同一人物間の歩容特徴交換によるデータ拡張を用いた歩容認証 第２報 生成画像の品質評価実験

○吉野 弘毅　中嶋 一斗（九州大学）　岩下 友美（NASA/Caltech JPL）　倉爪 亮（九州大学）

1. はじめに

生体情報を用いた個人識別は，暗証番号や筆跡などに比べて利便性と安全性に優れており，スマートフォンや入退室管理などでの本人確認手段として活用が進んでいる．特に，人の歩く様子を撮影した歩容映像は，対象者に特別な動作を行わせることなく非接触で取得できることから，大規模な犯罪捜査やスムーズな入退室管理に有効な生体情報として期待されている．一般的な歩容認証は，学習データ中には含まれない人物を対象として，事前に取得した評価専用のデータベース中の歩容映像と照合して人物識別する，オープンセット認識問題であるため，認証対象者に特化した特徴や識別器を直接学習することはできない．そのため歩容認証では，学習専用の人物の歩容映像を使って，学習されない認証対象者にも適応する相関性の高い特徴，つまり歩容に依存する特徴のみを抽出する特徴抽出器を作ることが目的となる．しかし，歩容映像には服装や背景等の歩容に固有でない情報（以降，共変量）が多く含まれ，認証を困難にしている．

従来多くの歩容認証手法は，背景差分手法や骨格抽出手法等の前処理によって共変量を除去してきた．例えば，歩容認証で広く利用される Gait Energy Image (GEI) [2] は，歩行者の輪郭のみを表すシルエット画像列の時間平均として得られ，周期的な輪郭の動きを表現している．しかし，背景差分手法では，例えば輪郭形状の変化に乏しい正面・背面方向の歩容映像を用いた場合，識別に重要な輪郭内部の動きが除去されてしまい，骨格抽出手法では歩容の重要な構成要素である体型情報が完全に除去されてしまうなど，前処理に伴う個人識別に有用な情報の欠落による識別率の低下が問題であった．

近年，データの潜在的に独立な属性に従って，ニューラルネットワークの内部表現を分離する分離表現学習（Disentanglement Representation Learning）が活発に研究されている．この手法は，属性ラベルを教示する必要なく解釈可能な特徴表現を学習できる．よって例えば歩容認証においては，人の外観や動きに関する特徴などを，歩容映像から End-to-End に推定・除去することも可能である．Zhang ら [3] は，共変量の一つである衣服等の外観の特徴を分離表現学習により分離・除去する手法を提案している．このように既存手法では，分離表現学習を用いることによって，前処理に依存せずに歩容映像中の共変量を除去してきたが，一方で分離性能はデータセット中のバリエーションに依存している．学習データのバリエーションを増やすことができれば，共変量を分離する精度の向上による識別精度の向上が期待できる．

この問題に対して，我々の先行研究 [1] では，分離表現学習の分離性能向上に伴う識別精度の向上を目的として，分離表現学習に基づく見えに関する共変量の特徴（以下，スタイル特徴）の除去，および敵対的生成に基づくデータ拡張による学習データの量とバリエーションを増加させる歩容認証手法を提案している．先行研究による実験では，生成画像品質の定性的評価および識別精度の定量的評価によって，生成画像の品質は高いものの，識別率の向上には課題があることが確認された．しかし，生成画像の品質の評価は生成画像の目視での確認によってのみ行われており，実際には生成画像の品質が識別精度に悪影響を及ぼしている可能性が考えられる．

