

撮影条件を考慮した同一人物識別のための基礎的分析

藤井 純一郎† 吉田 龍人† 吉村 直也‡ 芝 慎太郎‡ 孔 全‡ 伊藤 天詞†

†八千代エンジニアリング株式会社 ‡ウーブン・バイ・トヨタ株式会社

E-mail: jn-fujii@yachiyo-eng.co.jp

1 はじめに

監視カメラ映像を用いた施設利用状況の把握やセキュリティ向上のためには、別カメラ・別フレームに映った同一人物を識別することが求められる。コンピュータービジョン分野において、このタスクは人物再同定(Person Re-identification; ReID)と呼ばれる。一般に ReID モデルは、同一人物の画像ペアから抽出される特徴量のコサイン類似度を最大化するように学習される。しかし実運用環境では、カメラと被写体の距離やアングル等の撮影条件の変動によって同一人物特徴量の類似度が低下し[1]、識別精度が低下することが問題となる。

この問題に対処するためには、同一人物識別の際に ReID 特徴量のコサイン類似度のみで判定するのではなく、撮影条件を考慮することが有効と考えられるが、撮影条件にはカメラ-人物間距離、カメラ高さ、人物の映る向きといった複数の要素が存在し、それらが複合的に絡み合う。これに対して、撮影条件の組み合わせを網羅した大規模なデータセットがあれば、機械学習により撮影条件を考慮した同一人物識別ができる可能性がある。実動画のベンチマークデータセットでは、実空間のカメラ設置上の制約やアノテーションの作業量の問題により、撮影条件のバリエーションが限られるが、近年は多様な撮影条件を含む合成データセットが公開されている。

そこで、本研究では撮影条件を考慮した同一人物識別に向けた初期検討として、多様な撮影条件を含む合成データセットを用いて、撮影条件を説明変数に加えることで同一/非同一人物の分類精度が向上するか確認する。さらに、撮影条件を変化させることで分類精度および類似度分布にどのように影響するか実験し、今後の撮影条件に対する頑健性向上に向けた考察を行う。

2 関連研究

ReID タスクに関するモデル、データセットを示す。

2.1 ReID モデル

ReID は、異なる時間帯や異なるカメラに映る同一人物を識別する技術である。深層学習を用いた ReID は、Li ら[2]によって提案されて以来、高精度なモデル[3, 4]だけでなく実用のための効率的で軽量なモデルも提案されている[5, 6, 7]。Bag of Tricks (BoT) [5]は、複雑な構造を用いずに単純な ResNet50 [8] バックボーンに対して学習率やデータ拡張の工夫、正規化層の追加等の学習テクニックを組み合わせることで、推論時の計算効率を維持した手法である。Omni-Scale Network (OSNet) [6]は、様々なスケールの受容野を扱う構造を持ち、一般的な畳み込みを空間方向畳み込みとチャネル方向畳み込みに分解して扱うことでパラメータ数を削減した手法である。Zhou ら [7] は、OSNet に Instance Normalization (IN) を組み込むことで異なるカメラ間の撮影条件の差異に対処した OSNet-AIN を提案した。

2.2 ReID データセット

ReID タスクのデータセットは、大きく静止画と動画のデータセットに分かれる。静止画ベースとしては Market-1501[9]、DukeMTMC-reID[10]などが知られている。これらは多くの研究でベンチマークとして使われてきたが、カメラの台数やシーンのバリエーションが限られる。動画ベースのデータセットも、当初は小規模なデータでバリエーションが限られたが、MARS[11]や LS-VID[12]が登場し、深層学習モデルの学習に十分な規模を持つベンチマークデータセットとして使われている。ただしカメラ台数は MARS6 台、LS-VID15 台と、撮影条件を評価するにはカメラ台数のバリエーションが不足する。

実動画でより多くの撮影条件で同一人物を撮影、ラベリングするには膨大な作業量が必要となるが、合成データであれば作業量の問題を回避できる。2024 年、2025 年の AI City Challenge[13]向けに公開された Physical AI Smart Spaces Dataset[14]は、

実動画に比べて多様な撮影条件で同一人物のラベルが付与されている。例えば 2025 年のデータセットでは、292 人の人物を 504 台のカメラで撮影した 23 シーン 42 時間分の動画が公開されている。

