F-014

複数の監視カメラを用いた人物同定の一検討
Evaluation of person re-identification method using multiple surveillance cameras

坂口 将生† 尼崎 太樹‡ 木山 真人‡ 岡本 利章†
Shoki Sakaguchi Motoki Amagasaki Masato Kiyama Toshiaki Okamoto

1 はじめに

近年，AI(Artificial Intelligence) 技術の急速な発達により，監視カメラの用途は犯罪の抑止だけでなく，人流解析，群衆カウント，人物追跡といった映像解析技術による活用が進んでいる．例えば，街や観光地といった人が多く集まる場所では，人流を適切にコントロールする重要性が増している．このような場所でこれらの映像解析は有用な手段となる．

こうしたなか，視野が重ならない複数カメラ間で同一人物を見つける技術に注目が集まっている．これは，Person re-identification(以下，人物同定) と呼ばれ，近年盛んに研究がなされている．これら人物同定の技術を使用し，監視カメラの映像から人物がどのように動いたのかを正確に追跡することで，警備員が人流を適切にコントロールすることが期待されている．

本研究では，国の重要文化財に指定されている熊本城の監視カメラを複数用いた人物同定を行う．現在の人物同定のデータセットは，人物がどの順番でカメラに映るかは未知である．一方，本研究で作成した熊本城データセットは，一本道に設置された複数カメラを使用しているため，人物が次に映るカメラが既知である．また，歩行者を正面から撮影したカメラと，側面から撮影したカメラが存在するという特徴がある．一般に，人物同定は，カメラの俯角や，人物の向き，似た服の人の存在等の理由で照合が困難な場合がある．特に，本実験で使用する，熊本城データセットは観光地ということもあり，これらの問題が顕著に現れる．そこで，上記で述べた熊本城データセットの特徴に着目し，時間情報と方角情報がこれら問題解決の助けになると考えた．本研究では，撮影される歩行者の向きの影響を検討するとともに，視覚情報のみの人物同定と，時間情報と方角情報を使用した人物同定を比較し，有効性の検証を行う．

本稿は以下の構成をとる．まず，2 章で人物同定の概要と空間的・時間的情報を活用した関連研究について紹介する．3 章では本研究で作成した熊本城データセットの概要と，特徴について述べる．4 章では本研究の提案システムについて述べる．5 章では評価条件と，提案システムの評価について述べる．最後に，6 章で本稿の結論について述べる．

2 関連研究

本章では，人物同定の概要と空間的，時間的情報を活用した既存モデルについて述べる．人物同定は，重複しない複数のカメラにわたって，照合したい人物 (以下，Query) を人物の特徴が格納された登録データベース (以下，Gallery) から見つけることを目的としている．近年，ディープニューラルネットワークの進化とビデオ監視の需要の増加により，人物同定が大きな関心を集めている．

人物同定の基本的なアプローチとして，視覚的特徴表現のための深層学習モデルが挙げられる [1][2]．これら深層学習モデルの性能を向上させるために，現在までに多くの手法が提案されている．最先端のアプローチでは，ベンチマークデータセット Market1501[3] で 90%以上の rank-1 精度と，高いパフォーマンスを達成している．しかし，これらの手法では，外観の曖昧さの問題にはほとんど対処することができない．

この課題に対応するために，空間的，時間的情報に注目した研究もなされている．st-ReID[4] では，教師あり人物同定のために，空間的，時間的情報を視覚的特徴表現に自然に統合する効果的な共同分析を追求している．st-ReID では，ラプラス平滑化やロジスティック関数を使用し，視覚的類似度と空間的，時間的確率分布のモデル化を行なっている．これら空間的，時間的情報を使用したモデルは，歩行者がいつ，どのタイミングで現れるか分からないことを前提としている．しかしながら，本研究で使用する熊本城データセットは一本道に設置されたカメラである．このため，次に映るカメラが既知ということから，歩行者がどのタイミングで現れるか分かっているという前提で人物同定を行う．

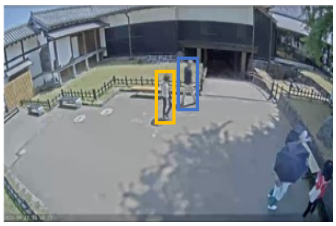
3 熊本城データセット

本章では，新たに作成した熊本城データセットの概要と特徴について述べる．データセットの収集には，熊本城に設置された視野が重ならない図 1 の A，B，C の 3 台の監視カメラを使用する．本データセットは，ベンチマークデータセット Market1501 を元に作成した．全て手作業でアノテーションをするには，莫大な時間と人手を要するため，以下に示す手順にて半自動で作成した．

まず，Multi Object Tracking のモデル，ByteTrack[5] を使用し，各カメラで撮影された映像ごとに人物の追跡を行う．その後，30 フレームおきに追跡された ID とそれに対応した bbox を収集する．ここでは，追跡 ID の間違い，低解像度の画像，重なり合った歩行者画像等が含まれているため，手作業にてこれらの修正を行う．最終的に，本データセットには 283 の ID に対応する 9,621 枚の bbox が含まれている．それぞれの bbox には，歩行者の ID，撮影されたカメラ，撮影されたフレーム数，向かっている方角が情報として存在し，各画像サイズは，(64，128) である．また，本データセットには，以下の特徴がある．