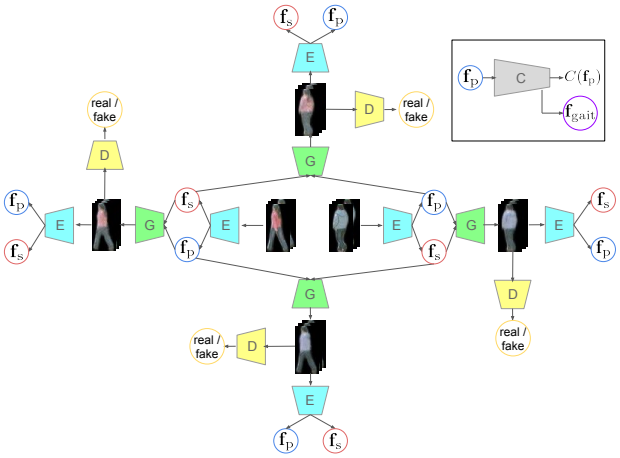

図 1　提案手法の概要図

そこで本研究では，先行研究の手法 [1] にいくつかの改善を加えた新たな手法を提案し，定量的評価および定性的評価によって生成画像品質の評価を行った．

2. 提案手法

提案手法の概要を図 1 に示す．本手法では，Zhang ら [3] と同様に，Mask R-CNN [6] によって人物領域のみを抽出した人物歩行映像 $\boldsymbol{I} = \{I_1, ..., I_T\}$ を入力データとする．また，提案モデルは先行研究 [1] と同様に，各フレームの歩行者画像 I_t から姿勢特徴および衣服や持ち物等のスタイル特徴を出力するエンコーダ E，エンコーダ出力から歩行者画像を生成する生成器 G，実画像と生成画像を識別する識別器 D，全フレームの姿勢特徴から人物 ID を識別する分類器 C から構成される．本章では，これらのモデルを同時学習する損失関数と人物認証のパイプライン，および損失関数における先行研究 [1] からの改善点について述べる．

表 1　比較手法

手法	敵対的損失 (5)	生成画像の識別損失 (6)	特徴一貫性損失 (3),(4)	AdaIN [9]
ベースライン [3]				
提案手法	✓	✓	✓	✓
w/o AdaIN (先行研究 [1])	✓	✓	✓	
w/o 一貫性損失	✓	✓		✓
w/o 生成画像の識別損失	✓		✓	✓
w/o 敵対的損失		✓	✓	✓

2.1　歩行者画像の姿勢特徴とスタイル特徴への分離

Zhang ら [3] に従い，同一人物で状態の異なる二つの映像 I^{c_1}, I^{c_2} から得られた姿勢特徴列 $\boldsymbol{f}_\mathrm{p}$ とスタイル特徴列 $\boldsymbol{f}_\mathrm{s}$ から，次の生成画像間損失 $\mathcal{L}_\mathrm{recon}$ と姿勢特徴間損失 $\mathcal{L}_\text{pose-sim}$ を計算する．

$$\mathcal{L}_\mathrm{recon} = \left\| G(\boldsymbol{f}_\mathrm{s}^{(t_1,c)}, \boldsymbol{f}_\mathrm{p}^{(t_2,c)}) - I_{t_2}^{c} \right\|_1 \tag{1}$$

$$\mathcal{L}_\text{pose-sim} = \left\| \frac{1}{n_1}\sum_{t=1}^{n_1} \boldsymbol{f}_\mathrm{p}^{(t,c_1)} - \frac{1}{n_2}\sum_{t=1}^{n_2} \boldsymbol{f}_\mathrm{p}^{(t,c_2)} \right\|_2^2 \tag{2}$$

さらに，生成した画像を再度エンコーダ E に入力し，エンコードする前後での両特徴の差分からエンコーダ E の一貫性を保証する $\mathcal{L}_\text{p-consis}$，$\mathcal{L}_\text{s-consis}$ を計算する．ここで，$E_\mathrm{p}(*)$ はエンコーダ E の出力のうち姿勢特徴を，$E_\mathrm{s}(*)$ はスタイル特徴を指す．

$$\mathcal{L}_\text{p-consis} = \left\| \boldsymbol{f}_\mathrm{p} - E_\mathrm{p}(G(\boldsymbol{f}_\mathrm{s}, \boldsymbol{f}_\mathrm{p})) \right\|_1 \tag{3}$$

$$\mathcal{L}_\text{s-consis} = \left\| \boldsymbol{f}_\mathrm{s} - E_\mathrm{s}(G(\boldsymbol{f}_\mathrm{s}, \boldsymbol{f}_\mathrm{p})) \right\|_1 \tag{4}$$

我々の先行研究 [1] では，(3), (4) 式は，平均絶対誤差（MAE）ではなく平均二乗誤差（MSE）を用いていたが，予備実験の結果，MAE の方が結果が優れていたため，本研究においては MAE を採用している．

