

OpenPose で取得した時系列骨格データからの深層生成モデルによる個人識別

浅見直人† Alessandro Moro‡ 池勇勲† 梅田和昇†

† 中央大学 ‡ RITECS Inc.

E-mail: asami@sensor.mech.chuo-u.ac.jp

1 はじめに

動画画像内に映った歩行者をその歩き方から識別する歩容認証が犯罪捜査などで利用されている。従来の歩容認証手法の多くはシルエット画像に基づくものである [1, 2] が、それらは荷物の所持の違いや服装の違いによる認証精度低下の問題がある。そこで本研究では、動画画像から骨格データを推定し、得られた時系列骨格データから人物を識別する。

一般に歩容特徴は、観測方向や歩行速度、歩行周期のばらつきなどの共変量を受け変化するため、共変量の影響による変化を吸収できる認証手法が求められている。また、年齢や性別といった人物ラベル以外のラベルの分類をサブタスクとして加えることで、歩容認証の精度を向上させた研究も存在する [3]。そのような共変量や人物ラベル以外のラベルは、歩容特徴を形作る潜在変数と考えることができる。そこで本研究では、様々なラベルを潜在変数とみなした深層生成モデルによる識別を目指し、その第一段階として人物ラベルのみを潜在変数とした混合ガウス分布を用いた変分オートエンコーダ (VAE: Variational autoencoder) によるクラス分類を提案する。

2 関連研究

一般的な生成モデルでは、仮定した確率モデルの潜在変数の事後分布を推論する。深層生成モデルは、その確率モデルや近似事後分布を深層学習を用いて表現したモデルである。深層生成モデルの代表的なものに VAE があり、VAE は潜在変数の近似事後分布をニューラルネットワークで表現する。深層生成モデルを用いてクラス分類を行っている研究は少なく、深層生成モデルを用いて半教師あり学習を行った例 [4] でも通常のクラス分類と同様な教師データによるクロスエントロピー誤差を利用している。

VAE に混合ガウス分布を用いた例としては、[5] と [6] が存在し、VAE を用いてクラスタリングを行っている。[5] はエンコーダによって混合ガウス分布に従う潜在変数の近似事後分布を取得し、k-means 法と同様の計算によってカテゴリ潜在変数の事後分布を計算して

いる。明示的に混合ガウス分布からサンプリングしているが、カテゴリ潜在変数が離散のため効率的な学習が困難である。一方、[6] は連続化したカテゴリ潜在変数を用いているが、非明示的に混合ガウス分布からのサンプリングを行っている。本研究では、クラス分類を目的とするため、エンコーダからカテゴリ潜在変数を直接推論し、さらに明示的な混合ガウス分布からのサンプリングを行うモデルを提案する。

3 生成モデルと認識モデル

本研究の生成モデルは $p(\mathbf{x}, \mathbf{z}, \mathbf{s}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{s}, \mathbf{z})$ とする。ここで $\mathbf{x}$ は骨格を表すデータ、 $\mathbf{z}$ は混合ガウス分布に従う潜在変数、 $\mathbf{s}$ は人物ラベルを意味するカテゴリ潜在変数である。潜在変数の事前分布 $p(\mathbf{s}, \mathbf{z})$ は式 (1) で表される混合ガウス分布である。

$$p(\mathbf{s}, \mathbf{z}) = p(\mathbf{z}|\mathbf{s})p(\mathbf{s}) = \prod_{i=1}^K \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_i, \Sigma_i)^{s_i} p(\mathbf{s}) \quad (1)$$

ここで、 K はクラス数、 Σ_i は各クラスの共分散行列、 $\boldsymbol{\mu}_i$ は各クラスの平均を表す。また、 $p(\mathbf{s})$ は Concrete distribution [7] と呼ばれる連続化したカテゴリカル分布である。近似事後分布 $q(\mathbf{s}, \mathbf{z}|\mathbf{x})$ を事前分布と同様の混合ガウス分布と仮定して VAE による潜在変数の推論を行う。VAE を作成するには、近似事後分布からのパラメータへの勾配逆伝播可能なサンプリングと、変分下限の計算が必要となる。

3.1 サンプリング

Concrete distribution $p(\mathbf{s})$ は Gumbel-Softmax trick によってサンプリングできる [7]。また、条件付き分布 $p(\mathbf{z}|\mathbf{s})$ は式 (2) で表される一つのガウス分布に変形できる。

$$p(\mathbf{z}|\mathbf{s}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\text{gmm}}, \Sigma_{\text{gmm}}) \quad (2)$$

その共分散行列は

$$\Sigma_{\text{gmm}} = \left\{ \sum_{i=1}^K s_i \Sigma_i^{-1} \right\}^{-1} \quad (3)$$

