Person Re-identificationの実用に向けたDNNモデルの特性評価

吉田　龍人[1]・大久保　順一[2]・藤井　純一郎[1]・高森　秀司[1]・天方　匡純[1]

[1]正会員　八千代エンジニヤリング(株)　技術創発研究所（〒111-8648 東京都台東区浅草橋 5-20-8）
E-mail: ry-yoshida@yachiyo-eng.co.jp (Corresponding Author)

[2]非会員　八千代エンジニヤリング(株)　技術創発研究所（〒111-8648 東京都台東区浅草橋 5-20-8）

動画を対象とする人流解析は狭域の解析に適した手法である．これに対してPerson Re-identificationが実現すると解析範囲を広域化できるため，手法の実用性がより一層向上する．Re-identificationではDNNモデルの出力する特徴ベクトルによって同一人物を照合するのが一般的だが，入力画像の変化が画像間の類似度に与える影響は十分に解明されていない．これを受けて，本研究では同一人物を規定の条件下で撮影し，同一人物照合に影響する要素を評価するためのデータセットを作成する．さらに本データセットをRe-identificationモデルで推論し，画像間の類似度を算出する．この結果に基づき入力画像の変化が類似度に与える影響を評価し，解析に用いたモデルの特性を明らかにする．

Key Words: *Pedestrian Tracking，Deep Learning，Re-identification，Metric learning*

1. はじめに

ストリートデザインガイドライン[1]などの活用により，まちなか空間を「自動車中心の通行空間」から「人中心の活動空間」に転換する取組が始まっている．これに伴い，人流計測技術の開発ニーズが高まり，人流計測手法の1つである動画解析にも注目が集まっている．一般に動画の人流解析は，物体検出と物体追跡を要素技術とする[2]．検出と追跡によって各人物の軌跡が取得でき，その軌跡を加工することで，高森[3]らが示すように多様な人流関連データを定量化できる．動画解析はカメラの画角内，つまり狭域での人流計測に適した手法であり，検出対象の追加・変更や属性情報の取得など現場ごとに柔軟なシステム構築ができる点に強みがある．これらに加えて，もし動画から複数のカメラにまたがって写る同一人物を照合できれば，カメラ間ODといった広域な人流データが取得可能となり，技術の実用性が向上する．

複数カメラ間での同一人物の照合はPerson Re-identification[4]と呼ばれる画像認識タスクで手法が検討されている．Re-identificationでは距離学習で構築したDNNモデルを活用することが一般的となっており，モデルが出力する特徴ベクトルの類似度で同一人物を照合する．同一人物照合を動画による人流解析システムに追加する場合，人物検出領域をDNNモデルで特徴ベクトル化し，カメラ間で類似度の高いベクトルを探索することが想定されるが，カメラの設置状況や人の向き，ポーズ，背景など撮影側で制御不能な入力画像の変化が照合精度を低下させる懸念がある．

既往のRe-identificationに関する論文は，新たなデータセットの構築[5]や損失関数（Loss）[6]，学習時のデータサンプリングの改善[7]などに焦点を当てたものが大半であり，それらはデータセット全体での照合精度を測る指標によって評価されていた．そのため個別の画像に生じた変化が画像間の類似度に与える影響は十分に評価されていない．過去にゲーム画像でその影響を評価した事例[8]はあるが，実際の画像によって評価した事例はなく，Re-identificationの現場適用時の留意点は明らかでない．

本研究ではRe-identificationタスクにおいて重要な役割を持つDNNモデルの特性評価を行うために，任意の条件で同一人物の画像を撮影し，DNNを用いた同一人物照合において影響を及ぼす要素を評価するためのデータセットを作成する．さらに本データセットに対して，学習データや学習方法の異なる2つのモデルによって画像間の類似度を算出する．この結果より入力画像の変化が類似度に与える影響を定量評価するとともに，解析に用いたモデルについて学習データセットやアルゴリズムの観点も踏まえてその特性を明らかにする．さらに各モデルを現場適用する上での留意点を整理する．

2. 先行研究

既存の学習データセットを用いて深層学習モデルを構築する場合，モデルアーキテクチャと学習アルゴリズム，損失関数の選定がモデル性能を左右すると一般に考えられる．このうちモデルアーキテクチャは，Tan[9]らが示すように特徴量抽出器となる DNN モデルのパラメータ数が増加するにつれて精度が向上するスケーリング則があるため，要求精度や計算量に応じて必然的に決定される．これらの観点を踏まえて，本章では Re-identification における学習アルゴリズム，損失関数について先行研究を整理する．

