スペクトル混合カーネルとガウス過程に基づく動画からの副詞の意味理解

谷口巴
お茶の水女子大学
g1620524@is.ocha.ac.jp

持橋大地
統計数理研究所
daichi@ism.ac.jp

長野匡隼　中村友昭
電気通信大学
n1832072@edu.cc.uec.ac.jp
tnakamura@uec.ac.jp

長井隆行　高野渉
大阪大学
nagai@sys.es.osaka-u.ac.jp
takano@sigmath.es.osaka-u.ac.jp

小林一郎
お茶の水女子大学
koba@is.ocha.ac.jp

1　はじめに

近年重要性が高まってきている家庭用ロボットには，日常生活において人と同じ感覚を共有した動作が期待される．動作に対する感覚は，自然言語では副詞を通じて表現されることが多い．ゆえに，特定の副詞を表現する複数の動作に共通する特徴を見つけることができれば，ロボットはその副詞の意味を本質的に理解したといえる．副詞認識についての先行研究として，Pang ら [1] が表情認識や画像情報を用いてアプローチしているが，動作との関係性を捉えることは難しく，実際にロボットを動かすことを考えると実用的ではない．本研究ではロボットに適用する前段階として，人の動作の特徴を通じて副詞の意味を理解することを試みる．具体的には，人の動作を GPLVM [2] で圧縮して得られた非線形な潜在空間での軌跡を，スペクトル混合カーネル [3] を用いたガウス過程で表現する．さらに各次元の軌跡を構成する複数の周波数成分を特定し，副詞との対応関係を捉える，周波数空間でのマルチモーダルなトピックモデルを提案する．これにより，動作について副詞を用いて表現することや、副詞表現から動作を生成することが可能となる．

2　動作と副詞の結合トピックモデル

2.1　ガウス過程潜在変数モデル (GPLVM)

本研究では，動作から得られる高次元の姿勢情報を低次元に圧縮してモデル化するため，ガウス過程に基づく教師なし学習であるガウス過程潜在変数モデル (GPLVM) [2] を用いる．GPLVM とは，ガウス過程に基づく非線形な確率的主成分分析であり，N 個の D 次元観測値をまとめた行列 $\boldsymbol{Y}$ について，式 (1) を最大化するような低次元の入力 $\boldsymbol{X}$ を計算する．

$$p(\boldsymbol{X},\boldsymbol{Y}) = p(\boldsymbol{Y}|\boldsymbol{X})p(\boldsymbol{X}) \tag{1}$$

ここで，$\boldsymbol{X}$ は未知であるため $p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ と仮定して，$p(\boldsymbol{Y}|\boldsymbol{X})$ を考える．$\boldsymbol{Y}$ はガウス過程に従い，$\boldsymbol{X}$ がわかれば出力の各次元が独立であると仮定すると，データ全体 $\boldsymbol{Y}$ の確率は $\boldsymbol{y}^{(1)}...\boldsymbol{y}^{(D)}$ の積であるため，$p(\boldsymbol{Y}|\boldsymbol{X})$ は以下の式で表される．ただし $\boldsymbol{K}_{\boldsymbol{X}}$ は共分散行列，$k(x,x')$ はガウス過程のカーネル関数であり，$K_{i,j} = k(x_i, x_j)$ で定義される．

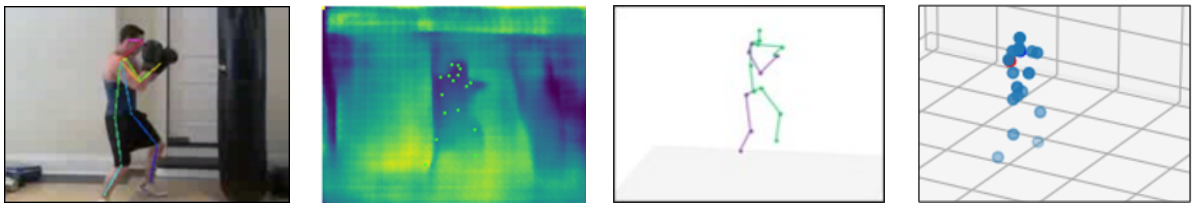
$$p(\boldsymbol{Y}|\boldsymbol{X}) = (2\pi)^{-\frac{ND}{2}} |\boldsymbol{K}_{\mathbf{X}}|^{-\frac{D}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}_{\boldsymbol{X}}^{-1}\boldsymbol{Y}\boldsymbol{Y}^{\mathbf{T}})\right) \tag{2}$$