で表され、その平均は

$$\boldsymbol{\mu}_{\text{gmm}} = \Sigma_{\text{gmm}} \left\{ \sum_{i=1}^K s_i \Sigma_i^{-1} \boldsymbol{\mu}_i \right\} \quad (4)$$

で表される。この式はカテゴリ潜在変数 s が連続でも成り立つ。結局、 z はガウス分布の reparameterization trick によってサンプリングできる。

3.2 変分下限

変分下限 $\mathcal{L}[\phi, \theta]$ は式 (5) で与えられる。

$$\mathcal{L}[\phi, \theta] = \mathbb{E}_{s, z \sim q_\phi(s, z|x)} [\log p_\theta(x|z)] - D_{KL}[q_\phi(z, s|x) || p(z, s)] \quad (5)$$

ここで、 ϕ はエンコーダのニューラルネットワーク、 θ はデコーダのニューラルネットワークを表す。式 (5) の第 1 項はデコーダの出力の再構成誤差を表し、第 2 項は近似事後分布 $q_\phi(s, z|x)$ と事前分布 $p(s, z)$ の KL ダイバージェンスの負の値である。近似事後分布と事前分布の KL ダイバージェンスは式 (6) に変形できる。

$$\mathbb{E}_{s \sim q_\phi(s|x)} \left[\log \frac{q_\phi(s|x)}{p(s)} \right] + \mathbb{E}_{s \sim q_\phi(s|x)} \left[D_{KL}[q_\phi(z|s, x) || p(z|s)] \right] \quad (6)$$

式 (6) の期待値計算の中身は解析的に計算できるので、式 (5) の第 2 項は Concrete distribution によるモンテカルロ推定で近似計算できる。提案手法の構造を図 1 に示す。図 1 における丸で囲われた変数は生成モデルにおける確率変数である。エンコーダは Concrete distribution のパラメータ π 、各クラスの平均ベクトル μ_i 、対角行列で表される各クラスの共分散行列 Σ_i を出力する。この π は ArcFace[8] と呼ばれる分類器の出力から得る。この π の事後分布の推論 $q_\phi(s|x)$ がクラス分類を表す。

3.3 教師データの与え方

VAE は教師あり学習の枠組みではなく、クラス分類問題を解くためには教師データを与える必要がある。教師データの与え方として、本研究では Concrete distribution の事前分布を用いる。具体的には、事前分布のパラメータを以下のように与える。

- カテゴリ潜在変数 s の事前分布 $p(s)$ のパラメータ π_{prior} を式 (7) で与える。

$$\pi_{\text{prior}} = \text{Softmax} \left(\frac{t + \epsilon}{T} \right) \quad (7)$$

ここで、 t は教師データを表すベクトル、 T は Softmax 関数の温度パラメータ、 ϵ は微小な正のノイズを表すベクトルである。

- 条件つき分布 $p(z|s)$ のパラメータは以下のように与える。ArcFace の識別層の重みベクトルが各クラスを代表するベクトルになる [8] ことを利用して、各クラスの平均として ArcFace の重みベクトルを与える。また共分散行列として通常の VAE と同様に単位行列を与える。

4 骨格データの取得

本章では、識別に用いる骨格データの取得方法を述べる。まず、天井に設置された魚眼カメラから動画を撮影し、深層学習を用いた一般物体認識手法である YOLO[9] を用いて人物領域を取得する。YOLO はあらかじめ魚眼画像で学習を行い、魚眼画像特有の歪んだ画像から人物領域を取得できる。さらに得られた人物領域について、正像変換 [10] を用いて魚眼画像を透視投影画像に変換する。最後に、変換された歪みの無い人物領域画像から深層学習を用いた姿勢推定手法である OpenPose[11] を用いて骨格座標を取得する。取得された骨格座標の例を図 2 に示す。骨格座標は首の座標が原点になるよう平行移動し、首から腰の左右の中心までの長さで全体を正規化する。図 3 に 2 フレーム毎の骨格データを示す。骨格座標データは時系列方向に並べることで、 x 座標 y 座標 2 チャンネルの 2 次元データにしてからエンコーダの CNN に入力される。

5 評価実験

本章では、提案手法とシステムの性能評価実験について述べる。

5.1 実験データ

本実験では、図 4 に示す二つのルートについて成人男性 5 人が歩行した動画から取得した骨格データを用いてクラス分類を行った。一人当たり合計で 3000 フレームの骨格データを訓練データ、1800 フレームの骨格データをテストデータに用いた。識別に利用するフレーム長は 10, 20 の 2 パターンで学習を行い評価した。学習は以下のモデルすべてそれぞれ 3000 エポック行い、学習係数は初期値 0.0001 として減衰させ、Gumbel Softmax-trick と事前分布の温度パラメータ T は初期値 3.0 として減衰させた。これら学習係数と温度パラメータの初期値の決定と減衰は実験的に行った。