一般に Re-identification のモデルは距離学習によって構築される [4]．距離学習とは，基準データとなるクエリーに対して同じクラスに属するデータを正例，異なるクラスに属するデータを負例と見なしたときに，特徴空間上でクエリー・正例間の距離を最小化し，クエリー・負例間の距離を最大化させるようモデルを構築するタスクである．距離学習は教師画像のラベルの有無によって教師ありと自己教師ありに大別される．ここで教師あり距離学習は Pairwise Loss を用いるアプローチと SCE (Softmax Cross Entropy) Loss を用いるアプローチに区分される．

Pairwise Loss はクエリーと正例・負例の組合せから Loss を算出する手法で，クエリー・正例・負例の 3 つを入力とする Triplet Loss[6]など多様な手法がある．これらの手法は Mining[7]と呼ばれるデータサンプリングの方策によって学習の収束性やモデルの性能が変化する点に難がある．その一方で特徴量抽出器の出力する特徴ベクトルだけで Loss が算出できるため，分類器が学習に不要で学習パラメータが少量で済むといったメリットがある．

SCE Loss を用いるアプローチは通常の分類 AI を構築する手順と同様に，特徴量抽出器と分類器から成るモデルを学習させる手法である．SCE Loss を用いる手法は，Mining 方法を考慮する必要がなく，Pairwise Loss に比べて収束性が高いといった特徴がある [10]．SCE Loss の中でも，(1)式に示す ArcFace[11]が代表的な Loss として知られている．

$$\mathrm{L} = -\log \frac{e^{s(\cos(m_1\theta_{yi}+m_2)-m_3)}}{e^{s(\cos(m_1\theta_{yi}+m_2)-m_3)} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}} \quad (1)$$

ここで m_1，m_2，m_3 のそれぞれが SphereFace[12]，ArcFace，CosFace[13]で示された Angular Margin と呼ばれる学習時のペナルティ項を表している．通常の SCE Loss に Angular Margin を導入したことで，特徴空間上でクラス間の距離が確保されることが明らかにされており，距離学習タスクでの精度向上を果たした．既往研究 [14]では，シングルカメラでの人流解析結果を用いて，フレーム毎に切り出した各人物の画像を ArcFace で学習することで一定の実用性のあるモデルが構築できることが示されている．

自己教師あり学習では，あるデータに異なる Augmentation を加えたものの組合せを正例，それ以外のデータを負例として Pairwise Loss を用いて学習を実施する．Person Re-identification での自己教師あり距離学習では，Fu[5]らが LUPerson データセットを Moco v2[15]ベースのアプローチで学習する手法が高精度となることを示した．LUPerson データセットとは，400 万を超える画像で構成されるラベルなし画像データセットで，都市空間を撮影した膨大な動画から人の領域を物体検出によって切り出して作成したものである．よって吉田 [14]らの手法と異なり，検出された矩形すべてに異なる ID が付与されている．

Fu らが用いた Moco v2 の前身の手法である Moco[16]は動的な辞書構造を用いて自己教師あり距離学習を行う手法である．Moco v2 では SimCLR[17]に用いられた全結合層および Augmentation を Moco に適用することで精度向上を果たした．Moco v2 では(2)式に示すクエリーq に対する正例を k+，負例を k-とした InfoNCE Loss[18]によって学習を行う．

$$\mathrm{L} = -\log \frac{exp(q\cdot k^{+}/\tau)}{\exp(q\cdot k^{+}/\tau) + \sum_{k^-} exp(q\cdot k^{-}/\tau)} \quad (2)$$

InfoNCE Loss は Pairwise Loss の派生手法であり，1 つのアンカーに対して 1 つの正例と膨大な負例を組み合わせた入力データから Loss を算出する．InfoNCE Loss は，距離学習データセットのアンカー・正例の組合せが少なく，アンカー・負例の組合せは膨大であるという特性に合わせて，アンカーと負例の組合せをより集中的に学習させる仕組みとなっている．

3. 実験方法

本実験では入力画像の変化が DNN モデルの出力，ひいては画像間の類似度に与える影響を評価するとともに，解析に使用した DNN モデルの特性を明らかにする．具体的には同一人物を任意の条件下で撮影し，それぞれの条件下で撮影された画像間で類似度を算出することで，その影響を定量評価する．

使用するモデルは，(1) 既往研究 [14]に倣ってシングルカメラでの人流解析結果を ArcFace で学習した ResNet50[19]ベースのモデルと，(2) LUPerson データセットを Moco v2 によって学習した ResNet50 ベースのモデルである（以降，(1)を ArcFace モデル，(2)を Moco モデルと称する）．ArcFace モデルは，①教師データを撮影したカメラと同じカメラの画像で実験を行う，②画像の撮影条件が比較的実験条件に近いといった観点から，本実験にて高い性能を発揮する可能性があるため採用した．Moco モデル