まず，各 ID の歩行者画像は，3 台のカメラで必ず撮影されている点である．また，それぞれのカメラは，一本道に設置されており，必ず歩行者は A→B→C または，C→B→A の順番に撮影される．すなわち，A から歩いてきた歩行者が B に撮影されず，C に撮影されることはない．

† 熊本大学大学院自然科学教育部 Graduate School of Science and Technology，Kumamoto University
‡ 熊本大学大学院先端科学研究部 Faculty of Advanced Science and Technology，Kumamoto University



(a) カメラ A 画像

(b) カメラ B 画像

(c) カメラ C 画像

図 1: 熊本城データセット使用監視カメラ [6]

次に，3 台のカメラそれぞれで撮影された歩行者の向きが似ているカメラと異なるカメラが存在する点である．図 1 の監視カメラ画像から，A のカメラと C のカメラでは，撮影される歩行者の向きが異なっているが，B のカメラと C のカメラでは，撮影される歩行者の向きが似ている．つまり，A から C の方向に向かう歩行者の場合，B と C のカメラでは正面から撮影された画像に対し，A のカメラでは側面または背後から撮影された画像である．

4 提案システム

本章では，本研究における提案システムについて述べる．4.1 節では本実験の概要について述べる． 4.2 節では使用した ReID モデルと，提案システムにおけるフィルタリング手法について述べる．

4.1 概要

本実験では，熊本城データセットで学習し評価を行う．その際，C のカメラを Gallery で固定し，A と B のカメラを Query とした場合の結果をそれぞれ比較する．ここでは，撮影される歩行者の向きの影響を検討する．また，視覚情報のみを使用した場合，視覚情報に加え，時間情報を使用した場合，視覚情報と時間情報に加え，方角情報を使用した場合の 3 つで結果をそれぞれ比較する．その際，時間情報は，Train データから得られた統計情報を使用し，フィルタリング処理を行う．また，方角情報は，データセット内に含まれるラベルを使用する．ここでは，本データセットにおける時間情報や方角情報の有効性について検討する．

4.2 人物同定

本節では，本実験で使用した ReID モデルと，フィルタリング手法について説明する．

4.2.1 MGN[7]

本研究では，ReID モデルとして MGN(Multiple Granularity Network) を使用する．MGN は，大域的な特徴表現のため の 1 つのネットワークと局所的な特徴表現のための 2 つのネットワークからなるディープネットワークアーキテクチャで設計されている．なお，ネットワークのバックボーンには，ImageNet[8] で事前学習した ResNet-50[9] を使用する．最終的な出力として，大域的特徴と局所的特徴の両方の情報を結合させた 2048 次元の特徴量を得る．また，Query と Gallery 内の画像との比較には，コサイン類似度を用いる．本実験では，Query と Gallery 内の全ての画像に対し，コサイン類似度でスコア化し，上位 5 枚を抽出する．

4.2.2 フィルタリング手法

本実験では，無関係な Gallery 画像を排除するために，時間情報と方角情報からフィルタリング処理を行う．フィルタリングを行う時間間隔については Train データから得られた統計情報を用いる．

まず，視覚情報に加え，時間情報を使用する場合は，統計情報から得られたフレーム間隔でフィルタリング処理を行う．ここで，Query の歩行者が撮影されたフレーム数を Q_t，Gallery 内の歩行者が撮影されたフレーム数を G_t と表したとき，$t_1 \leq |Q_t - G_t| \leq t_2$ の条件でフィルタリング処理を行う．

次に，視覚情報と時間情報に加え，方角情報を使用する場合は，方角情報が一致しているかどうかのフィルタリング処理を行う．Query の方角情報を Q_{dir}，

表 1: データセット概要

	日時	ID	camA(枚)	camB(枚)	camC(枚)
Train	2022 年 5 月 19 日 (木)13 : 00 ~ 14 : 00	162	2240	1759	1925
Test	2022 年 6 月　3 日 (金)13 : 00 ~ 14 : 00	121	1481	1176	1040

Gallery 内の歩行者が撮影された方角情報を G_{dir} と表したとき，$Q_{dir} = G_{dir}$ の条件でフィルタリング処理を行う．また，方角情報が与えられたときの時間情報は，$t_1 \leq Q_t - G_t \leq t_2$，もしくは，$t_1 \leq G_t - Q_t \leq t_2$ の条件でフィルタリング処理を行う．

5　評価

本章では 4 章で説明した提案システムの学習と性能評価における条件と結果について述べる．5.1 節では，評価項目，データセットやパラメータについて述べる．5.2 節では，5.1 節で述べた条件で評価を行い，5.3 節では，考察を述べる．

