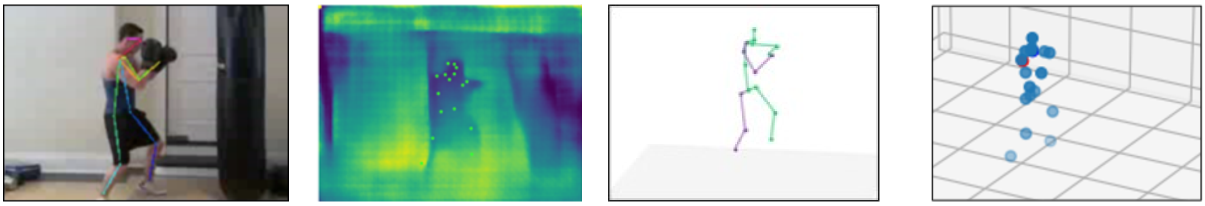
$$
\sigma^q = \frac{\max(\mathbf{X}^q) - \min(\mathbf{X}^q)}{6K^+} . \tag{9}
$$

The mean parameter μ of the Gaussian distribution is sampled from the posterior distribution given by the following equation. Here, λ is defined as $\lambda = 1/\sigma^2$, where σ^2 represents the variance of the Gaussian distribution:

$$
p(\mu | \mathbf{Y}) = \mathrm{N}(\mu | m, (\beta\lambda)^{-1}) . \tag{10}
$$

Let us assume that β_0 and m_0 are the parameters of the prior distribution, and they are defined as follows:

$$
\beta = M + \beta_0, \ m = \frac{1}{\beta} \left(\sum_{m=1}^{M} x_m + \beta_0 m_0 \right) . \tag{11}
$$



(a) 2D pose estimation (Cao et al., 2021) **(b)** Depth estimation (Laina et al., 2016) **(c)** 3D pose estimation (Martinez et al., 2017) **(d)** Direction normalization (See text)

Figure 4: Through the four sequential procedural stages, three-dimensional human joint points data are extracted from a two-dimensional video.

To estimate the mean $\mu_{k_{new}}$ for the Gaussian distribution associated with the new topic directly is not possible since there is no data belonging to the cluster. To address this, the mean is sampled from a Gaussian distribution using suitable parameters, allowing it to be learned to some extent, and then estimated as same as the mean of existing topic.

Estimation of scaling parameter α

To better estimate the number of topics that best fit the data, we adopt a gamma distribution as the prior distribution for the scaling parameter α:

$$\begin{aligned} &p(\alpha|\pi, s, Z, c_1, c_2) \\ &\quad = Ga(\alpha|c_1 + K^+ - s, c_2 - \log\pi). \end{aligned} \tag{12}$$

π and s are sampled as follows:

$$\begin{aligned} &p(\pi|\alpha, s, Z, c_1, c_2) \\ &\quad = Beta(\pi|\alpha + 1, N + M), \end{aligned} \tag{13}$$

$$\begin{aligned} &p(s|\alpha, \pi, Z, c_1, c_2) \\ &\quad = Bernoulli\left(s \,\middle|\, \frac{N+M}{N+M+\alpha}\right). \end{aligned} \tag{14}$$

3 Experiments

We begin by providing a description of the datasets utilized in our experiments. We then proceed to conduct an experiment involving HDP-SMLDA, where we examine the adverbs and frequency components, and generate adverbs based on the frequency components within the trained model.

3.1 Experimental settings

Data set

We conducted an experiment utilizing a dataset containing walking motions called 100 Walks[2]and another dataset comprising dancing motions called AIST++[3].

100 Walks 100 Walks, the video available on YouTube, is in a two-dimensional format. However, for our experiment, we required three-dimensional pose information as input data. To overcome this limitation, we divided the video into 100 segments at the motion breaks and applied four different methods for three-dimensional pose estimation.