は，Fu らの論文中に未学習のデータセットに対する高い適合性が示されており，本実験の画像でも高い性能を発揮することが期待されたため採用した．

Moco モデルは Fu らが GitHub にて公開する ResNet50 をベースとする学習済みモデル [20]を用いた．ArcFace モデルはGitHub にて公開されている顔認識タスクに特化したライブラリの insightface[21]を利用して，新たにモデルを構築した．それぞれ既存の実装を用いたため，ResNet50 をベースとしたアーキテクチャでありながら構造に違いがある．**表-1** にそれぞれのモデルの構造の違いを示す．Moco モデル出力から類似度を算出する際は，計算量削減などの観点から Global Average Pooling の処理を行い，2048 次元の特徴ベクトルに変換したものを用いた．

図-1 から**図-5** に入力画像の変化の影響を評価するために撮影した画像を示す．撮影条件はカメラからの水平距離を 4m，カメラ高さを 112cm，カメラ俯角を 0°，人の向きを背面向き（カメラ正対を 0°として 180°向き），立ち姿を直立，服装を白いスウェットシャツ，灰色のパンツ，ヘルメットとしたケースを標準条件とし，比較項目だけ条件を変えるようにして画像の撮影を行った．**図-1** では水平距離を，**図-2** では人の向きを，**図-3** では人の服装を，**図-4** では背景を，**図-5** ではカメラ高度を標準条件から変更した．向きの影響を様々なパターンで評価するためリュックサックを背負った状態の画像も別途撮影した．背景変更時の比較条件の 1 つであるモニターとは，単に黒色のモニターの裏面を画面奥側に設置した状態を指し，パネルなし/ありのときのグレーの背景と明確な違いが生じることを狙いとしている．

標準条件のうち，カメラ高度は概ね人が画像中央に位置するように定め，水平距離を変化させた際に人とカメラ間の俯角が変化しないよう考慮して設定した．撮影時は背景の影響を排除するため，背景変更時以外は人の奥側にパネルを物理的に設置した．さらに被撮影者の姿勢の揺らぎなどが算出される類似度に影響することを考慮して条件ごとに 5 枚ずつ画像を撮影した．カメラ高度を変更した際は，人とレンズ間の斜距離を 4m に統一するよう水平距離を変更するとともに，人が画像中央に存在するようにカメラ俯角を調節した．

Re-identification モデルには物体検出によって人の範囲を切り出した画像を入力することから，上述の手順で撮影した画像から人の範囲を目視で切り出した．さらにモデルの入力サイズに合わせて，元の画像の縦横比を維持するよう幅 128px，高さ 384px にリサイズした．リサイズによって幅 128px，高さ 384px を下回った箇所は R:128, G:128，B:128 の値で補った．

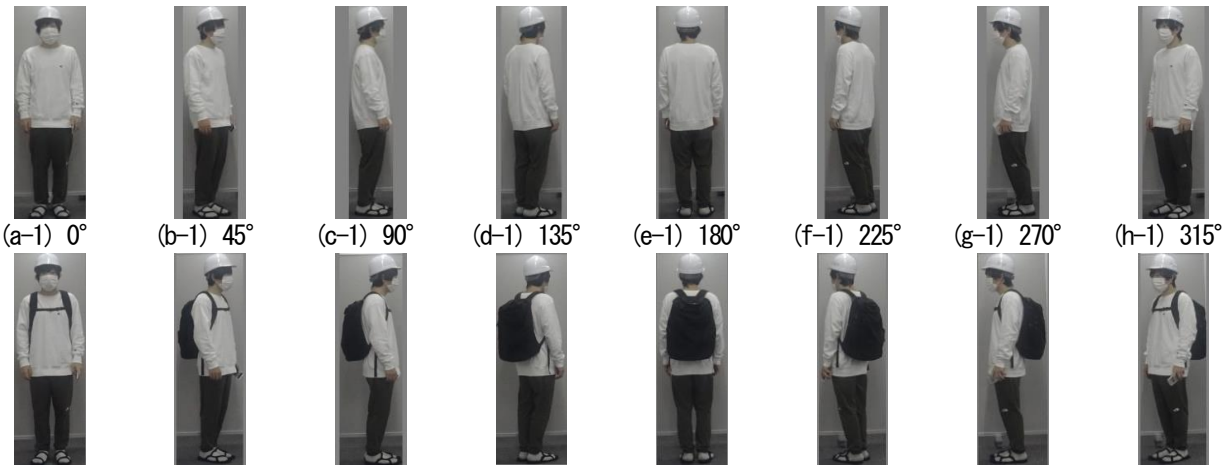
図-1 から**図-5** のそれぞれの画像を，各 Re-identification モデルによって特徴ベクトル化し，撮影条件間でコサイン類似度を算出した．この結果に基づき入力画像の変化が類似度に与える影響を評価した．なお類似度は異なる撮影条件下の画像だけでなく，同一撮影条件下の画像同士でも取得した．異なる撮影条件下で撮影した画像は 5 枚×5 枚=25 組の類似度の平均値によって評価し，同一撮影条件下で撮影した画像は ${}_5C_2 = 10$組の画像で算出した類似度の平均値によって評価した．