3 実験

本研究は ReID モデル自体の精度向上を目的としたものではなく、ReID 特徴量に撮影条件を加味することによる同一人物識別精度向上の可能性を実験により確認するものである。そのため、同一人物画像を類似度順に上位 K 枚を検索し、その中に正解が含まれていたかを評価する ReID タスクではなく、撮影条件を加味する最もシンプルな問題設定として、2枚の人物画像ペアに対して同一/非同一人物の2値分類タスクを考える。ReID 特徴量の類似度のみで分類した場合をベースラインとして、以下の3つの撮影条件を説明変数に加えた場合に精度が向上するか確認した(実験1)。

- カメラ-人物間距離
- カメラ高さ
- 人物の映る向き

また、各撮影条件のバリエーションを絞り込んだ実験を行い、撮影条件の違いがどの程度精度に影響を与えているか評価した(実験2)。

3.1 データセット

データセットには Physical AI Smart Spaces Dataset の中から、1シーンあたりのカメラ台数が多く撮影条件がより多様な MTMC_Tracking_2025 のデータセットを用いた(図 1)。ただし全データを用いるとデータ量が膨大となるため、GroundTruth のラベルに基づき、以下の手順でサンプリングを行った。

- 90 フレームごとの BoundingBox を抽出
- BoundingBox の座標が画像領域外にかかるものは全身が見切れているとみなし除外
- BoundingBox 領域の DepthMap を取得し、Depth の最大値と平均値の差が一定の値を越えたものはオクルージョンとみなし除外
- 過小な BoundingBox を除外

サンプリングした結果、ID 数(個人数)50 人、画像枚数 19,416 枚の画像が抽出された(図 2)。ここから更に各実験では撮影条件で絞り込んだデータを用いる(後述)。



図 1: Physical AI Smart Spaces Dataset の動画の例。



図 2: 同一 ID の人物画像例。GroundTruth の BoundingBox で Crop した画像を 10 サンプル抽出したもの。

MeTRAbs の姿勢推定出力(SMPL24形式)

人物の映る向き θ

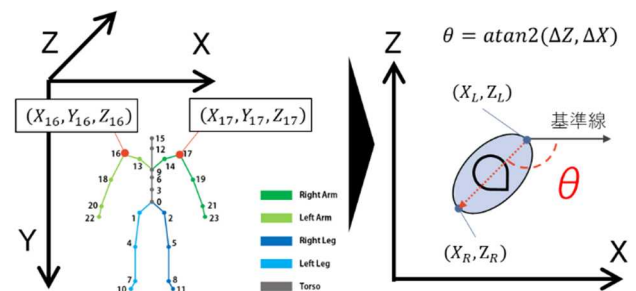


図 3: 人物の映る向き θ の算出方法。MeTRAbs の 3 次元姿勢推定結果の左肩・右肩の座標をもとに角度を算出した。

3.2 評価方法

画像ペアに対する同一/非同一人物の 2 値分類精度を、分類タスクで一般的な Accuracy/ Precision/ Recall で評価した。分類モデルは予備実験で高い精度を発揮した XGBoost[15]を用い、説明変数は以下を入力した。

- CosineSimilarity: ReID 特徴量の類似度
- $\Delta d(c,p)$: カメラ-人物間距離の差分
- $\Delta \theta c$: カメラ高さの差分
- $\Delta \theta p$: 人物の映る向きの差分

CosineSimilarity の算出に用いる ReID 特徴量の抽出には、撮影条件の違いに頑健と考えられる OSNet-AIN を用いた。 $\Delta d(c,p)/ \Delta \theta c/ \Delta \theta p$ は画像ペア間の GroundTruth 同士の差分を直接算出した。 $\Delta \theta p$ は Crop 画像に対して 3 次元姿勢推定モデルである MeTRAbs[16]を用いて相対 3 次元座標を推論し、その方向ベクトルのなす角を算出した(図 3)。

3.3 実験 1

ReID 特徴量の類似度のみで分類した場合をベースラインとして、撮影条件を説明変数に加えた XGBoost によって精度が向上するか確認した。撮影条件別に同一 ID の画像を 50 枚抽出し、非同一 ID の画像もほぼ同数抽出した。したがって同一 ID の画像ペア(正例)は各実験で $50 \times 49 = 2450$ 組存在する。XGBoost を使わずに、Accuracy が最も高くなる類似度閾値で分類した場合をベースラインとして、前述の説明変数を用いて XGBoost を学習した結果と比較した。学習においては 5 分割交差検証を行い、XGBoost のパラメータは、予備実験において分類精度が最も高かった $\text{depth}=3, \text{n_estimator}=100$ を採用した。