2.2　姿勢特徴とスタイル特徴による歩行者画像の生成

異なる映像の姿勢特徴とスタイル特徴を入力し，生成器 G から仮想歩行映像を生成する．生成した各フレームは識別器 D に入力し，次の敵対的損失 $\mathcal{L}_\mathrm{adv}$ を計算する．ただし，敵対的損失の定義には RaLSGAN [7] を用いた．

$$\begin{aligned}\mathcal{L}_\mathrm{adv} = \sum_{i,j\in\{c_1,c_2\}} \Big(&\left(D(\boldsymbol{I}^j) - \mathbb{E}\left[D\left(G(\boldsymbol{f}_\mathrm{s}^i, \boldsymbol{f}_\mathrm{p}^j)\right)\right] - 1\right)^2 \\ &+ \left(D\left(G(\boldsymbol{f}_\mathrm{s}^i, \boldsymbol{f}_\mathrm{p}^j)\right) - \mathbb{E}\left[D(\boldsymbol{I}^j)\right] + 1\right)^2 \Big)\end{aligned} \tag{5}$$

2.3　姿勢特徴列による人物識別

得られた姿勢特徴列を分類器 C に入力し，人物 ID の推定確率を出力する．分類器 C は，3 層の LSTM と 1 層の全結合層から構成される．人物 ID の推定確率から次の交差エントロピーを計算する．ここで，$\boldsymbol{y}$ は one-hot 表現された正解ラベルを示す．

$$\mathcal{L}_\mathrm{id} = \frac{1}{\sum_{t=1}^{n}\omega_t}\sum_{t=1}^{n} -\omega_t \boldsymbol{y}^\mathrm{T} \log\left(C(\boldsymbol{f}_\mathrm{p})\right) \tag{6}$$

学習は次のマルチタスク損失 $\mathcal{L}$ を最小化する．

$$\begin{aligned}\mathcal{L} = &\lambda_\mathrm{recon}\mathcal{L}_\mathrm{recon} + \lambda_\text{pose-sim}\mathcal{L}_\text{pose-sim} \\ &+ \lambda_\text{p-consis}\mathcal{L}_\text{p-consis} + \lambda_\text{s-consis}\mathcal{L}_\text{s-consis} \\ &+ \lambda_\mathrm{adv}\mathcal{L}_\mathrm{adv} + \lambda_\mathrm{id}\mathcal{L}_\mathrm{id}\end{aligned} \tag{7}$$

3.　実験

実験では，CASIA-B [5] データセットを用いて，生成画像品質の定量的評価および定性的評価を行った．CASIA-B [5] データセットは，基準となる NM，バッグをもった BG，ロングコートを着た CL の三つの画像セットから構成される．学習には，全被験者 124 名のうち前半の 74 名分の歩行映像を用いた．

表 1 に，実験に用いた手法を示す．提案手法を構成する各要素の有効性を詳しく検証するために，提案手法とベースラインの他に，提案手法から各要素を一つずつ取り除いた手法についても実験を行った．ただし，本実験で用いたベースラインは，Zhang [3] らの論文を元に筆者らが実装したものである．AdaIN [9] とは，スタイル変換タスクに用いられる正規化処理のことで，我々の先行研究 [1] では使用していなかったが，本研究では生成画像の品質を向上させるために生成器に用いた．なお，学習の際にはそれぞれの歩行映像から連続する 20 フレームをランダムに切り取ったものを用い，評価の際には各映像の全フレームを用いた．

3.1　生成画像品質の定量的評価

本実験では，データセットの全画像中からランダムに 5,000 枚選択し，ランダムな組み合わせで生成した画像に対して，品質を評価する指標によって生成画像品質の定量的評価を行った．評価指標として，生成画像品質の測定のために広く用いられている Fréchet Inception Distance (FID) [10] を採用した．FID は，実画像の分

表 2　各手法における FID の値

手法	FID
ベースライン [3]	169.8
提案手法	<u>118.7</u>
w/o AdaIN (先行研究 [1])	130.5
w/o 敵対的損失	205.7
w/o 生成画像の識別損失	**<u>106.3</u>**
w/o 一貫性損失	138.6

布と生成画像の分布間の距離を計算しており，この値が小さいほど生成画像の品質が高い．各手法においてFID を測定した結果を表 2 に示す．最も小さい値を太字かつ下線付きに，二番目に小さい値を下線付きにしている．