1. Estimate 2D skeletal coordinates from video data using Openpose (Cao et al., 2021) (Figure 4(a))
2. Estimate the depth of the video per frame using FCRN-depth prediction (Laina et al., 2016) (Figure4(b))
3. Estimate 3D skeletal coordinates from video data using results of 1 and 2 ,and 3d-pose baseline (Martinez et al., 2017) (Figure 4(c))
4. Normalize human body orientation using a rotation matrix (Figure 4(d))

AIST++ The AIST Dance DB (Tsuchida et al., 2019) is a curated dataset consisting of original dance videos. These videos have been carefully selected and include dance performances accompanied by copyright-cleared music. The dataset is created and maintained by the National Institute of Advanced Industrial Science and Technology (AIST). Li et al. (2021) conducted annotations on the AIST Dance DB dataset, specifically focusing on three-dimensional human keypoints and developed a dance generation model. These annotations provide valuable information for each dance video in the dataset. Additionally, they released the annotated dataset called AIST++, which consists of 1,199 simple Basic Dance motions annotated with three-dimensional pose information for 16 joint

[2]https://www.youtube.com/watch?v=HEoUhlesN9E

[3]https://google.github.io/aistplusplus_dataset/

points in the COCO format. The dataset consists of 10 different choreographies, each representing a specific genre of dance. For each choreography, there are 20 different dancers who perform the dance in the corresponding video. The dancers follow the specified choreography while dancing to genre-specific music. The music tempo varies across the dataset and is set at six different levels.

Annotation of adverbs

We employed a crowdsourcing system called Lancers[4] to gather annotations from multiple annotators for the Japanese adverbs associated with the human motions in the videos. We requested each annotator to provide as many Japanese adverbs as possible for human motions of each video. To ensure the quality of the annotations, we considered only those adverbs that appeared at least three times across all the videos and discarded the rest as noise. For the 100 Walks dataset, we assigned 20 annotators to annotate every 100 videos. In the case of the AIST++ dataset, we assigned 5 annotators to annotate every 50 videos. This approach allowed us to collect a diverse range of adverbs associated with the motions while maintaining the quality of the annotations. The details of the adverb dataset are presented in Table 1, where the 100 Walks dataset is referred to as "walk" and the AIST++ dataset is referred to as "dance". The metric "average adverbs" represents the mean number of adverbs annotated per video. In comparison to data set used in prior research (Pang et al., 2018; Malmaud et al., 2015), we have amassed a more extensive corpus of adverbs in both datasets.

Calculation of direction vectors

We utilize the direction vectors connecting each joint as input data to reconstruct the original pose information. To account for individual differences such as arm length, we compute unit vectors. For the 100 Walks dataset, we compute 16 direction vectors, while for the AIST++ dataset, we compute 14 direction vectors. The resulting vectors are then combined, with their three-dimensional coordinates arranged in the column direction for each frame. Consequently, the data dimensions are 48 and 42 for the respective datasets.

[4] https://www.lancers.jp/

	Videos	Adverbs	average adverbs
walk	100	264	12.93
dance	1199	1767	16.18

Table 1: Details of the data.

Extraction of frequency components from human motions

Frequency components were extracted from the preprocessed video data utilizing the following two steps. Experiments were conducted by varying the number of kernel mixtures, denoted as M_d, within the range of 4 to 12.

1. Reduce high-dimensional pose data to low-dimensional latent variables using GPLVM. Figure 1 shows the case of reducing pose data into three-dimensional latent variables.
2. Extract frequency components for each dimension from the three-dimensional latent variables using SM kernel. Figure 2 shows the case of using four bases of Gaussian distribution.