実験結果を表1に示す。撮影条件を説明変数に加えることで若干の精度改善は見られたが、いずれの指標も1%以内の改善にとどまり、精度に大きな差はなかった。説明変数の Importance, SHAP 値を確認すると、CosineSimilarity の影響が突出していた。そのため、単純に ReID 特徴量の類似度と撮影条件を同列で説明変数に加えるだけでは、撮影条件の違いを考慮した同一人物同定精度の改善効果は限定的であると考えた。

3.4 実験 2

撮影条件を加味した精度改善の可能性を探るため、各撮影条件の違いが同一人物分類精度にどの程度精度に影響を与えているか評価した。具体的には撮影条件に関する3つの説明変数を、独立してバリエーションを制限するフィルタリングを複数パターン適用し、精度の違いを評価した。同一 ID の画像枚数は、実験1の 50 枚に対して、類似の撮影条件を揃えられる 20 枚とした。その他の実験条件は実験 1 と同様である。また、各撮影条件の実験において分類精度が最も高かったフィルタリングのケースについては、説明変数の CosineSimilarity の分布を可視化した。

表 1: 実験 1 結果。太字は最高精度。

	Accuracy	Precision	Recall
類似度のみ (ベースライン)	0.842	0.841	0.840
撮影条件加味 (XGBoost)	0.848	0.846	0.847

A) カメラ-人物間距離

分類モデルの学習に用いる画像を、カメラ-人物間距離でフィルタリングして精度を比較した(表 2)。フィルタは近傍、遠方、その両方でフィルタリングするケースに分けて評価した。

近傍フィルタ、遠方フィルタのいずれにおいても、フィルタリングを強めてカメラ-人物間距離のバリエーションを絞り込むほど分類精度が高くなる結果となった。近傍と遠方で比較すると、遠方フィルタを適用し、遠い画像同士の方が同一人物の分類精度が高かった。この傾向は両方でフィルタリングした場合も同様で、実験ケースの中では最も遠い 15-20m でフィルタリングしたケースが最も精度が高かった。直感的には近傍の画像の方が特徴量がリッチで分類精度が高くなると考えられるが、ReID モデルの入力時に画像はリサイズするため、必ずしも近傍の高解像度の画像であっても ReID 特徴量がリッチになるとは限らない。また、遠方の背景が類似しているなどの、データセット由来のバイアスによる可能性も否定できない。遠方の画像の方が分類精度が高くなる理由については、今後さらなる精査が必要である。

Accuracy が最高となった 15-20m のケースの CosineSimilarity の分布を図 4 に示す。Negative ペアの分布はフィルタリング前後でほとんど変わらないのに対し、Positive ペアはコサイン類似度が向上していることが確認できる。カメラ-人物間距離が近い画像では同一人物画像の ReID 特徴量のばらつきが抑えられ、それが分類精度向上に寄与していると考ええる。

B) カメラ高さ

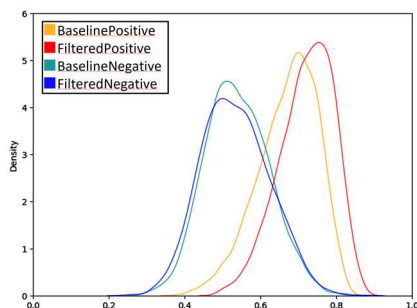
分類モデルの学習に用いる画像を、カメラ高さでフィルタリングして精度を比較した(表 2)。

A)の実験と同様に、フィルタリングを強めてカメラ高さのバリエーションを絞り込むほど分類精度が高くなる結果となった。

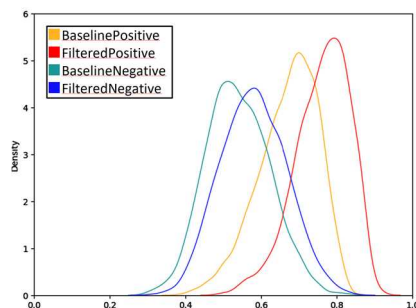
最高精度となった 2.3 m 以下のケースの CosineSimilarity の分布を図 4 に示す。カメラ高さでフィルタリングした場合、A)と異なり Negative ペアの分布もコサイン類似度が向上している。しかし、Positive ペアの分布は Negative ペア以上にコサイン類似度が向上しており、それが分類精度向上に寄与していると考えた。

表 1:実験結果一覧表. 下線はフィルタ種別ごとの最高精度, 太字は全ケース中の最高精度.

