Spectral Mixture Kernel を用いた動作を表す副詞の意味理解へ向けた取り組み

谷口巴　(指導教員：小林一郎)

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文中の副詞を通じて表現される．このことから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理解したといえる. 本研究では，ロボットを使う前段階として，人の動作を対象とし，副詞の意味を人の動作特徴を通じて理解することを試みる．具体的な方法として，Gaussian Process Latent Variable Model [1] を用いて動画像中の人の多次元である姿勢情報を 3 次元に圧縮した後，Spectral Mixture Kernel [2] を用いて, 各次元の時系列データを構成する複数のカーネルを特定し，特定の副詞に共通するカーネルを発見することで，その副詞の意味理解を目指す.

2　提案手法

2.1　姿勢情報の抽出と正規化

副詞認識の先行研究として，Pang ら [3] により人の動作を副詞で表現した Adverbs Describing Human Actions (ADHA) データセットが作成されている．本研究では，このデータセット中の人の動作が映る動画像データとそれぞれに付けられた副詞のキャプションデータを使用する．動画像のキャプション情報は 3 人のアノテータ別に用意されており、各人一つの動画に対して平均 1.81 個の副詞を付けている．32 個の動作に対して各々400 個程度の動画像データが収められており，その中で上半身が確実に映っている「punch」の動作に着目して動作と副詞の意味の対応関係を捉えることにする．人動作データの前処理として，以下の 4 つの手法を施行した．

1. **2 次元骨格座標抽出**
   Cao ら [4] が提案した OpenPose を用いて，動画像内に映る人物の 2 次元骨格座標を抽出する．

2. **深度解析**
   Laina ら [5] が提案した FCRN-Depth Prediction を用いて動画像内に映る物体の深度を解析することにより 3 次元の骨格座標を抽出する．

3. **3 次元骨格座標推定**
   Martinez ら [6] が提案した 3d-pose-baseline を用いて人物の 3 次元骨格座標を推定する．本研究では 17 点の関節点が推定される．特徴量として前述の 2 次元骨格座標と深度解析結果を用いる．推定した 3 次元座標から 16 個の単位方向ベクトルを計算することにより各関節角情報が得られる．他, 対象人物の違いによってデータが異なることを防げる.

4. **姿勢の正規化**
   人物の動作の方向を揃えるため，右腰の座標を基準に回転行列を使って姿勢の正規化を行う．動画像内では対象とする人物の下半身が映っていないものが多数存在したため，上半身を対象に 1 フレームで算出された 8 個からなる正規化した方向ベクトルを 1 単位として扱う.

2.2　GPLVM を用いた潜在空間への次元圧縮

次に作成したデータセットを入力として Gaussian Process Latent Variable Model(GPLVM) を用いて 3 次元の潜在空間に写像する。ここで GPLVM とは，確率的主成分分析にガウス過程を導入したものである.$N \times D$ 行列の出力を $\boldsymbol{Y}$, 潜在的な入力を $\boldsymbol{X}$ として

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

を最大化するような $\boldsymbol{X}$ を見つけることを目標とする. 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．$\boldsymbol{Y}$ はガウス分布に従い，$\boldsymbol{X}$ 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 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 $\boldsymbol{K_X}$ は共分散行列，$k(x, x')$ はガウス基底関数であり，$K_{X_{(i,j)}} = k(x_i, x_j)$ で定義される.

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{2\pi^{\frac{ND}{2}}|\boldsymbol{K_X}|^{\frac{D}{2}}} \exp(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{K_X^{-1}YY^T}))$$

よって，L-BFGS 法などにより基底関数のハイパーパラメータを最適化することで，$\boldsymbol{X}$ が算出される.

2.3　SMKernel を用いた基底分解

Wilson ら [2] はガウス過程で使用する基底を，既存の基底および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 Spectral Mixture 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．ホボナーの定理より任意の $k(\tau)$ は以下の形で表される.

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

$k(\tau)$ 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である.

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

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.

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

パラメータ w，μ，σ は通常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 次元の潜在変数から各副詞で使用されている複数の基底を調べる．

3 実験

SMKernel を用いて複数の時系列データにおいて各々で使用されている基底を調査した実験について述べる

3.1 実験設定

対象データにはキャプションとして副詞 fast がつけられたデータを 6 個と slowly がつけられたデータを 6 個連続で繋げたものと，heavily がつけられたデータを 6 個と lightly がつけられたデータを 6 個連続で繋げたもの 2 つを用いた．それぞれの副詞における動画の例を図 1 に示す．キャプションについては 3 人のアノテータの中で 2 人以上がつけた副詞を採用する．一つの動画に対し 2 つ以上の副詞がキャプション付けされている．$N \times 3$ 行列の潜在変数 $\boldsymbol{X}$ について，データを 1 次元ごとに分解してそれぞれ入力する．基底の数は 4 に固定している．