クラス分類実験について、提案手法ではデータを入力したときの近似事後分布のパラメータ π によってクラス分類を行った。クラス分類実験では提案手法と比較するモデルとして、識別のみを行うモデルを用意した。識別のみのモデルでは、図 1 における ArcFace の出力と教師ラベルとのクロスエントロピー誤差で学習を行う。また、提案手法の Concrete distribution を一定温度以下で離散化したものも比較を行う。

歩容認証などの生体認証では、データベース上に登録された特徴と取得された特徴を比較することで本人認証と個人認証を行うことが多く、本システムでも認証手法としての利用を想定し、提案手法を特徴抽出器として利用したときの特徴抽出性能についても評価を行った。具体的には、以下に示す特徴について k 近傍法を

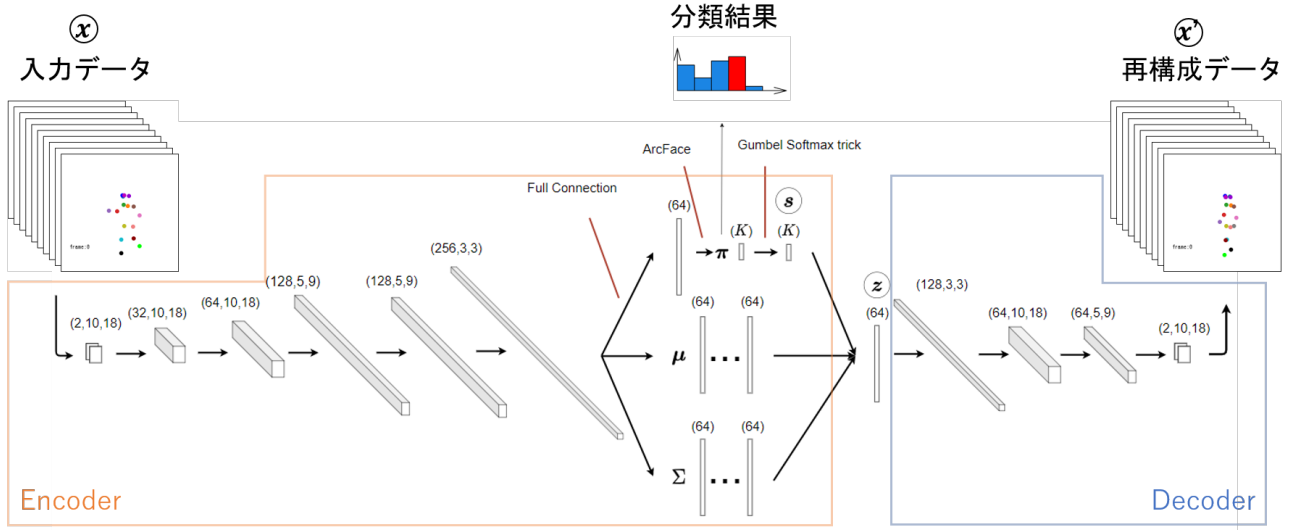


図 1 本研究における混合ガウス分布を用いた VAE の構造



図 2 取得された骨格座標データの例

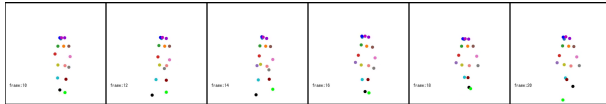


図 3 2 フレーム毎の骨格データ

用いてクラス分類を行った。

1. ArcFace(分類器) への入力ベクトル
2. 近似事後分布からサンプリングされた潜在変数 z
3. 近似事後分布 $q_\phi(z|s, x)$ の平均ベクトル μ_{gmm}

本実験では $k=5$ とした。また、特徴抽出性能についても識別のみのモデルと提案手法を離散化したものと比較した。識別結果は全て accuracy で評価した。

5.2 実験結果

表 1 に近似事後分布 $q_\phi(s|x)$ のパラメータ π によるクラス分類の結果を示す。各フレーム長において最良値を太字で示す。クロスエントロピー誤差で学習した識別のみのモデルに対して、提案手法である混合ガウス分布を用いた VAE は同等かそれ以上の識別性能を持つことが分かる。また、表 2 に k 近傍法によるクラス