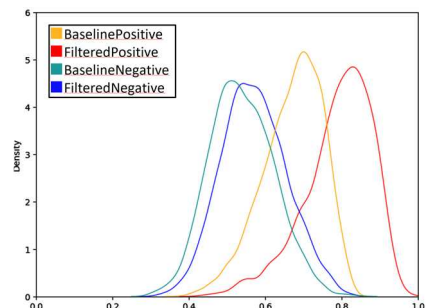
実験ケース			[参考]Train			Test		
			Accuracy	Precision	Recall	Accuracy	Precision	Recall
フィルタなし(ベースライン)			0.797	0.805	0.777	0.798	0.798	0.792
実験 A カメラ-人物間距離	近傍 フィルタ	20m 以下	0.809	0.814	0.793	0.810	0.815	0.795
		15m 以下	0.817	0.823	0.800	0.818	0.829	0.794
		10m 以下	0.849	0.853	0.839	<u>0.851</u>	<u>0.847</u>	<u>0.851</u>
	遠方 フィルタ	10m 以上	0.817	0.799	0.841	0.819	0.803	0.840
		15m 以上	0.848	0.839	0.857	0.850	0.839	0.860
		20m 以上	0.873	0.859	0.888	<u>0.874</u>	<u>0.859</u>	0.892
	近傍 遠方 フィルタ	5-10m	0.853	0.872	0.823	0.855	0.865	0.836
		10-15m	0.851	0.851	0.845	0.852	0.853	0.845
		15-20m	0.875	0.875	0.872	<u>0.877</u>	<u>0.881</u>	<u>0.866</u>
実験 B カメラ高さ	高さ フィルタ	3.0m 以下	0.811	0.814	0.799	0.812	0.803	0.821
		2.5m 以下	0.847	0.851	0.835	0.848	0.850	0.840
		2.3m 以下	0.850	0.839	0.861	<u>0.852</u>	<u>0.846</u>	<u>0.856</u>
実験 C 人物の映る向き	向き フィルタ	±90 度	0.816	0.819	0.804	0.817	0.819	0.807
		±60 度	0.827	0.825	0.823	0.828	0.835	0.811
		±30 度	0.852	0.850	0.849	<u>0.854</u>	<u>0.854</u>	<u>0.848</u>
実験 D(各実験の最高精度のフィルタ組合せ)			0.896	0.890	0.875	0.898	0.893	0.874



実験 A: 近傍+遠方フィルタ 15-20m



実験 B: 高さフィルタ 2.3m 以下



実験 C: 向きフィルタ ±30 度

図 4: 各実験の最高精度を記録したケースにおける ReID 特徴量のコサイン類似度分布(横軸はコサイン類似度)

C) 人物の映る向き

分類モデルの学習に用いる画像を, 人物の映る向きでフィルタリングして精度を比較した(表 2). なお, 向きは正面向きを0度として, そこからの角度でフィルタリングしている.

A)の実験と同様に, フィルタリングを強めて人物の映る向きのバリエーションを絞り込むほど分類精度が高くなる結果となった.

最高精度となった ± 30 度のケースの CosineSimilarity の分布を図 4 に示す. カメラ高さでフィルタリングした場合, A)と異なり Negative ペア

の分布もコサイン類似度が向上している. しかし, Positive ペアの分布は Negative ペア以上にコサイン類似度が向上しており, それが分類精度向上に寄与していると考ええる.

D) 3変数のフィルタ組み合わせ

撮影条件の3変数すべてのフィルタを組み合わせた. フィルタは正例のサンプル数が確保できた「カメラ-人物間距離 15m 以下, カメラ高さ 3m 以下, 人物の映る向き ± 40 度」とした. その結果, 最も高い精度を記録した(表 2), 撮影条件が近い画像に絞り込むほど同一人物分類精度が向上することが確認できた.