本研究では $\boldsymbol{X} \to \boldsymbol{Y}$ の GPLVM のカーネル関数として RBF カーネルを使用し，L-BFGS 法を用いて $\boldsymbol{X}$ およびカーネルのハイパーパラメータを最適化する．図 3 に，実際の動作から計算した $\boldsymbol{X}$ の例を示した．

2.2　スペクトル混合カーネル

上で得られた潜在空間 $\boldsymbol{X}$ での動作の軌跡について，その特徴を捉えることを試みる．Wilson ら [3] はガウス過程で使用する基底を，既存の基底やその組み合わせに限定せず，フーリエ領域で混合ガウス分布を考えることでデータから自動的に学習できるスペクトル混合カーネル (Spectral Mixture Kernel, SM kernel) という手法を提案した．ここではガウス過程の基底として，値が $\tau = x - x'$ だけに依存する定常基底関数 $k(\tau)$ を考える．ボホナーの定理より，任意の $k(\tau)$ は以下の形で表される．

$$k(\tau) = \int_{\mathbb{R}^D} e^{2\pi i s^{\mathrm{T}}\tau} \psi ds \tag{3}$$



(a) Openpose による画面座標推定 **(b)** FCRN-depth による深度推定 **(c)** 3 次元の骨格座標の推定 **(d)** 回転行列による方向正規化

図 1: 動画データの前処理による動作の骨格座標の抽出.

$k(\tau)$ は周波数領域での確率密度 $\psi(s)$ と等価なので, $\psi(s)$ に関して混合ガウス分布を考える. ガウス分布の各要素は, もとの領域では以下の基底関数を考えていることと等価となる.

$$k(\tau|\sigma,\mu) = \exp(-2\pi^2\tau^2 v^2)\cos(2\pi\tau\mu) \tag{4}$$

すなわち基底として, 次の Q 個の基底関数の混合を考えていることになる. ただし, μ_q^d と v_q^d は q 個目の基底における入力 $\boldsymbol{X}$ における d 次元目の平均と分散である.

$$k(\tau) = \sum_{q=1}^{Q} w_q \cos(2\pi\tau^{\mathrm{T}}\mu_q) \prod_{d=1}^{D} \exp(-2\pi^2\tau_d^2 v_q^d) \tag{5}$$

パラメータの重み $\mathbf{w}$, 平均 $\boldsymbol{\mu}$, 分散 $\boldsymbol{\sigma}$ は通常のガウス過程のハイパーパラメータ最適化で学習できる. 本研究ではこの手法を用いて, 各動画について GPLVM で圧縮した 3 次元の潜在変数 $\boldsymbol{X}$ から, 副詞と関係があると予想される $Q=4$ 個の周波数成分を抽出して観測値とする. 例を図 4 に示した. [1)]

2.3 スペクトル混合潜在ディリクレ配分法 (Spectral Mixture LDA)

動作から抽出された周波数成分は, その動作に付与された副詞と関係があると考えられる. そこで, 潜在ディリクレ配分法 (LDA) [4] を拡張し, 各動作 d に K 次元の潜在的な「トピック分布」θ_d があると仮定する. このとき, 動作に付与された副詞

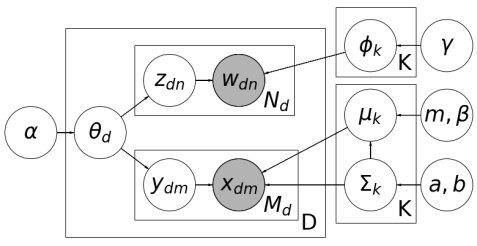


図 2: SMLDA のグラフィカルモデル.

1) $\boldsymbol{X}$ での軌跡を直接フーリエ変換することもできるが, その場合は関数がどこを通るか (関数の位相) と関数の特徴を分離することができない. スペクトル混合カーネルを用いることにより, 純粋に軌跡の特徴だけを抽出することができる.

$\{w_{dn}\}$ $(n = 1 \ldots N_d)$ および動作の周波数成分 $\{x_{dm}\}$ $(m = 1 \ldots M_d)$ は, 次のモデルで生成されたと考える.

1. Draw $\theta_d \sim \mathrm{Dir}(\alpha)$.
2. For $n = 1 \ldots N_d$,
   – Draw $z_{dn} \sim \theta_d$; Draw $w_{dn} \sim \mathrm{Mult}(\phi_{z_{dn}})$.
3. For $m = 1 \ldots M_d$,
   – Draw $y_{dm} \sim \theta_d$; Draw $x_{dm} \sim \mathcal{N}(\mu_{y_{dm}}, \Sigma_{y_{dm}})$.

