(42) マルチカメラでの同一人物照合を目的としたDNNモデルの出力ベクトル分析

吉田　龍人[1]・大久保　順一[2]・藤井　純一郎[1]・高森　秀司[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

[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今後の活用が期待される手法であるが，解析対象範囲の狭さに欠点がある．これに対して複数カメラに写った同一人物の照合が実現すれば，解析範囲が広域化し，より情報量の多いデータが取得可能となる．同一人物の照合は DNN モデルの出力する特徴量ベクトルを用いることが一般化しているが，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．

本研究では任意の距離に立つあらゆる方向を向いた同一人物を固定カメラで撮影し，画像認識での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に対するDNN モデルの出力に対して種々の分析を実施し，距離や向きといった変化をもたらす要素が類似度に与える影響を明らかにする．

Key Words: *pedestrian tracking，deep learning，Re-Identification，metric learning*

1. はじめに

まちづくり，交通，防災などの地域課題を可視化する上で，より情報量の多い人流データを獲得することが課題となっている．中でもビデオカメラは人流データを取得するツールの 1 つとして活用が期待されている．一般にビデオによる人流解析は，物体検出と物体追跡を要素技術とする．人流解析により動画上での各人物の軌跡が取得でき，さらにその軌跡を加工・分析することで通行者数などのデータが獲得できる．ビデオ解析は拡張性にも強みがあり，検出対象物体の追加・変更や属性情報の取得など現場に合わせて柔軟にシステムが構築できる．一方で，ビデオによる人流解析には広範囲の OD データを取得できないといった欠点がある．これに対して，複数のカメラに共通して写った人物を照合することができれば，カメラ間 OD データなど広域な人流データも収集可能となり，ツール活用のより一層の広がりが期待される．

複数のカメラに写った同一人物の照合は Person Re-Identification[1)]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量ベクトルのユークリッド距離やコサイン類似度などによって同一人物を照合する．同一人物照合を物体検出・物体追跡で構成される人流解析システムに追加する場合，人物検出領域を距離学習 DNN モデルでベクトル化し，カメラ間で類似度の高いベクトルを探索することが容易に想像されるが，人・カメラ間の距離や人の向き，ポーズ，背景，周辺の照度など制御不能な入力画像の変化が特徴量ベクトルに変化をもたらし，マッチング精度の低下を引き起こすことが懸念される．既往の研究にて Re-Identification のためのモデル構築手法は多数検討されているが，入力画像の変化が画像間の類似度に与える影響など Re-Identification の現場適用に向けた研究は十分に実施されていない．過去にはゲーム画像を用いて分析を行った事例 [2)]が示されたが，実際の画像によって評価を実施した事例はない．

本研究では固定カメラによって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種々の DNN モデルで特徴量ベクトルを取得し，特徴量ベクトルの分析を通じて入力画像の変化が類似度算出結果に与える影響を明らかにする．

2. 先行研究

一般に Re-Identification のモデルは距離学習によって構築される [1]．距離学習とは特徴量空間上でアンカーと正例の距離を最小化，負例の距離を最大化させるモデルを構築するタスクである．距離学習はさらに教師ありと自己教師ありに区分される．

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大別される．Pairwise Loss は正例・負例のペアを作って学習を行う手法で，特徴量抽出器のみで学習ができるため，学習パラメータが少量で済む．SCE Loss は通常の画像分類 AI を構築する手順と同様に，特徴量抽出器および分類器から成るモデルを学習させる手法で，Pairwise Loss に比べて収束性が高い．Pairwise Loss と SCE Loss は概念的に一見異なるが，クラス内距離の最小化およびクラス間距離の最大化を行い，特徴量ベクトルと正解ラベルの相互情報量を最大化をする点で同じであることが理論的に示されており [3]，本質的に似た学習方法だと考えられる．したがって実用上はデータセットの ID 数が多い場合は分類器が不要な Pairwise Loss，それ以外の場合は SCE Loss の利用が推奨される．SCE Loss の中でも，(1)式に示す ArcFace[4]が代表的な手法となっている．

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

ここで m_1，m_2，m_3 のそれぞれが 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量空間上でクラス間のマージン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．

Person Re-Identification での自己教師あり距離学習では LUPerson データセットを Moco v2 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が Fu [5]らによって示されている．Moco v2 [6]とは前身の手法である Moco に SimCLR [7]に用いられた MLP 層および Augmentation を導入した距離学習手法である．Moco v2 では query に対する正例を k+，負例を k-とした(2)式によって学習を行う．