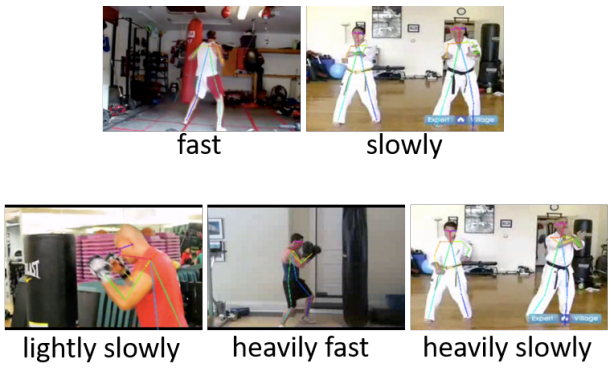

図 1: 入力動画情報

3.2 実験結果

最適化された μ と σ をパラメータとしてガウス分布を出力する. 本実験では使用される基底の種類に着目するため，重み w は無視する．各副詞について 6 個の動画を解析したが，中でも基底の特徴がわかりやすい 5 つの解析結果を以下に紹介する．また，lightly と heavily についてそれぞれ fast または slowly と共起している頻度が高いことが判明した．そのため実験結果のキャプションは lightly と fast，lightly と slowly，heavily と fast，heavily と slowly が共起している動画をそれぞれ異なるキャプションとして区別する．

3.3 考察

前述の $k(\tau)$ の式より μ の値が大きいほど周期が大きくなることから，値の変動が低速な動画データほどスペクトルは右側に多く見られると推測できる．fast と slowly の実験結果ではまさにそれが明白となってい

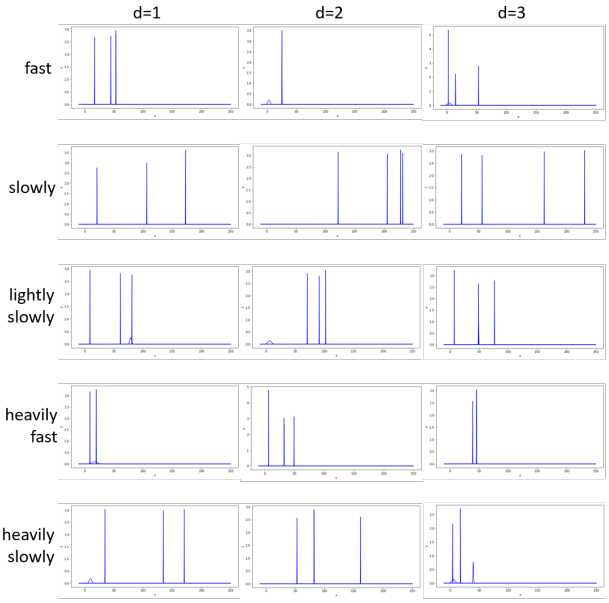

図 2: 各副詞に使用される基底情報

る．lightly と slowly がつけられた動画である．lightly といえば軽やかに動くイメージで値が素早く動いているはずではあるが slowly の要素が混同されてスペクトルも fast と比べて少々右に寄っているのが確認できる．最後に heavily ついては fast と slowly の以上の 5 つの結果から fast と slowly について使われる基底が異なっている様子が確認でき，lightly と heavily についてはそれと共起する副詞によって使用する基底が変化する事が判明した．

4 まとめと今後の課題

動画像から人物の 3 次元骨格座標を抽出した後，GPLVM を用いて 24 次元の時系列データを 3 次元の潜在変数に圧縮した．また SMKernel を用いて異なった副詞がキャプション付けされている動作に関して使用される基底の違いを観察した．今後は動画内における人の動作回数を正規化して実験し，提案手法の正当性を検証していく．

参考文献

[1] M. K. Titsias and N. D. Lawrence (2010) Bayesian Gaussian Process Latent Variable Model. Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR: WCP 9, pp. 844-851.

[2] A. G. Wilson and R. P. Adams, “Gaussian process kernels for pattern discovery and extrapolation,” in Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp. 1067–1075.

[3] Bo Pang, Kaiwen Zha and Cewu Lu. (2017) Human Action Adverb Recognition: ADHA Dataset and A Three-Stream Hybrid Model , arXiv preprint arXiv:1802.01144, 2017.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh (2018) OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CoRR, Vol.abs/1812.08008.

[5] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, Nassir Navab (2016) Deeper Depth Prediction with Fully Convolutional Residual Networks, CoRR, Vol.abs/1606.00373.

[6] Martinez, Julieta and Hossain, Rayat and Romero, Javier and Little, James J. (2017),A simple yet effective baseline for 3d human pose estimation,ICCV.