Three motions from the training data of the 100 Walks dataset, processed through Gaussian Process Latent Variable Model (GPLVM), are visualized in the three-dimensional latent space, as depicted in Figure 1. In our approach, we employ the radial basis function (RBF) as the kernel function of GPLVM. To optimize the values of $\boldsymbol{X}$ and the hyperparameters of the kernel, we utilize the L-BFGS method (Liu and Nocedal, 1989). Due to the repetitive nature of walking motions, the latent variables exhibit circular patterns, as observed in the figure. For $M_d = 4$, the Gaussian distribution is depicted in Figure 2 with optimized mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ parameters for the first dimension of each motion, using the SM kernel. The estimated variance is exceptionally small, resulting in the Gaussian distribution being represented as a delta function in the figure. From Equation (3), we observe that a larger mean μ value corresponds to a shorter period. Therefore, it can be inferred that the spectral components representing the basis are more likely to be found on the left side of the spectrum for motion data with slower fluctuations. Thus, (a) contains more fast motion components, (c) contains more slow motion components, and (b) lies in between as an intermediate case. The SM kernel is optimized with weights as parameters, representing the significance of each frequency component. At each iteration, the frequency components used as motion features in each video are sampled using the weights.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
wildly	happily	regularly	gracefully	strongly	dancing	practiced	rhythmically
strongly	rhythmically	rhythmically	smoothly	wildly	stepping	settled	stylishly
clearly	lightly	dynamically	seemly	confidently	happily	waving	comfortable
passionately	bouncily	cheerfully	lightly	quickly	dynamically	quickly	flowing
classy	cheerfully	boldly	spinning	boldly	disappointed	dynamically	cool
Topic 9	**Topic 10**	**Topic 11**	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**
spacisouly	dynamically	bouncily	cool	sharply	finely	checking	lightly
smoothly	wildly	spreading	sharply	machinelike	spinning	comically	shaking
slowly	waving	totteringly	spacisouly	comically	suffering	carefully	waving
machinelike	big	steadily	happily	firmly	avoiding	cautiously	finely
quietly	sharply	settled	machinelike	strangely	rhythmically	seemly	robotlike

Table 2: AIST++ dataset($M_d = 4$): Top 5 adverbs in each topic estimated by HDP-SMLDA. Each topic corresponds to each topic in Figure 5. Compared to LDA, HDP-SMLDA takes into account not only co-occurrence of adverbs but also similarity of motions when classifying adverbs.

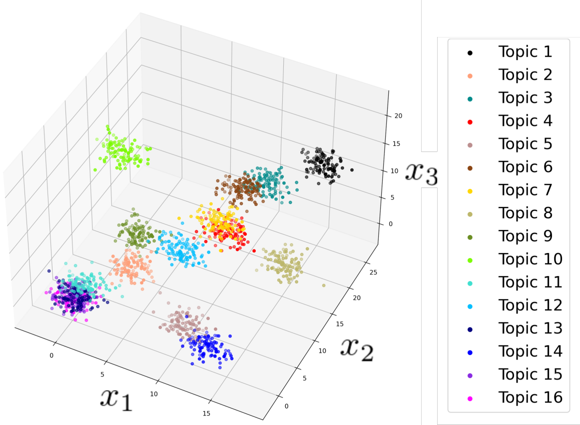


Figure 5: The relationship between topics and motion features can be visualized by plotting 100 samples extracted from the Gaussian distribution associated with each topic learned through HDP-SMLDA.

3.2 Result

For the AIST++ dataset with $M_d = 4$, Table 2 displays the top five words for each adverb, along with their corresponding Normalized Pointwise Mutual Information (NPMI) values (Bouma, 2009) calculated from the learned topic-word distribution. Figure 5 visually represents the 100 samples in a three-dimensional space, obtained from the Gaussian distribution associated with the mean μ_k of each learned topic. Each sample represents a frequency component that symbolizes a specific topic, and the proximity of the samples indicates similarity in their frequency components. It is important to note that since the scales are not estimated, the dispersion of the points in the figure remains constant. To evaluate the performance of this model, perplexity is used as a metric. Table 3 presents the perplexity of each topic model during training. Additionally, the perplexity for the Unigram model is calculated using the word distribution prior to training.

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

Table 3: Perplexity at training in each topic model.

Generation of adverbs from frequency To verify the accurate association between frequencies and adverbs, we performed an experiment where we generated adverbs based on the frequency components extracted from an evaluation video (Figure 6), utilizing the learned word distribution. Table 4 presents both the ground truth adverbs and the top seven adverbs with the highest probabilities, calculated through HDP-SMLDA. Through the estimation of M_d from 4 to 12, we observed that, for the majority of evaluation videos, the estimation with $M_d = 10$ yielded more suitable adverbs as the top choices.