表 2 より，提案手法は，ベースラインよりも生成画像の品質が高くなっており，先行研究よりもさらに品質が向上していることがわかる．また，各手法における値を比較すると，生成画像を識別学習に用いない時のみ，つまり生成画像をデータ拡張に使用しなかった場合のみ品質が向上しているものの，生成画像の識別損失以外の要素を提案手法から取り除いた時に品質が低下している．特に，敵対的損失を取り除いた時には，ベースラインよりも大きくなっており，提案手法において，敵対的損失が生成画像品質の向上に大きく貢献していることがわかる．

3.2 生成画像品質の定性的評価

同一人物に対し，歩行方向や服装が異なる二つの映像から，姿勢特徴およびスタイル特徴を交換して生成した結果を図 2 に示す．1, 2 行目は，それぞれ生成画像における外観特徴，姿勢特徴の元となった画像である．左端 3 列は撮影角度が同じ場合の画像であり，右端 3 列は 1, 2 行目で撮影角度が異なっている．そのため，分離表現学習による分離性能が高い場合，2 行目の姿勢をしていながら 1 行目の服装をしている画像が生成されると期待される．

図 2 を見ると，ベースライン [3] では，姿勢の大局的な特徴は再現できているものの，輪郭と色がぼやけている．また，右端 3 列の画像を見ると，脚と腕の長さや角度などの細部が，姿勢特徴の元画像のものと異なっている．このことから，ベースラインは，姿勢特徴およびスタイル特徴の元となった画像の撮影角度が同じ場合には，色や輪郭がぼやけるものの特徴の分離・抽出が可能であるが，撮影角度が異なる場合には，分離性能が低下することがわかる．それに対して，提案手法は，輪郭や色がぼやけることなく，両特徴の元となった画像の撮影角度が異なる場合においても，精度よく特徴の分離・抽出ができている．また，先行研究 [1] と比較しても，提案手法の方が手足などの細部まで表現できているため，分離性能が向上していることがわかる．

提案手法から各要素を取り除いた手法の結果を見ると，敵対的損失を取り除くとベースラインと似た生成画像になっていることから，敵対的損失が分離性能の向上に大きく貢献していることがわかる．また，生成画像品質の定量的評価では最良の結果となった，生成画像を識別学習に用いない設定では，手足の先など細部が再現できていないだけでなく，服装に 2 行の要素が混入してしまっていることから，分離性能は提案手法よりも低下していると考えられる．

4. まとめと今後の予定

本研究では，分離表現学習に基づいて歩容映像から姿勢特徴と服装や持ち物などの外観特徴に分離し，同一人物の異なる歩容映像間で両特徴を交換して新たな画像を生成するデータ拡張を用いた歩容認証手法を提案した．CASIA-B [5] を用いた実験では，生成画像品質の定量的評価および定性的評価を行い，提案手法がベースライン [3] や先行研究 [1] と比較して，生成画像の品質を向上させること，また分離表現学習における特徴分離の性能を高めることをそれぞれ確認した．

今後は，提案手法による識別精度の検証を行う予定である．また，本提案手法は同一人物間での特徴交換に限定していたが，異なる人物間での特徴交換に基づくデータ拡張を用いた手法に発展させていく．

謝　辞　本研究の一部は JSPS 科研費 JP20H00230 の助成を受けたものである．

参　考　文　献

[1] 吉野 弘毅, 中嶋 一斗, 岩下 友美, 倉爪 亮. “同一人物間の歩容特徴交換によるデータ拡張を用いた歩容認証” 画像の認識理解シンポジウム (MIRU), IS3-3-28, 2020.

[2] Ju Han and Bir Bhanu. “Individual Recognition Using Gait Energy Image.” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(2):316–322, 2005.

[3] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. “Gait Recognition via Disentangled Representation Learning.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4710–4719, 2019.

[4] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. “Gait Recognition via Semi-supervised Disentangled Representation Learning to Identity and Covariate Features.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13309–13319, 2020.

[5] Shiqi Yu, Daoliang Tan, and Tieniu Tan. “A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition.” In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, vol. 4, pp. 441–444, 2006.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask R-CNN.” In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

[7] Alexia Jolicoeur-Martineau. “The relativistic discriminator: a key element missing from standard GAN.” In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[8] Yang Feng, Yuncheng Li, and Jiebo Luo. “Learning effective gait features using LSTM.” In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, pp. 325–330, 2016.