表-1 モデル構造

	ArcFace		Moco	
	Output (H, W, CH)	Layer	Output (H, W, CH)	Layer
Conv1	$384, 128, 64$	$3 \times 3, 64, \mathrm{Stride}1$	$96, 32, 64$	$7 \times 7, 64, \mathrm{Stride}2$ $3 \times 3 \mathrm{Max\ Pool}, \mathrm{Stride}2$
Conv2	$192, 64, 64$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$96, 32, 256$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Conv3	$96, 32, 128$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 4$	$48, 16, 512$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$
Conv4	$48, 16, 256$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 13$	$24, 8, 1024$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 3$
Conv5	$24, 8, 512$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$24, 8, 2048$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
Output	512	512-d FC	$24, 8, 2048$	-
その他相違点	• Skip Connection に$1 \times 1, -, \mathrm{Stride}2$を適用 • 活性化関数に PReLU[22]を使用 • Conv2~5 は BN, Conv, BN, Activation, Conv, BN という構成 ※Conv1 は Conv, BN, Activation		• Skip Connection に$1 \times 1, -, \mathrm{Stride}1$を適用 • 活性化関数に ReLU を使用 • Conv2~5 は Conv, BN, Conv, BN, Conv, BN, Activation という構成 ※Conv1 は Conv, BN, Activation	

(a) 2m　(b) 4m　(c) 6m　(d) 8m　(e) 10m　(f) 12m　(g) 14m

図-1 距離の影響評価用画像



(a-1) 0°　(b-1) 45°　(c-1) 90°　(d-1) 135°　(e-1) 180°　(f-1) 225°　(g-1) 270°　(h-1) 315°

(a-2) 0°　(b-2) 45°　(c-2) 90°　(d-2) 135°　(e-2) 180°　(f-2) 225°　(g-2) 270°　(h-2) 315°

図-2 向きの影響評価用画像（1：リュックなし，2：リュックあり）

4. 実験結果

図-6 から図-11 に本実験にて撮影した画像に対して ArcFace モデルおよび Moco モデルでコサイン類似度を算出した結果を示す．いずれもヒートマップ化した相関行列で撮影条件間の類似度を表現している．対角行列は同じ撮影条件の画像同士で算出した類似度で，非対角行列は異なる条件の画像間で算出した類似度である．例えば図-6 (a)の 1 行 2 列目および 2 行 1 列目の要素は，ArcFace モデルが出力する特徴ベクトルによって算出した図-1 (a)に例示する 2m 画像 5 枚と図-1 (b)に例示する 4m 画像 5 枚で構成される 25 組の類似度を平均した値である．これらはコサイン類似度の対称性により同一の値が示されている．

(1) 距離の影響評価結果

図-6 は図-1 に示す距離に変化を与えた画像間で類似度を算出した結果である．距離の差が開くほど類似度が低下する傾向が確認された．ArcFace は Moco に比べて距離の変化に頑健であった．

(2) 向きの影響評価結果

図-7 および図-8 は図-2 に示すカメラから 4m の地点で 1 回転した人物の類似度を示している．図-7 はリュックサックを背負っていない状態の画像での類似度であり，

(a) 白服 1　(b) 白服 2　(c) 灰色服　(d) 白服 1 ヘルメットなし

図-3 服装の影響評価用画像

(a) パネルなし　(b) モニター　(c) パネルあり

図-4 背景の影響評価用画像

(a) 112cm　(b) 172cm　(c) 232cm

図-5 カメラ高度の影響評価用画像

図-8 はリュックサックを背負った状態の画像の類似度である．ArcFace ではリュックサックを背負った状態の類似度が顕著に低下した．両モデルにおいて 90° に対する 270° 画像のように水平反転を加えたような関係性を持つ画像間で比較的高い類似度が示された．

(3)　服装変化の影響評価結果

図-9 は図-3 に示す異なる服装をした画像間で類似度を算出した結果を示している．いずれのモデルにおいてもグレーのスウェットを着用した(c)にてその他 3 ケースの画像との類似度が低下した．ArcFace モデルではヘルメットを脱帽した(d)でも他の画像との類似度が低下した．