このグラフィカルモデルを図 2 に示した. このマルチモーダルなトピックモデルを, 本論文ではスペクトル混合 LDA (Spectral Mixture LDA; SMLDA) と呼ぶ. SMLDA では、周波数成分と副詞は動作毎に同じトピック分布 θ_d を共有している. ここで ϕ_k, $\mathcal{N}(\mu_k, \Sigma_k)$ はそれぞれ, k 番目のトピックに対応する副詞の多項分布および周波数のガウス分布であり, 互いの情報を用いて, 各動画について 1 つ 1 つの副詞と周波数成分にトピックを割り当てていく.

副詞と周波数についてのサンプリング ギブスサンプリングにより, 副詞と周波数のトピック分布を学習していく. 副詞 w_{dn} のトピック z_{dn} は, 次式を用いてサンプリングする. ここで V は語彙数を表す.

$$p(z_{dn}=k|\boldsymbol{W},\boldsymbol{X},\boldsymbol{Z}_{\backslash dn},\boldsymbol{Y},\alpha,\beta,\gamma) \propto (N_{dk\backslash dn} + M_{dk} + \alpha)\frac{N_{kw}+\beta}{N_{k\backslash dn}+\beta\boldsymbol{V}} \tag{6}$$

ハイパーパラメータである α と γ は不動点反復法により, 以下の式を用いて更新する. ここで登場する Ψ はディガンマ関数 $\Psi(x) = d/dx \log\Gamma(x)$ である.

$$\alpha^{new} = \alpha\frac{\sum_{d=1}^{D}\sum_{k=1}^{K}\Psi(N_{dk}+M_{dk}+\alpha) - DK\Psi(\alpha)}{K\sum_{d=1}^{D}\Psi(N_d+M_d+\alpha K) - DK\Psi(\alpha K)} \tag{7}$$

$$\gamma^{new} = \gamma\frac{\sum_{k=1}^{K}\sum_{v=1}^{V}\Psi(N_{kv}+\gamma) - KV\Psi(\gamma)}{V\sum_{k=1}^{K}\Psi(N_k+\gamma V) - KV\Psi(\gamma V)} \tag{8}$$

周波数成分 x_{dm} のトピック y_{dm} に関しては, 副詞の単語分布をガウス分布の確率密度関数に置き換え, 以下の式を用いてサンプリングする.

$$p(y_{dm}=k|\boldsymbol{W},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y}_{\backslash dm},\alpha,\beta,\gamma) \propto \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(x_{dm}-\mu_k)^2}{2\sigma_k^2}\right)\frac{N_{dk}+M_{dk\backslash dm}+\alpha}{N_d+M_d-1+\alpha K} \tag{9}$$

パラメータである μ と σ は以下の事後分布からサンプリングする．ここで $\lambda = 1/\sigma^2$ である．

$$p(\mu|\mathbf{Y}) = \mathrm{N}(\mu|m, (\beta\lambda)^{-1}),\ p(\lambda|\mathbf{Y}) = \mathrm{Gam}(\lambda|a, b) \quad (10)$$

ただし a_0, b_0, β_0, m_0 を事前分布のパラメータとして

$$a = \frac{M}{2} + a_0,\ b = \frac{1}{2\sum_{m=1}^{M} x_m^2 + \beta_0 m_0^2 - \beta m^2} + b_0,$$
$$\beta = M + \beta_0,\ m = \frac{1}{\beta}\left(\sum_{m=1}^{M} x_m + \beta_0 m_0\right). \quad (11)$$

3 実験

3.1 使用するデータ

YouTube に掲載されている, 100 種類の異なる歩行動作を集めた動画[2]を用いて実験を行った．クラウドソーシングシステム Lancers[3]を用いて，20 名のアノテーターに動画の各動作について思いつく限り自由に副詞をアノテーションしてもらうよう依頼した．全動画で 3 個以上出現した副詞に限定し，満たない副詞はノイズとして除去した．また「ゆっくり」という副詞は多くの動画に付けられていたため，各動画につき 3 個以上付けられた場合のみ採用した．この結果，1 つの動画につき平均 12.93 個の副詞がアノテーションされたデータが得られた．

データの前処理 上記の動画データから，以下のように 4 段階で 3 次元の骨格座標の推定を行った．

1. Openpose [5] を用いて動画データから 2 次元の骨格座標を推定 (図 1(a))
2. FCRN-depth prediction [6] を用いて動画データの深度を推定 (図 1(b))
3. 1,2 の推定結果と 3d-pose baseline [7] を用いて動画データから 3 次元の骨格座標を推定 (図 1(c))
4. 歩いている人の体の向きを合わせるため, 回転行列を用いて正規化 (図 1(d))

周波数成分の抽出 前処理したデータから各関節間ごとに方向ベクトルを算出し, 入力データとした. 以下の 2 つの手法を用いて, 前処理した動画データから周波数成分を抽出した.