[9] Xun Huang, and Serge Belongie. “Arbitrary style transfer in real-time with adaptive instance normalization.” In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. “Gans trained by a two time-scale update rule converge to a local nash equilibrium.” In *Proceedings of Neural Information Processing System (NIPS)*, pp. 6629–6640, 2017.

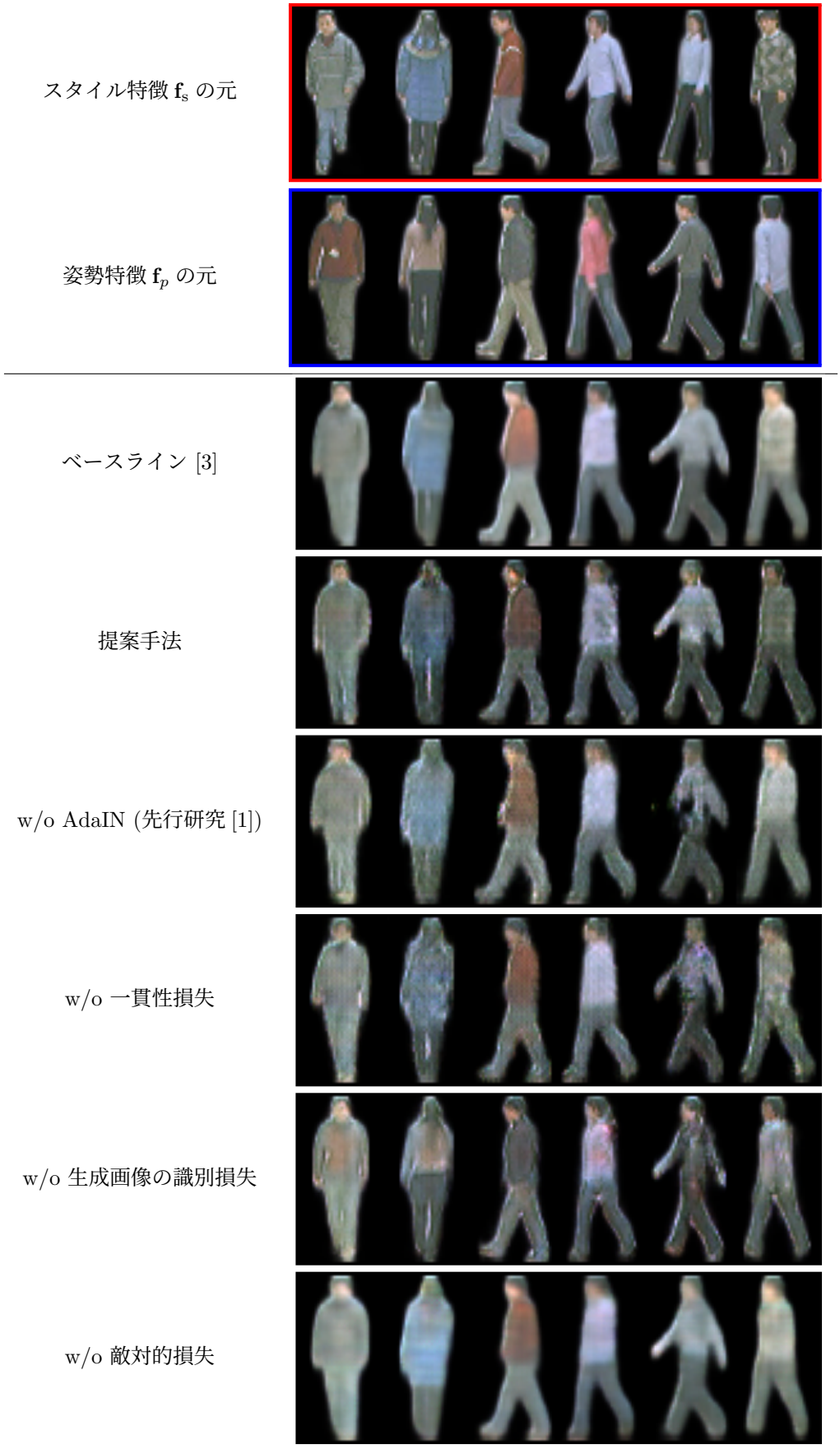

図 2　各手法における同一人物間の姿勢特徴交換による生成画像