(4)　背景変化の影響評価結果

図-10 は図-4 に示す異なる背景の画像間で類似度を算出した結果を示している．Moco モデルでは微小な類似度低下にとどまったが，ArcFace モデルでは顕著な類似度低下が確認された．パネルは撮影フロアと同系色であったことから，設置の有無による影響はやや小さかったが，モニターを画面奥側に設置した際は顕著に類似度が低下した．

(5)　カメラ高度変化の影響評価結果

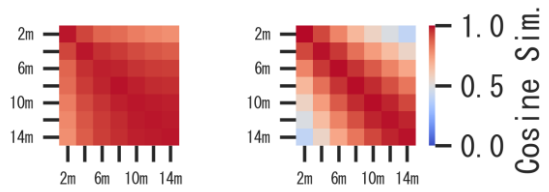
図-11 は図-5 に示すカメラ高度の異なる画像間で類似度を算出した結果を示している．図-5 の各画像は定性的には類似しているように感じるが，高度の変化に伴い類似度が下がる傾向が確認された．背景変更時と同様に ArcFace モデルで顕著な類似度低下が確認された．

5.　考察

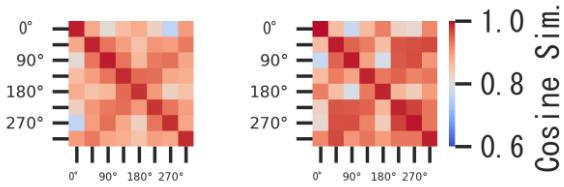
本章では実験結果を踏まえて，(1) 学習データとモデル性能，(2) 学習アルゴリズムとモデル性能，(3) Re-identification の現場適用といった観点から考察を行う．

(1)　学習データとモデル性能に関する考察

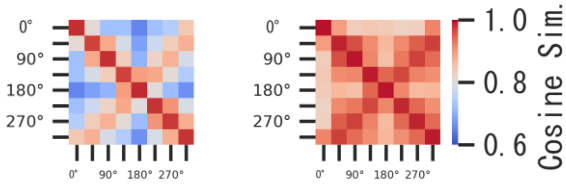
Re-identification で類似度を算出する画像間には体型，服装，髪型，姿勢，向きなどの被撮影者の身体的変化と，人とカメラ間の距離，画角，背景，明るさ，カメラの歪みなどの撮影条件変化が発生する．このうち Re-identification の高精度化のためには身体的変化のうち，体型，服装，髪型などの変化情報のみを特徴ベクトルに抽出し，その他情報はできる限り排除する必要がある．これを本実験の条件に当てはめると，服装の変更時には画像間の類似度が低下し，その他変化が発生した場合は類似度が 1 に近い値を示すことが理想であると言える．



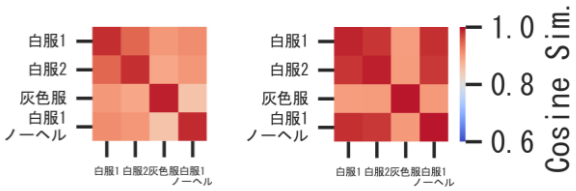
(a) ArcFace　(b) Moco
図-6 距離の影響評価結果



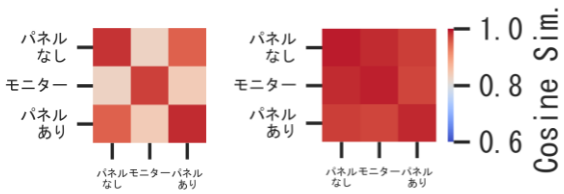
(a) ArcFace　(b) Moco
図-7 回転の影響評価結果（リュックなし）



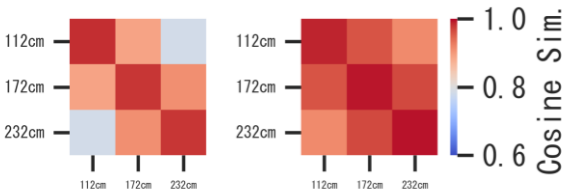
(a) ArcFace　(b) Moco
図-8 回転の影響評価結果（リュックあり）



(a) ArcFace　(b) Moco
図-9 服装変化の影響評価結果



(a) ArcFace　(b) Moco
図-10 背景変化の影響評価結果



(a) ArcFace　(b) Moco
図-11 カメラ高度変化の影響評価結果

しかし実際は，類似度が低下すべき服装の変更時よりも，その他条件の変更時の方がより類似度が低下するケースが確認され，Re-identification を行う上で排除すべき特徴を DNN モデルが多数抽出していることが明らかとなった．この原因は単にモデルの性能に依るものであり，モデル性能の向上が必要であると考えた．