3.3 Discussions

In Figure 5, the arrangement of the 16 Gaussian distributions evenly spans the width of the data. Notably, Topic 5 and Topic 14 exhibit proximity to each other, indicating a similarity in the content of the motions, as supported by Table 4 showcasing the top adverbs associated with each topic. Topics 1, 8, and 10 appear more distanced from the other topics. Notably, these three topics demonstrate pronounced adverb features in terms of frequency. While there may be an apparent overlap between the content of Topics 1 and 10, a closer examination of the top 20 words reveals that Topic 1 encompasses emotionally driven dances such as “bravely” and “heavily”, while Topic 10 represents adverbs associated with more vigorous movements

Figure 6: A video for evaluation. In the video, the dancer is dancing jazz ballet.

Ground truth	HDP-SMLDA ($M_d = 4$)	HDP-SMLDA ($M_d = 10$)
passionately	strongly	rhythmically
cheerfully	wildly	smoothly
rhythmically	clearly	stylishly
smoothly	boldly	flowing
flowing	confidently	cheerfully
strongly	sharply	sadly
boldly	dynamic	happily

Table 4: Ground truth adverbs of the dance video (Figure 6) and Top 7 adverbs estimated by HDP-SMLDA.

like "sharply" and "refreshed". This distinction suggests that the model successfully clusters adverbs based on both semantic and motion-related features derived from frequency components. The perplexity values from Table 3 indicate significantly lower values compared to those obtained from LDA training data, signifying the valuable contribution of frequency components in adverb topic classification. Although increasing the number of mixtures in the kernel was expected to reduce perplexity, the experiment yielded unfavorable results. On the other hand,regarding the generation of adverbs from frequency components, it was observed that when $M_d = 10$, the model was able to estimate more suitable adverbs compared to when $M_d = 4$. This observation raises the possibility that the annotators may have encountered difficulty in identifying the precise vocabulary during the annotation process or that the model could generate correct synonyms that did not align perfectly with the ground truth.

Comparison with neural network models We conducted additional experiments to compare the representative neural network model's performance. Given that our study involves annotations of multiple adverbs per video, multi-label learning becomes necessary. In typical class classification learning, the model calculates the error by back-propagating the difference between the output probability and the input label. However, in our case, training is performed by back-propagating the average of errors for all adverb labels annotated to the video. We conducted experiments using two different models, Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP)(Rumelhart and McClelland, 1987), with four different data inputs:

1. Input data processed by GPLVM to LSTM
2. Input original data to LSTM
3. Input frequency ($M_d = 4$) to MLP
4. Input frequency ($M_d = 10$) to MLP

	LSTM (3D/Original)	MLP ($M_d = 4/10$)	HDP-SMLDA ($M_d = 4/10$)
walk	210 / 402	253 / 284	**89** / 117
dance	1068 / 1794	994 / 1027	**320** / 382

Table 5: Perplexity at evaluating in each model.

Table 5 displays the perplexity scores for each model during evaluation. Comparing the data processed by GPLVM with the original data, it is evident that the processed data yielded lower perplexity, indicating the effectiveness of data dimensionality reduction in class classification. All neural network models received high scores, which does not necessarily indicate effective learning of adverbs. Nonetheless, our proposed method demonstrated the highest scores on both datasets, highlighting its superior performance. Thus,our model showcases the ability to accurately estimate adverbs even with limited data.

4 Conclusions

We have proposed a joint topic model named 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. Within our framework, adverbs are modeled as a composite distribution within the frequency space of their kernels in a Gaussian process that represents the latent trajectory of motions. Consequently, it becomes feasible to estimate the constituents of sensory adverbial motions. When compared to the simple Neural Net model, our model exhibits superior performance on classification of adverbs. Our approach considers motions as a mixture of diverse frequency components, leading to the successful generation of appropriate adverbs from motion features in our empirical investigations.