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

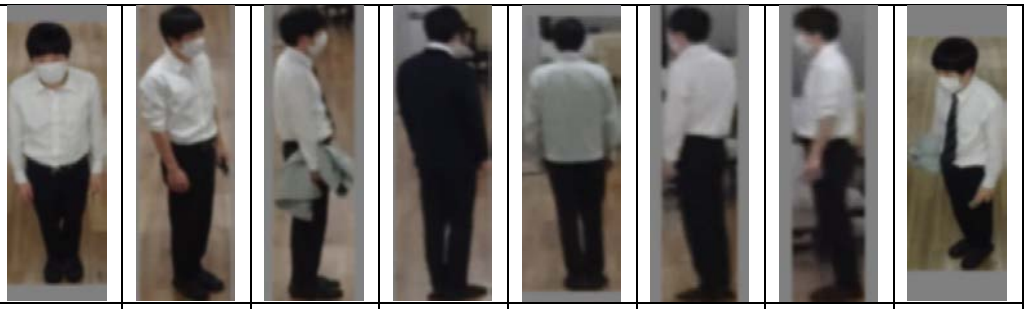
Fu らは精度向上のために Moco v2 のデフォルトの学習手順に対して，画像の Augmentation 手法の変更や温度パラメータτのチューニング実施した．

この他に超大規模データで学習された基盤モデルの特徴量抽出器は，入力データをより解釈性の高い特徴量空間に写像することが期待される．億単位の画像とテキストラベルのペアで学習した CLIP[8]は Zero-shot Transfer を可能としており，CLIP Image Encoder が Re-Identification の特徴量抽出器として実用性を持つ可能性がある．

3. 実験方法

入力画像の変化が照合精度に与える影響を明らかにするため，2.4m程度の高さに設置したカメラによって同一人物を**図-1** に示す様々なパターンで撮影した．カメラと人の水平距離は 2m 刻みで 2m～14m の計 7 パターンを撮影した．カメラに対する向きはカメラ正対を 0° として，45° 刻みで計 8 パターンを撮影した．更に，見た目の違いによる影響を評価するため，**表-1** に示す計 5 パターンを撮影した．立ち姿は全て直立である．入力画像は人の写る箇所のみを切り出し，元の画像の縦横比を維持するよう 128×384px にリサイズした．リサイズはアスペクト比を保つようにし，不足する箇所はグレーで補った．なお実験ではぼかしを実施していない元の画像を用いた．

7×8×5=280 枚の各画像は ArcFace で学習した ResNet50，LUPerson で学習した ResNet50，CLIP で学習した ViT-L/14 の 3 つのモデルで推論して特徴量ベクトル化した．なお ArcFace は既往の研究の手法 [9]によって自前のデータで学習したモデルを使用し，LUPerson および CLIP は論文の著者らが GitHub にて公開する学習済みモデルを用いた．



距離 (m)	2	4	6	8	10	12	14	2
向き (°)	0	45	90	135	180	225	270	315
服装	①	②	③	④	⑤	①	②	③

図-1 実験画像例 ※Gaussian Blur によるぼかしを実施

4. 実験結果

(1) 実験１：UMAP による次元削減

入出力の関係を定性的に評価するため，UMAP [10]によって全画像の特徴量ベクトルを 2 次元に削減した．**図-2**, **図-3**, **図-4** は ArcFace，LUPerson，CLIP での結果である．プロット上に対象画像を描画し，人とカメラ間の水平距離に応じて画像に枠線を付けた．

すべてに共通して，分布が水平距離に強く相関し，距離に合わせて各点が隣接した．一方で向きとの相関は弱く，様々な向きの画像が無相関に分布した．

Person Re-Identification 用モデルとして人の画像だけで学習した ArcFace・LUPerson の両モデルの結果は概ね類似した．形成されるクラスター数などに若干の違いはあるが，ジャケット画像が独立したクラスターを形成し，定性的に違いの明らかな作業着画像がワイシャツ画像と特徴量空間上で混同する点で同じであった．

CLIP の出力ベクトルの次元削減結果は，ジャケット着用画像も他の服装と特徴量空間上で混同し，距離学習で構築したモデルに比べて明確な境界が表れないといった特徴が見られた．次元削減結果が入力画像の特性に従って分布しているという観点から，CLIP の Image Encoder にも特徴量抽出器としての十分な有用性が確認されたが，服装の明らかに異なる画像間のマージンが確保されにくいという観点から，Person Re-Identification には距離学習で構築したモデルが適していることを確認した．