1. GPLVM を用いて 48 次元の姿勢データを 3 次元の潜在変数に圧縮する (図 3)
2. SM kernel を用いて 3 次元の潜在変数から, 各次元について周波数成分を抽出する (図 4)

2） https://www.youtube.com/watch?v=HEoUhlesN9E
3） https://www.lancers.jp/

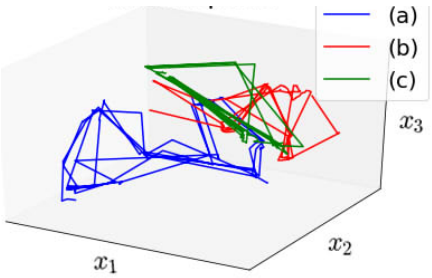


図 3: GPLVM による動作の非線形次元圧縮．ここでは (a)–(c) の異なる歩行動作が, 低次元の潜在空間 $\boldsymbol{X}$ の軌跡として表現されている．

学習データのうち, 3 個の動作を GPLVM で圧縮した 3 次元の潜在空間にプロットしたものを図 3 に示す. 歩く動作は繰り返しの動作であるため, 潜在変数は図のように円を描くような動きになる. SM kernel によって各動画の 1 次元目について最適化された平均 $\boldsymbol{\mu}$ と分散 $\boldsymbol{\sigma}$ をパラメータとしてガウス分布を描画したものを図 4 に示した.[4] 式 (5) より平均 μ の値が大きいほど周期が小さくなることから，値の変動が低速な動画データほど基底を表すスペクトルは左側に多く見られると推測できる. よって, (a) は遅い動きの成分が多く, (c) は速い動きの成分が多く，(b) はその中間的な動きということがわかる. SM kernel では構成されるカーネル関数に関して重み $\mathbf{w}$, 平均 $\boldsymbol{\mu}$, 分散 $\boldsymbol{\nu}$ が推定されるが, 関数の種類に着目するため, 平均 $\boldsymbol{\mu}$ のみを周波数成分として今後使用する. 本実験ではこの抽出された周波数成分と動画に付与された副詞の集合を入力データとした. 共にトピック数を 4, 十分収束するよう, MCMC の繰り返し数は 1000 と大きい値に設定した. 混合ガウス分布の部分に関して, 分散はデータの幅に合わせて 4 つのガウス分布が均等に配置されるよう, σ=3.75 と固定して平均のみ学習する.

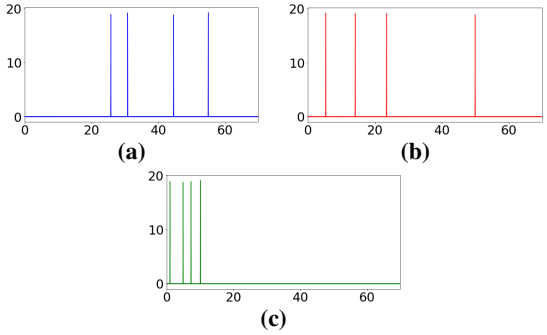


図 4: 図 3 の各動作について, 1 次元目の軌跡を解析したスペクトル混合カーネルによる周波数表現.

4） 推定された分散がきわめて小さいため, 図ではガウス分布がデルタ関数状に描画されている.

表 1: SMLDA で得られたトピック別副詞上位 7 語.

Topic 1	Topic 2	Topic 3	Topic 4
けだるげに	辛そうに	軽やかに	ゆっくりと
普通に	痛そうに	軽快に	恐る恐る
だるそうに	重たげに	足早に	寒そうに
後ろ向きで	硬く	テンポよく	慎重に
大股で	ぎこちなく	急いで	避けながら
ぶらぶらと	小刻みに	颯爽と	そろりそろりと
疲れて	不自然に	リズミカルに	怯えながら

表 2: LDA だけで得られたトピック別副詞上位 7 語.