3.5 考察

実験2では、フィルタをかけて撮影条件を近づけるほど精度が向上していることから、今回実験した3要素の撮影条件の違いが精度低下を招くことが確認できた。3つの撮影条件A)B)C)の説明変数のスケールは異なるため横並びの比較はできないが、B)C)に比べてA)のフィルタによる精度改善が大きい。同一人物識別精度の頑健性向上に向けては、カメラ-人物間距離の違いを考慮することが有効であると示唆される。

一方で図4の3つのグラフを比較すると、ReID特徴量のコサイン類似度分布はC)向きフィルタ±30度による改善が高いように見えるが、精度評価ではA):近傍+遠方フィルタ15-20mの方が高い。この理由としては、C)ではPositiveペアだけでなくNegativeペアも類似度が上がってしまっているのに対して、A)ではNegativeペアの類似度が変わらないことによるものと推察した。さらに本研究では、同一/非同一人物の分類モデルの説明変数としてReID特徴量のコサイン類似度に加えて撮影条件の違いも入力していることも要因と考えた。すなわち、分類モデルでカメラ高さや人物の映る向きは同一人物識別に寄与するのに対して、カメラ-人物間距離は相対的に寄与度が低い可能性がある。ただし分類モデルは各ケースで別々に学習しており、各ケースのフィルタによりサンプリングされる人物画像も異なることから、今回の実験からは結論づけることはできない。

4 まとめ

本研究では、撮影条件に対するモデルの頑健性向上に向けた基礎的知見を得ることを目的として、多様な撮影条件を含む合成データセットを用いて、条件の差異がReID特徴量に基づく人物識別精度および類似度分布に及ぼす影響を分析した。その結果、撮影条件の違いが識別精度低下を招いていることを確認し、今後の精度改善に向けた考察を行った。

ただし本研究の結果は、あくまで対象としたデータセットで確認できたものであり、別のデータセットや実動画で同じ結果が得られるとは限らないことに留意が必要である。今後は、本研究で得られた知見に基づき、撮影条件に対して頑健な同一人物識別手法を考案し、合成データではない実際の監視カメラの動画で検証することが課題である。

参考文献

- [1] 吉田龍人, 大久保順一, 藤井純一郎, 高森秀司, 天方匡純: Person Re-identification の実用に向けた DNN モデルの特性評価, AI・データサイエンス論文集, Vol. 4, No. 3, pp. 385–392, 2023.
- [2] W. Li, R. Zhao, T. Xiao, and X. Wang: Deep-reid: Deep filter pairing neural network for person re-identification, *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 152–159, 2014.
- [3] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang: Transreid: Transformer-based object re-identification, *In Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15013–15022, 2021.
- [4] S. Li, L. Sun, and Q. Li: Clip-reid: exploiting vision-language model for image re-identification without concrete text labels, *In Proceedings of the AAAI conference on artificial intelligence*, Vol. 37, pp. 1405–1413, 2023.
- [5] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang: Bag of tricks and a strong baseline for deep person re-identification, *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 1487–1495, 2019.
- [6] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang: Omni-scale feature learning for person re-identification, *In Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3702–3712, 2019.
- [7] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang: Learning generalisable omni-scale representations for person re-identification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 9, pp. 5056–5069, 2021.
- [8] K. He, X. Zhang, R. Ren, and J. Sun: Deep residual learning for image recognition, *In Proceedings of the IEEE conference on*

computer vision and pattern recognition, pp. 770–778, 2016.

- [9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian: Scalable person re-identification: A benchmark, *In Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015.
- [10] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi: Performance measures and a data set for multi-target, multi-camera tracking, *In Proceedings of Lecture Notes in Computer Science*, Vol. 9914, pp. 17–35, 2016.
- [11] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian: MARS: A video benchmark for large-scale person re-identification, *In Proceedings of the European Conference on Computer Vision*, pp. 868–884, 2016.
- [12] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang: Global-local temporal representations for video person re-identification, *In Proceedings of the IEEE international conference on computer vision*, pp. 3958–3967, 2019.
- [13] S. Wang, D. C. Anastasiu, L. Zheng, et al.: The 8th AI City Challenge, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7261–7272, 2024.
- [14] Y. Wang, T. Meinhardt, O. Cetintas, et al.: BEV-SUSHI: Multi-target multi-camera 3d detection and tracking in bird’s-eye view, *arXiv:2412.00692*, 2024.
- [15] T. Chen and C. Guestrin: XGBoost: A scalable tree boosting system, *In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
- [16] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe: MeTRAbs: Metric-scale truncation-robust heatmaps for absolute 3D human pose estimation, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Vol. 3, No. 1, pp. 16–30, 2021.