(2) 実験２：ArcFace による類似度計測

ArcFace で学習した ResNet50 によって，一部抜粋した画像の類似度を算出した．**図-5**(ⅰ)は 8m の 0° の画像を基準として同じ服装の 8m 画像同士で類似度を算出した結果であり，同一人物をただ回転させた場合の類似度の変化を表している．**図-5**(ⅱ)は(ⅰ)と同様に 8m の画像を基準に 6m の画像との類似度を算出した結果である．さらに**図-5**(ⅲ)は 8m の標準画像を基準として 8m の画像の類似度を算出した結果で，標準以外の行は標準の服装との類似度を算出したことを意味している．

以上より向きの変化時よりも距離の変化時により類似度が低下することが確認された．なお唯一，**図-5** の全ケースで発生した(c)の 90 度画像での極端な類似度低下は，右手に持った手荷物によるオクルージョンが要因だと推測される．

見た目の変化による類似度変化は服装の違いに依存し，ワイシャツ-ジャケット画像間では類似度は大きく低下するが，ワイシャツ-作業着画像間では若干の類似度低下程度にとどまった．

表-1 見た目のパターン

条件	概要
(a)	ワイシャツ（ネクタイ無）※標準とする
(b)	ワイシャツ（ネクタイ無，袖まくり）
(c)	ワイシャツ（ネクタイ有）&手荷物（右手持ち）
(d)	ジャケット（ネクタイ無）
(e)	作業着（ネクタイ無）