モデルの性能はモデルアーキテクチャによっても変化するが，学習データセットに依るところが大きいと判断する．本実験で用いた ArcFace モデルはシングルカメラで撮影した動画，つまり一定の条件で撮影した動画を人流解析し，得られた物体検出結果の矩形を基に作成したラベル付きデータセットで学習を実施した．そのため，ArcFace モデルのデータには，①同一人物データの距離のバリエーションが多い，②同一人物データの向きのバリエーションが少ない，③撮影高度のバリエーションがないといった特徴がある．これを踏まえて画像間の類似度を算出した結果を見ると，①距離のバリエーションが多いため距離の変化に強い，②向きのバリエーションに少ないため向きの変化に弱い，③撮影高度のバリエーションがないため高度の変化に弱いといったデータセットの特性を反映した結果であることが明らかとなった．

ArcFace モデルに対して Moco モデルは膨大なカメラで撮影した動画，つまり様々な条件で撮影した動画から検出された人の画像を基に作成したラベルなしデータによって学習が実施されている．そのため Moco モデルのデータセットには，①同一人物データの距離のバリエーションがない，②同一人物データの向きのバリエーションがない，③撮影高度のバリエーションが多いといった特徴がある．これを踏まえて画像間の類似度を算出した結果を見ると，Moco モデルでは②向きのバリエーションがないにも関わらず向きの変化に強いという結果も確認されたが，①距離のバリエーションがないため距離の変化に弱い，③撮影高度にバリエーションが多いため高度の変化に強いというデータセットの特性を反映する傾向にある結果が確認された．以上よりデータセットにバリエーションを与えることで，モデルが変化に頑健になることが想定された．Moco モデルにてバリエーションのない向きの変化に頑健になった理由として，深層学習でのモデルのパラメータがデータセットに最適化するように決定される特性を踏まえると，LUPerson データセットにおいて人の向きが同一性の判定に用いる特徴として重要ではなかったからだと推察されるが，その根拠は定かではない．

距離変化に対する頑健性確保という観点では，クラス内で様々な距離のバリエーションの画像を追加できない自己教師あり学習には技術的限界点があると考える．教師の作成コストが低いという観点から自己教師あり学習の実用性も多分に存在するが，より高精度なモデルを構築するためには，コストを許容してでも教師あり学習のアプローチを取るべきであると考える．なお生成モデルなどを活用して，距離変化を再現するように画質を向上・低下させる前処理がもし適用できれば，自己教師あり学習でも距離変化に対する精度向上の可能性は期待される．

(2)　学習アルゴリズムとモデル性能に関する考察

Re-Identificaton モデルの特性は単に学習に用いたデータセットの違いだけでなく，学習手法の違い，特に Hard Negative Samples の学習方策にも起因すると考える．

Xuan[7]らは Triplet Loss による学習実施時の Mining 方法として，アンカーと高い類似性を示す負例（以降，Hard Negative Samples と称する）を集中的にサンプリングしたときの方がより局所領域の特徴量を抽出するモデルが形成されることを明らかにした．局所領域の特徴量を抽出するモデルは着目領域に同一性があれば高い類似度を示すため，着目領域以外の変化には鈍感であることが想定される．

ArcFace をはじめとする SCE Loss のアプローチの場合，局所データによって Loss を算出する Pairwise Loss のアプローチと異なり，データ空間全体で Loss を算出する．したがって学習データセットの内容から直接的に Hard Negative Samples の学習方策が定まると予測される．つまりデータセットのクラス間の類似度が低い場合は大域，高い場合は局所特徴を抽出するモデルが形成されると想定する．実験結果からは ArcFace モデルが画像全体の変化に敏感に反応したことから，大域の特徴を満遍なく抽出していると予測した．これを検証するために ArcFace モデルの類似度算出時の着目点を可視化することは今後の課題とする．

本研究で用いた Moco モデルは Negative Samples をランダムに取得するアルゴリズムで学習されているため，Mining 方法自体は Hard Negative Samples を重視していない．しかし Fu らは Moco モデル学習時の Contrastive Loss における温度パラメータτを，Hard Negative Samples を重視するとされる[23]0.07 という低い値に設定していた．したがって LUPerson モデルが局所的な特徴量を抽出するモデルとなっていた可能性がある．全結合層によって特徴ベクトルを形成する ArcFace と異なり，Global Average Pooling によって特徴ベクトルを形成する Moco モデルの場合，類似度評価時の着目領域を明らかにする Stylianou[24]らの手法を適用することができる．**図-12** はヘルメット着脱前後の画像において Moco モデルの着目点を可視化した例である．図は赤色領域ほど類似度算出に影響のあった領域であることを意味する．この結果より Moco モデルが足から胸にかけての比較的広範囲の特徴量を重視していることが確認されたが，本画像ペアにて着目すべき頭