5 Limitations

The primary limitation to the generalization of these results lies in the scarcity of datasets containing adverbially annotated human motions. There is no other way to annotate adverbs by ourselves to capture the meaning of adverbs which describe human motions, and it is difficult to make comparisons with other models because there are few studies working on the same research topic. Another limitation is that even if the adverbs output by the model are correct, such as synonyms, the model may judge that it has output the wrong one unless it is an exact match. We think this can be resolved by representing the adverbs in embedding vectors to evaluate output.

6 Ethical considerations

All datasets used in the experiments are either publicly available or have been licensed for use by the authors. In addition, all copyrights to the data generated using crowdsourcing were transferred to the authors.

References

Michael Ahn et al. 2022. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*.

David M. Blei and Michael I. Jordan. 2003. Modeling annotated data. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*.

S. Bochner, S. Trust, M. Tenenbaum, and H. Pollard. 1959. *Lectures on Fourier Integrals*. Annals of Mathematics Studies. Princeton University Press.

Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of GSCL*, pages 31–40.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

Aakanksha Chowdhery et al. 2022. Palm: Scaling language modeling with pathways.

Trevor Cohn, Daniel Preoţiuc-Pietro, and Neil Lawrence. 2014. Gaussian processes for natural language processing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials*, pages 1–3, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. 2020. Action Modifiers: Learning from Adverbs in Instructional Videos.

Danny Driess et al. 2023. Palm-e: An embodied multimodal language model. *CoRR*, abs/2303.03378.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR.

Hiroshi Ito, Hideyuki Ichiwara, Kenjiro Yamamoto, Hiroki Mori, and Tetsuya Ogata. 2022. Integrated learning of robot motion and sentences: Real-time prediction of grasping motion and attention based on language instructions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 5404–5410.

Aleksandra Kalinowska, Patrick M. Pilarski, and Todd D. Murphey. 2023. Embodied communication: How robots and people communicate through physical interaction. *Annual Review of Control, Robotics, and Autonomous Systems*, 6:205–232.

Siddharth Karamcheti, Suraj Nair, Annie S. Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. 2023. Language-driven representation learning for robotics.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *3DV*, pages 239–248. IEEE Computer Society.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(1-3):503–528.

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nicholas Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's cookin'? interpreting cooking videos using text, speech and vision. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 143–152, Denver, Colorado. Association for Computational Linguistics.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. 2017. A Simple yet Effective Baseline for 3D Human Pose Estimation. In *ICCV 2017*, pages 2640–2649.

T.P. Minka. 2003. Estimating a dirichlet distribution. *Annals of Physics*, 2000(8):1–13.

Bo Pang, Kaiwen Zha, and Cewu Lu. 2018. Human Action Adverb Recognition: ADHA Dataset and A Three-Stream Hybrid Model. *CoRR*, abs/1802.01144:2438–2447.

Matthias Plappert, Christian Mandery, and Tamim Asfour. 2018. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26.

Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian processes for machine learning.* Adaptive computation and machine learning. MIT Press.

David E. Rumelhart and James L. McClelland. 1987. *Learning Internal Representations by Error Propagation*, pages 318–362.

Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. ProgPrompt: Generating situated robot task plans using large language models. In *International Conference on Robotics and Automation (ICRA)*.

Michael L. Stein. 1999. *Interpolation of spatial data*. Springer Series in Statistics. Springer-Verlag, New York. Some theory for Kriging.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

T. Taniguchi, D. Mochihashi, T. Nagai, S. Uchida, N. Inoue, I. Kobayashi, T. Nakamura, Y. Hagiwara, N. Iwahashi, and T. Inamura. 2019. Survey on frontiers of language and robotics. *Advanced Robotics*, 33(15-16):700–730.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55.

Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. 2019. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. 2023. Chatgpt for robotics: Design principles and model abilities. Technical Report MSR-TR-2023-8, Microsoft.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Andrew Gordon Wilson and Ryan Prescott Adams. 2013. Gaussian Process Kernels for Pattern Discovery and Extrapolation. In *ICML 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1067–1075. JMLR.org.

Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. 2018. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robotics and Automation Letters*, 3(4):3441–3448. Publisher Copyright: © 2016 IEEE.