Topic 1	Topic 2	Topic 3	Topic 4
ゾンビ風に	急いで	楽しそうに	慎重に
だらだらと	ロボット風に	堂々と	こっそりと
気だるく	ちょこまか	力強く	不安げに
疲れた	速足で	リズミカルに	そろそろと
のんびりと	慌てて	サッサッ	警戒しながら
疲れ果てて	すばやく	スッスッ	おどおど
辛い	そわそわ	芝居がかった	静かに

3.2 実験結果

学習されたトピック–単語分布から各副詞について NPMI [8] を計算した上位 7 語を表 1 に示す. 比較のため，周波数成分の情報を使わず LDA で解析した結果を表 2 に示した．動作の情報を同時に用いることで, 副詞のトピックがより明確になっていることがわかる. また学習された平均 $\boldsymbol{\mu}$ を用いてガウス分布を描画したものを図 5 に示す. 収束しているか確認するため, MCMC の各繰り返しごとに計算したパープレキシティをプロットしたものを図 6 に示す. かなり早い段階で収束していることがわかる.

3.3 モデルの評価

データを訓練用と評価用に 8 対 2 に分割した. 評価用のデータについて，モデルに周波数データを与えた際の副詞のパープレキシティを算出した. また，モデルの評価のため，次の 2 つのモデルと比較する.

1. 評価用と同じ θ を用いて，副詞を一様分布を使ってランダムにサンプリングしたもの
2. θ そのものもランダムに生成し，副詞を一様分布を使ってランダムにサンプリングしたもの

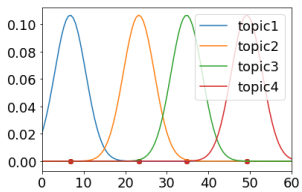


図 5: 学習した μ を用いて描画したガウス分布.

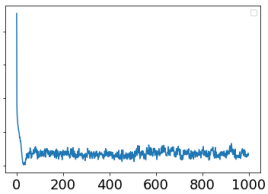


図 6: SMLDA の MCMC とパープレキシティ.

この結果，SMLDA のパープレキシティは 41.55, 上記のランダム 1 とランダム 2 はそれぞれ 90.75 および 92.84 であり，モデルが正しく副詞を予測していることがわかった. また評価用の動画 (図 7) に関して，周波数データを与えたとき，SMLDA が推定した θ を図 8 に，確率の高い副詞を表 3 に示す.

図 7: 評価用の動画の例.

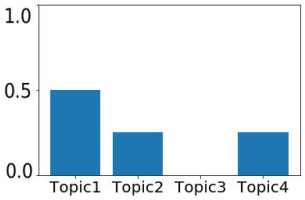


図 8: 動作から推定された θ.

副詞	予測確率
ゆっくり	0.098
普通に	0.034
ゆっくりと	0.028
慎重に	0.022
大股で	0.022
ぎこちなく	0.020
痛そうに	0.016

表 3: 動画から予測される確率の高い副詞上位 7 語.

3.4 考察

表 1 では，副詞を動作の情報も用いて意味的にクラスタリングすることに成功している．図 5 では，狙い通りデータの幅を均等に分けるようにガウス分布が配置されている形となっている．このことから，SMLDA は副詞情報，周波数成分情報ともに互いの情報を用いて，クラスタリングが成功していることが確認できる．また SMLDA はランダムに副詞を生成するモデルよりパープレキシティが低く，動作を表す周波数成分から副詞を生成するモデルとして有効であることが示された．最後に表 3 より，動画に対して妥当な副詞がサンプリングされていることが確認できた.

4 まとめと今後の課題

歩行動作をする人間の骨格座標を GPLVM を用いて 3 次元の潜在変数に圧縮し, SM kernel を用いて周波数成分を抽出した. 次に SMLDA を用いて, 動画にアノテーションされた副詞データと周波数データを, お互いの情報を使いながら 4 つのトピックにクラスタリングした. 最後にパープレキシティを用いて評価し, モデルの有効性を示した. 現在は 3 次元に圧縮した潜在変数を 1 次元ずつスペクトル混合カーネルを用いて解析しているが, 今後はこれを直接 3 次元で同時に行うことを検討している.

謝辞 本研究は，科学研究費補助金・基盤 (B) (18H03295) の支援を受けて行った.

参考文献

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

[2] Michalis K. Titsias and Neil D. Lawrence. Bayesian Gaussian Process Latent Variable Model. In *AISTATS 2010*, pp. 844–851.

[3] Andrew Gordon Wilson and Ryan Prescott Adams. Gaussian Process Kernels for Pattern Discovery and Extrapolation. In *ICML 2013*, pp. 1067–1075.

[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 994–1022, 2003.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 43, No. 1, pp. 172–186, 2021.

[6] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *3DV*, pp. 239–248, 2016.

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

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