部の特徴量は重視されておらず，ヘルメット着脱によって類似度が低下しない原因が確認された．同様に，**図-10** (b)で背景変化時に Moco モデルであまり類似度が低減しなかった理由も，背景領域が類似度評価の際に重視されなかったためだと考える．なお着目点は画像によっても変化するため，常に足や胸が着目されるわけではないことに留意する必要がある．

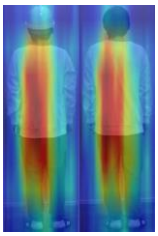


図-12 着目領域の可視化結果

ArcFace モデルの学習では Margin の与え方もモデルの特性に寄与することが予測される．Margin は学習データセットにおけるクラス内類似度を小さくし，クラス間類似度を大きくさせるように作用するものであるが，この Margin によって形成されたモデルのパラメータが入力データの微小な変化に強力に発火し，データごとに特異な特徴空間に写像させる効果をもたらした可能性がある．実際に ArcFace モデルは画像に微小な変化が起こるだけで類似度が大きく減少する傾向にあった．この考察については Margin を変えて複数構築した ArcFace モデルの性能比較実験によって，解明を行うことを今後の課題とする．

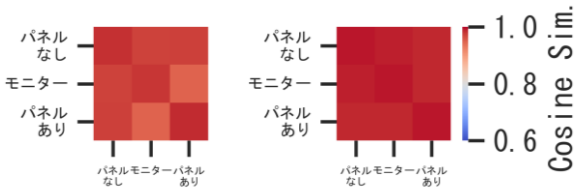
(3) Re-identification の現場適用に関する考察

ArcFace モデルはリュックサックを背負った際の回転時や背景の変化時など Re-identification のタスク上で高い類似性を示すことが望まれる画像間でも類似度が低下しており，入力画像の変化に応じて特異な特徴ベクトルを形成することを確認した．Moco モデルでは高い類似度を示すことが望まれる画像間において正しく高い類似度を示す例がいくつか見られたが，ヘルメットの着脱時や白スウェット-白シャツ間など，人の目には異なる特徴量を持つ画像間でも類似度が低下しない例が見られた．これらの結果を踏まえて，ArcFace モデルは異なる画像間で高い類似度を示し難い分，類似度が高い場合は精度よく同一人物と判定できる Precision 優先型のモデルであり，Moco モデルはその反対の Recall 優先型のモデルであると判断した．入力画像が一定の条件下で撮影できる場合は ArcFace モデルが極めて高い精度を示すことが想定されるが，各人物が不定の行動をする屋外環境での人流解析では Moco モデルの方が向き等の変化にも頑健であることから高い実用性を持つと判断した．

今後さらに高精度な Re-identification を実現するには，単にクラス内データに多様性のあるデータセットでのモデルの学習だけでなく，画像の前処理の実施といった対策が想定される．例えば**図-13** は背景の異なる**図-4** の画像に対して，zero-shot segmentation が可能な SAM[25]によって，人物領域だけを自動的に抽出して類似度を算出した結果である．背景をマスクすることで，**図-10** に比べてクラス間の類似度が 1 に近い値を示すことが確認された．このように変化が生じた箇所に対して，その影響を排除するような前処理を実施することで，さらなる精度向上が見込まれる．

他の対策例として，水平反転を活用して入力する人の向きを0～180°の範囲に制約するなどの処理が検討される．**図-7** および**図-8** では学習時に実施した水平反転の前処理の効果から，画像の反転の関係にある画像間で比較的高い類似度が示された．この結果より反転した人物の同一性を識別するためにモデルのパラメータの一部が割かれているとも想定されるため，画像反転を使って入力時の向きを疑似的に0～180°に制約することで，向き以外の変化に頑健なモデルが構築できると予測される．

今後は屋外空間を対象とする Re-identification タスクにおいて画像上で起こりうる変化を整理し，それぞれに即した前処理方法の検討およびその評価に取り組むことを課題とする．



(a) ArcFace　　(b) Moco

図-13 背景マスクの効果検証結果

6. 終わりに

本研究ではマルチカメラでの同一人物照合の精度向上に向けて，既定の条件下で撮影した画像を用いて，異なるアプローチで学習した 2 つの Re-identification モデルの特性評価を行った．モデルには(1) テスト環境に類似した撮影条件のデータセットを ArcFace によって学習したモデルと，(2) 様々な動画に写った人物の画像で構成される LUPerson データセットを Moco v2 のアルゴリズムで学習した学習済み公開モデルの 2 つを用いた．ArcFace モデルは画像全体の微小変化を反映した特徴ベクトルを出力する傾向にあり，異なる画像間で高い類似度を示し難い分，類似度が高い場合は精度よく同一人物と判定できる Precision 優先型のモデルであることを確認した．これ

に対して Moco モデルは Recall 優先型のモデルであることを確認した.

