深層学習と単眼カメラを用いた骨格による人物識別手法の構築

浅見直人[1]　高橋正裕[2]　橋本慧志[2]　Alessandro Moro[3]　池勇勳[1]　梅田和昇[1]

1 中央大学理工学部精密機械工学科　〒112-8551 東京都文京区春日 1-13-27

2 中央大学理工学研究科精密工学専攻　〒112-8551 東京都文京区春日 1-13-27

3 RITECS Inc. 〒190-0023 東京都立川市柴崎町 3-5-11

概要

近年、個人認証の中でも、認証者の負担とならない動画像を用いた認証手法が研究されている。本研究では人物の歩行に着目し、歩行動画像から人物を識別するシステムの構築を目指す。天井に設置された魚眼カメラで撮影された動画像から、骨格推定アルゴリズムである OpenPose によって骨格座標点を推定し、取得された時系列骨格座標データにより人物を識別する手法を提案する。さらに、潜在変数を考慮した深層生成モデルによる識別を行い、部分的には従来の識別モデルと同等の識別結果が得られたことを確認する。

キーワード：人物識別 深層学習 深層生成モデル

1. 序論

個人認証に用いられる特徴の中でも，骨格特徴や歩容特徴は他の多くの生体特徴と異なり，離れたセンサからでもデータを取得することが可能である．歩容を用いた認証には人物のシルエットに基づく特徴量が用いられることがある[1]が，服装や荷物の影響を受けてしまう問題がある．そこで，本研究では人物の骨格情報に着目する．骨格情報による人物識別に関してはいくつかの研究がされており，その有用性が示されている[2]．本研究では，一般的な単眼カメラを利用した骨格による個人識別手法の構築を目指す．

2. 提案手法

2.1　骨格特徴の抽出

まず，天井に設置された魚眼カメラから動画を取得し，YOLO[3]を用いて人物領域の推定を行う．推定された人物領域の画像を正像変換[4]し歪みを取り除いた後, OpenPose ライブラリ[5]を利用して骨格情報を取得する．

2.2　深層生成モデルの利用

一般的に，画像内での人物領域は部屋の様々な場所に存在し，人物は様々な方向に歩行するため，日常生活の歩行データを用いて識別を行う場合，骨格時系列データのクラス内分散は大きくなる．また，2 次元に射影された画像内での骨格の長さ情報を正規化することは困難である．そのような骨格時系列データを大量に集めても，データはクラス内で多峰性の分布となってしまう．しかし，多峰性のデータが様々な潜在的関係性のもと生成されていることを考慮すれば，潜在変数を用いた生成モデルによる識別が可能である．本研究では，深層生成モデルによる識別を考え，その第一段階として潜在変数が混合ガウス分布に従う変分オートエンコーダ（VAE: Variational autoencoder）を提案する．

2.2.1　混合ガウス分布による VAE

一般に VAE では，潜在変数$\boldsymbol{z}$からデータ$\boldsymbol{x}$が生成される生成モデル$p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z})$を深層学習で表現し，以下の変分下限$\mathcal{L}[\boldsymbol{\phi},\boldsymbol{\theta}]$を上げるように学習する．

$$\mathcal{L}[\phi,\theta]=\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})+\log\frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z})}\right] \quad (1)$$

ここで，$q_\phi(\boldsymbol{z}|\boldsymbol{x})$は近似事後分布であり，$\theta,\varphi$はニューラルネットワークを用いて表現されていることを意味する．変分下限のモンテカルロ推定を行う際，近似事後分布$q_\phi(\boldsymbol{z}|\boldsymbol{x})$からの勾配逆伝播可能なサンプリングと尤度関数の値が計算出来れば, 任意の分布で VAE を作成できる．本研究では，カテゴリカル分布に従う潜在変数$\boldsymbol{s}$と混合ガウス分布に従う潜在変数$\boldsymbol{z}$を持つ生成モデルからデータが生成されると仮定する．カテゴリカル分布は各クラスへ分類される確率$\boldsymbol{\pi}$をパラメータとし，1 つのみ1で他は0の要素を持つベクトル（one-of-K 表現）を生成する確率分布である．カテゴリカル分布は連続値のカテゴリが従う ExpRelaxedCategorical 分布[6]を用い，ガウス分布はクラス数分用意する．ExpGumbelSoftmax trick[6]を用いてサンプリングした実数値を持つ連続カテゴリ$\boldsymbol{s}$’ を one-of-K 表現に強制して得られた離散カテゴリ$\boldsymbol{s}$と対応するガウス分布から$\boldsymbol{z}$をサンプリングする．