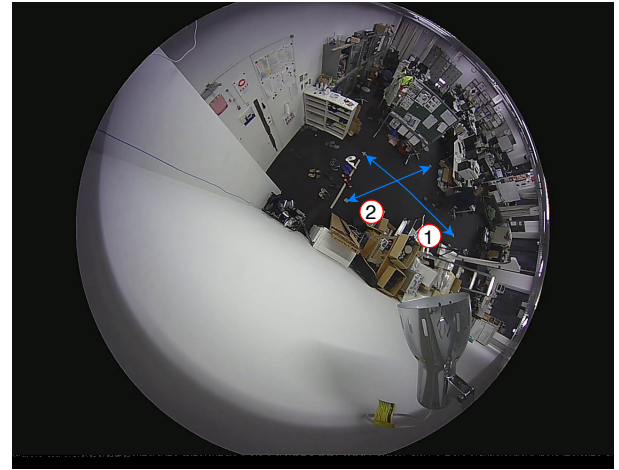


図 4 対象者の歩行ルート

表 1 クラス分類性能評価 [%]

model	timestep [frame]	
	10	20
混合ガウス分布 VAE(離散化)	92.99	94.48
混合ガウス分布 VAE	92.79	95.16
識別のみ	92.33	94.79

分類の結果を示す。各フレーム長において最良値を太字で示す。提案手法の近似事後分布から得られた特徴がクラス分類に有効であることが確認できる。また、特徴として潜在変数 z を用いるより、パラメータ μ_{gmm} を用いた方が識別に有効であることが確認できる。これは、近似事後分布の共分散 Σ_{gmm} によって潜在変数の空間内にクラス間で重複する部分が生じてしまうためだと考えられる。

表 2 特徴抽出性能評価 [%]

model	feature	timestep [frame]	
		10	20
混合ガウス分布 VAE(離散化)	ArcFace への入力	92.57	94.40
	潜在変数 z	90.14	86.59
	$q_\phi(z s, x)$ の平均	92.98	94.42
混合ガウス分布 VAE	ArcFace への入力	91.84	95.19
	潜在変数 z	83.39	82.91
	$q_\phi(z s, x)$ の平均	92.69	94.71
識別のみ	ArcFace への入力	91.84	94.63

6 おわりに

本研究では、時系列の骨格座標データによる個人識別システムに組み込むための深層生成モデルによるクラス分類を提案した。具体的には、人物ラベルをカテゴリとした混合ガウス分布を近似事後分布とした VAE を提案し、近似事後分布の推論によってクラス分類を行ったところ従来の識別モデルと同等かそれ以上の識別性能が得られたことを確認した。本研究では人物ラベルのみをカテゴリ潜在変数とした深層生成モデルを考えたが、実際の時系列骨格データには観測方向や人物行動、手荷物の有無など人物ラベルより強く影響する潜在因子が存在する。今後はそのような潜在因子を考慮した生成モデルを構築することで、より時系列骨格データを表現しやすい特徴の取得を目指す。また、本研究で利用した VAE はニューラルネットワークの重みパラメータの分布を考慮していない非ベイズのモデルである。今後は重みパラメータの分布を考慮したモデルを目指す。

謝辞

本研究の実施にあたり株式会社日立ハイテクソリューションズのご支援を頂いた。記して謝する。

参考文献

- [1] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," *International conference on Computer Vision Workshops*, pp. 1058-1064, 2009.
- [2] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," *International Conference on Biometrics*, pp 1-8, 2016.

- [3] 守脇幸佑, 村松大吾, 武村紀子, 八木康史, "追加ラベルを組み込んだ歩容特徴抽出器", 電子情報通信学会技術報告, Vol. 119, No. 214, pp. 31-35, 2019.
- [4] D.P. Kingma, S. Mohamed, D.J. Rezende, and M. Welling, "Semi-supervised Learning with Deep Generative Models," *Advances in Neural Information Processing Systems*, pp.3581-3589, 2014.
- [5] N. Dilokthanakul, P.A. Mediano, M. Garnelo, M.C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," *arXiv preprint arXiv: 1611.02648*, 2016.
- [6] M. Collier, and H. Urdiales, "Scalable Deep Un-supervised Clustering with Concrete GMVAEs," *arXiv: 1909.08994*, 2019.
- [7] C.J. Maddison, A. Mnih, and Y.W. Teh, "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables," *International Conference on Learning Representations*, 2017.
- [8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *Conference on Computer Vision and Pattern Recognition*, pp. 4690-4699, 2019.
- [9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.
- [10] 森隆寛, 外村元伸, 大住勇治, 後藤敏, 池永剛, "キュービック補間に基づく魚眼画像の高画質補正アルゴリズムおよび専用ハードウェアエンジンの提案", 画像電子学会誌, Vol. 36, No. 5, pp. 680-687, 2007.
- [11] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Real-time multi-person 2d pose estimation using part affinity fields," *Conference on Computer Vision and Pattern Recognition*, pp. 7291-7299, 2017.