撮影条件ごとに類似度を算出した結果より，人とカメラ間の距離やカメラ高度などといった変化が同一人物の類似度を大きく低下させることが確認され，Re-identification の精度向上のためにはそれら変化に頑健なモデル構築や，変化の影響を低減させる画像の前処理が必要であることが明らかとなった.

参考文献

1) 国土交通省：ストリートデザインガイドライン ver2.0, 2021.
2) Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., and Kim, T.: Multiple object tracking: A literature review, Artificial Intelligence, Vol.293, 2021.
3) 高森秀司，吉田龍人，堀井大輔，菊池恵和，大久保順一：AIによる人流解析結果を介した歩道空間の特性把握の可能性に関する研究，AI・データサイエンス論文集 4巻2号，pp. 121-127，2023.
4) Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. H.: Deep Learning for Person Re-identification: A Survey and Outlook, PAMI2022, pp.2872-2893, 2022.
5) Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., and Chen, D.: Unsupervised Pre-training for Person Re-identification, CVPR, pp. 14750-14759, 2021.
6) Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., and Wu, Y.: Learning Fine-grained Image Similarity with Deep Ranking, CVPR2014, pp. 1386-1393, 2014.
7) Xuan, H., Stylianou, A., Liu, X., and Pless, R.: Hard Negative Examples are Hard, but Useful, ECCV2020, pp 126-142, 2020.
8) Xiang, S., You, G., Li, L., Guan, M., Liu, T., Qian, D., and Fu, Y.: Rethinking Illumination for Person Re-identification: A Unified View, CVPRW2022, pp. 4731-4739, 2022.
9) Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, PMLR2019, 2019.
10) Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P., and Ayed, I. B.: A Unifying Mutual Information View of Metric Learning: Cross-Entropy vs. Pairwise Losses, ECCV 2020, pp. 548-564, 2020.
11) Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., and Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition, CVPR2019, pp. 4690-4699, 2019.
12) Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L.: SphereFace: Deep Hypersphere Embedding for Face Recognition, CVPR2017, pp. 212-220, 2017.
13) Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W.: CosFace: Large Margin Cosine Loss for Deep Face Recognition, CVPR2018, pp. 5265-5274, 2018.
14) 吉田龍人，菊池恵和，堀井大輔，大久保順一，高森秀司：深層距離学習のための MOT 解析を使った教師画像自動生成，第 37 回人工知能学会全国大会，2023.
15) Chen, X., Fan, H., Girshick, R., and He, K.: Improved Baselines with Momentum Contrastive Learning, arXiv: 2003.04297, 2019.
16) He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R.: Momentum Contrast for Unsupervised Visual Representation Learning, CVPR2020, pp. 9729-9738, 2020.
17) Chen, T., Kornblith, S., Norouzi, M., and Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations, PMLR2020, 119: 1597-1607, 2020.
18) Oord, A., Li, Y., and Vinyals, O.: Representation learning with contrastive predictive coding, arXiv:1807.03748, 2018.
19) He, K., Zhang, X., Ren, S., and Sun., J.: Deep Residual Learning for Image Recognition, CVPR2016, pp. 770-778, 2016.
20) Fu, D.: LUPerson, (2021), https://github.com/DengpanFu/LUPerson.
21) Deep Insight: insightface, (2021), https://github.com/deepinsight/insightface.
22) He, K., Zhang, X., Ren, S., and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, ICCV2015, pp.1026-1034, 2015.
23) Wang, F., Liu, H.: Understanding the Behaviour of Contrastive Loss, CVPR2021, pp. 2495-2504, 2021.
24) Stylianou, A., Souvenir, R., and Pless, R.: Visualizing Deep Similarity Networks, WACV2019, 2019.
25) Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., and Girshick, R.: Segment Anything, arXiv: 2304. 02643, 2023.

(Received June 30, 2023)
(Accepted August 31, 2023)

Evaluation of metric learning model for Person Re-identification

Ryuto YOSHIDA, Junichi OKUBO, Junichiro FUJII and Shuji TAKAMORI

The analysis of pedestrian tracking in videos is suitable for narrow scope measurements. On the other hand, Re-identification enables the expansion of the scope, thereby enhancing the method's practicality. While Re-identification commonly matches the same individual based on similarity using feature vectors generated by DNN models, the full impact of changes in input images on the similarity has not been fully understood. In this study, a dataset is created by capturing images of the same individuals under specific conditions to evaluate the factors influencing Re-identification. Furthermore, similarity between images in this dataset is measured using a Re-identification model. Based on these results, the influence of changes in input images on similarity is evaluated, and the characteristics of the model are clarified.