2.2.2　教師データ

VAE では単純に教師データを与えることができないため, 本研究では生成モデルの事前分布$p_\theta(\boldsymbol{z})$に教師

データを与える．ここで教師データを事前分布のカテゴリカル分布のパラメータ$\boldsymbol{\pi}$として与える．N個の教師データ$\boldsymbol{s}_i\,(i=1\cdots N)$に微小なノイズを加え，温度付き Softmax 関数で正規化することで疑似的な$\boldsymbol{\pi}$を作る．各クラスの平均については，近似事後分布の推論に用いる Arcface[7]から重みベクトルをクラスに対応させて取得する．Arcface は距離学習の識別器の一つで，クラスを代表するベクトルどうしが，あるマージン以上に分離されるよう学習される．分散については単位行列を与える．結局，カテゴリ潜在変数$\boldsymbol{s}$の事後分布を推論することがクラス分類となる．

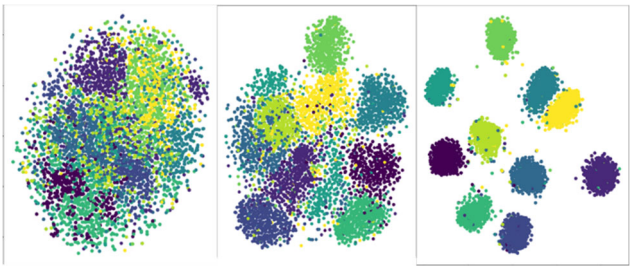
3. 実験

混合ガウス分布による VAE を用いて，手書き数字のデータセットである Mnist データセットと骨格時系列データについて識別を行った．Mnist データについて 10 クラス 60000 枚の訓練データで学習し，10000 枚のテストデータで識別を行い，提案手法と Arcface による識別のみとで比較した．10 epoch 学習させたときの識別結果の精度を表 1 に示した．Arcface による識別に比べて精度が落ちずに深層生成モデルによる識別が行えていることがわかる．図 1 は訓練データからランダムに 5000 枚画像を選び，その潜在変数を t-SNE[8]によって 2 次元へ変換し可視化したものである．各点の色は全部で 10 種類存在し，潜在変数に対応するクラスを表している．(a)はガウス分布の VAE による結果であり，(b)は事前分布の平均を$\mathbf{0}$, 分散共分散行列を単位行列に固定した混合ガウス分布による VAE (GMM-VAE)による結果であり，(c)が提案手法の VAE による結果である．提案手法による結果では，潜在変数がクラスごとに分離していることが分かる．

骨格時系列データについて，5 人の人物の歩行動画から取得した一人当たり約 880 フレームのデータで 600 epoch 学習し，一人当たり約 240 枚のバリデーションデータで評価した. 表 2 は識別結果の精度を示す. Arcface の識別のみの結果より，提案手法の結果の方が精度が低下してしまった．これは，骨格時系列データの再構成のための特徴抽出に時間がかかり，識別のための学習が適切に行えなかったためと考えられる．

4. 結論と今後の展望

本研究では，骨格情報から人物識別を行うシステムを構築するとともに，深層生成モデルによるクラス分類を提案した．今後の展望としては，人物以外例えば行動などのラベルを加えて潜在変数を増やした生成モデルでの識別を行う予定である．人物ラベルと行動ラベルのマルチラベル問題となり，深層生成モデルによって多峰性のデータ構造の再現を目指す．



(a)ガウス分布 VAE　(b)GMM-VAE（平均$\mathbf{0}$）(c)提案手法

図 1 t-SNE を用いた潜在変数の可視化

表 1 Mnist データの識別結果

	Arcface のマージン	
	0.1[rad]	0.5[rad]
提案手法	97.34[%]	97.02[%]
Arcface（**識別のみ**）	97.72[%]	97.79[%]

表 2 骨格時系列データの識別結果

	精度
提案手法	79.67[%]
Arcface（識別のみ）	89.02[%]

文　献

[1] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, “Gait Recognition Using a View Transformation Model in the Frequency Domain,” Proc. Conf. on ECCV, Lecture Notes vol 3953, pp 151-163, 2006.

[2] A. Sinha, K. Chakravarty, and B. Bhowmick, “Person Identification Using Skeleton information from Kinect,” Proc. Intl. Conf. on ACHI, pp. 101-108, 2013.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” Proc. of the IEEE Conf. on CVPR, pp. 779-788, 2016.

[4] 森隆寛，外村元伸，大住勇治，池永剛，”キュービック補間を用いた魚眼レンズ画像の高画質補正アルゴリズム”,情報科学技術フォーラム一般講演論文集, 5-1, pp.7-8, 2006.

[5] C. Zhe, T. Simon, and Y. Sheikh, “Realtime Multi-Person 2d Pose Estimation using Part Affinity Fields,” Proc. of the 2017 IEEE Conf. on CVPR, pp. 1302-1310, 2017.

[6] C.J. Maddison, A. Mnih, and Y.W. Teh, “The concrete distribution: A Continuous Relaxation of Discrete Random Variables,” Workshop on BDL, NIPS, 2016

[7] J. Deng, J Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive Angular Margin Loss for Deep Face Recognition,” Proc. of the IEEE Conf. on CVPR, pp. 4690-4699, 2019.

[8] L. Matten, G. Hinton, “Visualizing data using t-SNE,” Journal of machine learning research vol. 9, pp. 2579-2605, 2008.