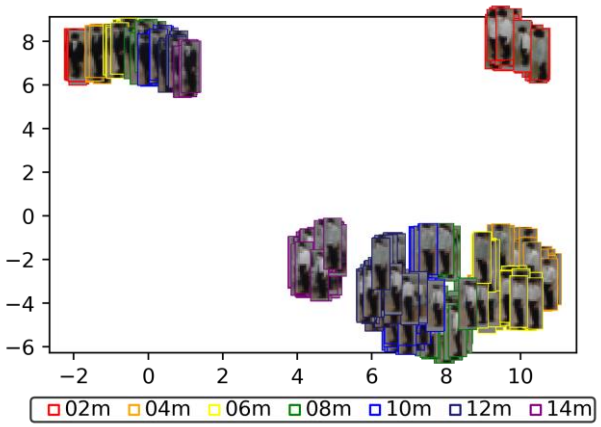


図-2 ArcFace 次元削減結果

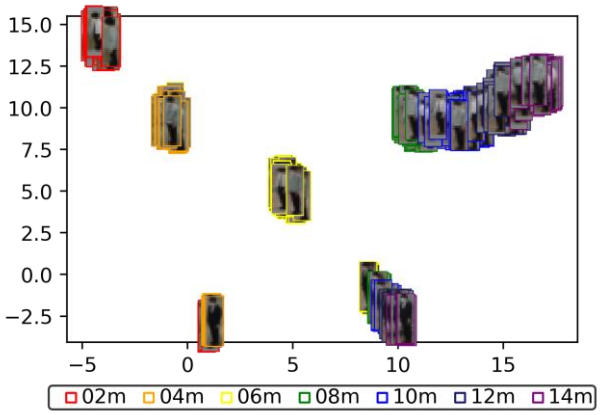


図-3 LUPerson 次元削減結果

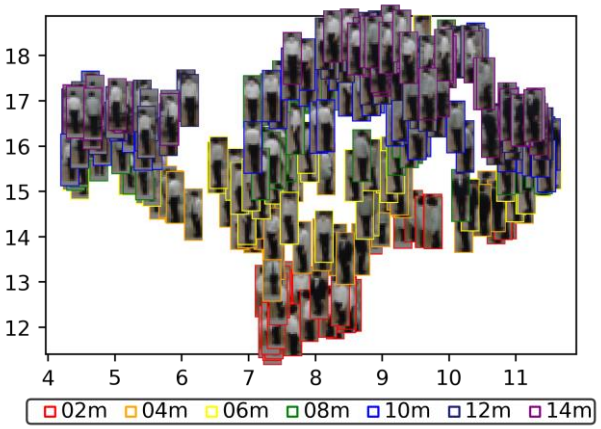


図-4 CLIP 次元削減結果

5. 考察

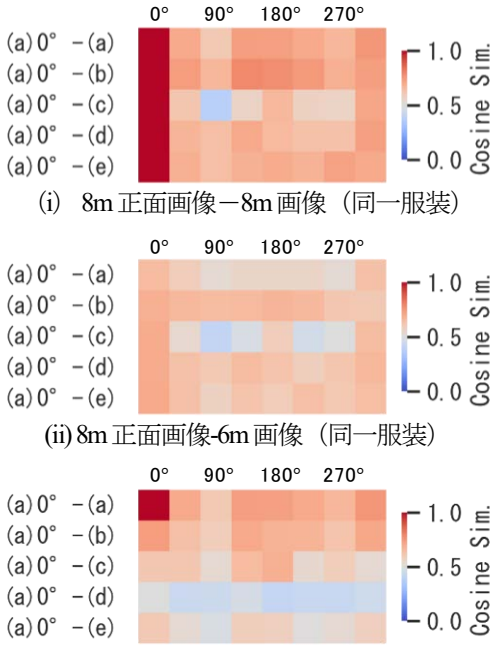
類似度に強く寄与した水平距離の変化は，背景，人とカメラ間の俯角，画質の変化などの要素を内包している．このうち背景に着目すると，2m，4m 画像で共通して写る背景が木目調の床のみであったのに対して，ArcFace, LUPerson の各次元削減結果にてそれらが全体から分離したクラスターを形成したことから，背景が類似度に強く寄与したとは想定しがたい．これに対して俯角はカメラ近傍ほど変化が大きくなるため，2m 画像で異なるクラスターが形成された次元削減結果にも合致し，類似度に強く寄与する要素である可能性が示唆された．画質の観点では被写体との距離とノイズ強度に相関性があることが定性的に画像から確認されており，ノイズが ResNet50 に対して敵対的攻撃に似た現象を引き起こし，距離と類似度に強力な相関をもたらしたと考えた．なお**図-4** より敵対的攻撃に頑健である ViT モデルでも距離と類似度に強い相関が確認されたため，ノイズ以外の影響である可能性も現段階では棄却できない．

いずれのモデルも向きの変化には寛容であり，人の回転による大域的なエッジ情報の変化は顕著な類似度低下を誘引しなかった．これは**図-5** の(iii)より(a)正面画像と(a)画像および(a)正面画像と(b)画像の類似度に大差がなかった点からも裏付けられる．ジャケット着用画像で極端に類似度が低下した点も踏まえると，いずれのモデルも画像の大域的エッジ情報ではなく大域的色情報に着目していると判断した．

6. おわりに

マルチカメラでの同一人物照合に向けて，既定の条件下で同一人物の撮影を行うとともに，その画像を用いて DNN の出力する特徴量ベクトルの分析を行った．本実験データで比較した項目のうち，服装の大域的な色情報以外に人とカメラ間の距離が画像間の類似度を大きく低下させる要素であることが明らかとなった．ただし，単なる距離の変化による影響とカメラと人を結ぶ俯角の変化の影響を分離して評価できていないため，これらを明らかにすることは今後の課題である．今回の結果を踏まえ，マルチカメラの同一人物照合の実現には，画像内の深度情報を組み込み，人とカメラ間の位置関係を考慮した上で同一性の判定を行う必要があると考えた．

本研究手法のように単に人の見た目の類似度だけ同一性を判定する場合，似た服装の人物を誤マッチングするリスクがあるため，骨格推定および骨格推定を活用した歩容推定技術などによって多角的な観点から同一人物を行う必要があると考える．



(i) 8m 正面画像－8m 画像（同一服装）

(ii) 8m 正面画像-6m 画像（同一服装）

(iii) 8m 正面画像-8m 画像（非同一服装）

図-5 類似度算出結果

参考文献

1) M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. H. Hoi.: Deep Learning for Person Re-identification: A Survey and Outlook, arXiv: 2001.04193, 2021.
2) S. Xiang, G. You, L. Li, M. Guan, T. Liu, D. Qian; Y. Fu: Rethinking Illumination for Person Re-Identification: A Unified View, CVPRW2022, 2022.
3) M. Boudiaf, J. Rony, I. M. Ziko, E. Granger, M. Pedersoli, P. Piantanida ans I. B. Ayed: A Unifying Mutual Information View of Metric Learning: Cross-Entropy vs. Pairwise Losses, ECCV 2020, 2020.
4) J. Deng, J. Guo, J.Yang, N. Xue, I.Kotsia and S. Zafeiriou.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition, CVPR2019, 2019.
5) D. Fu; D. Chen; J. Bao; Hao Yang; Lu Yuan; Lei Zhang; Houqiang Li; Dong Chen: Unsupervised Pre-training for Person Re-identification,CVPR2021, 2021.
6) X. Chen, H. Fan, R. Girshick and K. He: Improved Baselines with Momentum Contrastive Learning, arXiv: 2003.04297, 2019.
7) T. Chen, S. Kornblith, M. Norouzi, G. Hinton: A Simple Framework for Contrastive Learning of Visual Representations, PMLR 119:1597-1607, 2020.
8) A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever: Learning Transferable Visual Models From Natural Language Supervision, PMLR 139:8748-8763, 2021.
9) 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.
10) L. McInnes, J. Healy and J. Melville: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv: 1802.03426, 2018.