5.1　評価条件

5.1.1　評価項目

本研究では，Cumulative Matching Characteristics(以下，CMC)[10] と mean Average Precision(以下，mAP)[3] を使用する．CMC-k(以下，rank-k)[10] は，Gallery 内から検索された結果の上位 k 位に正しい ID が現れる確率を表す．rank-k は，検索された結果の上位 k 位のみを考慮するため、各クエリに対して Gallery 内の正解 ID が 1 つしか存在しない場合に有効である．しかし，大規模なカメラネットワークでは，Gallery 内には通常複数のグランドトゥルースが含まれており，CMC はモデルの性能を完全に反映することはできない．

もう一つの指標である mAP は，Gallery 内に正解 ID が複数存在する場合にモデルの平均的な性能を測定する．これはもともと画像検索で広く使われている指標である．各 Query について，Average Precision(以下，AP)[3] を計算し，全ての Query の AP の平均値，すなわち，mAP を計算する．

5.1.2　データセット

詳細な熊本城データセットの情報について，表 1 に示す．Train データとして，2022 年 5 月 19 日 (木) の 13:00 から 14:00 のデータを使用した．こちらには，162 の ID に対応する 5,924 枚の bbox が含まれている．また，Test データとして，2022 年 6 月 3 日 (金) の 13:00 から 14:00 のデータを使用した．こちらには，121 の ID に対応する 3,697 枚の bbox が含まれている．

5.1.3　統計情報

図 2 に，Train データ 162 名分のカメラ間に映るまでのフレーム間隔を示す．これより，フィルタリング処理を行うフレーム間隔を A-C 間では，t_1 を 6,000 に，t_2 を 34,000 フレームに設定し，B-C 間では，t_1 を 4,000 に，t_2 を 30,000 フレームに設定した．

また，Train データには，C から A の方角に向かう 67 名と，A から C の方角に向かう 95 名が存在する．それぞれで，A-C 間のカメラ間に映るまでのフレーム間隔を図 3 に，B-C 間のカメラ間に映るまでのフレーム間隔を図 4 示す．これより，C から A の方角に向かう歩行者よりも，A から C の方角に向かう歩行者の方が短いフレーム間隔で現れており，分布の間隔が短いことが確認できる．これは，A の方角には，天守閣があり，A から C の方角に向かう歩行者は天守閣を観光し終わり，帰宅する人が多い．一方で，C から A の方角に向かう歩行者は，これから天守閣を観光しようとしている歩行者が多く，通路の途中で立ち止まって天守閣を見たり，写真撮影を行なっている．このことがそれぞれで差が生まれている要因であると考えられる．そこで，方角情報と時間情報を組み合わせる際には，方角ごとにフィルタリング処理を行う時間間隔を変更した．A-C 間において，C の方角に向かっている場合は，t_1 を 6,000 に，t_2 を 26,000 フレームに設定し，A の方角に向かっている場合は，t_1 を 8,000 に，t_2 を 34,000 フレームに設定した．また，B-C 間において，C の方角に向かっている場合は，t_1 を 4,000 に，t_2 を 20,000 フレームに設定し，A の方角に向かっている場合は，t_1 を 8,000 に，t_2 を 34,000 フレームに設定した．

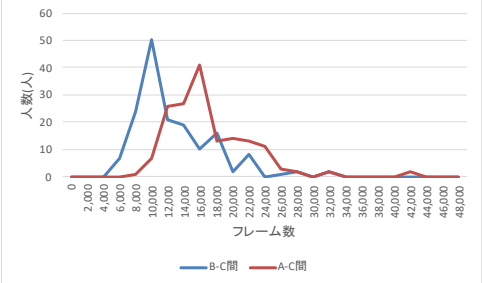


図 2: フレーム間隔

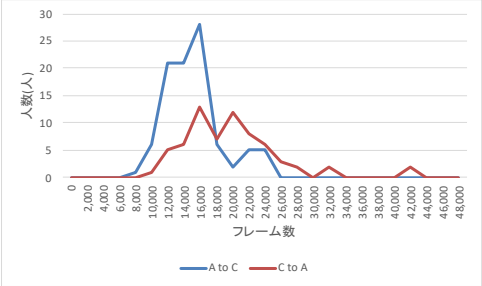


図 3: A-C 間の方角によるフレーム間隔の違い

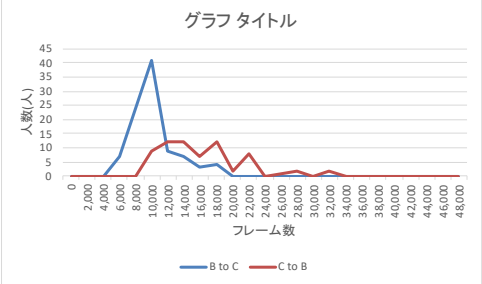


図 4: B-C 間の方角によるフレーム間隔の違い

表 2: 結果 (Query:A，Gallery:C)

	mAP	rank-1	rank-5
視覚情報	55.3	63.0	74.2
視覚情報+時間情報	64.8	70.7	80.8
視覚情報+時間情報+方角情報	79.6	82.9	89.0

表 3: 結果 (Query:B，Gallery:C)

	mAP	rank-1	rank-5
視覚情報	80.1	90.4	96.9
視覚情報+時間情報	86.2	94.6	98.2
視覚情報+時間情報+方角情報	93.2	97.6	99.1

5.2 評価結果

表 2 に，Query を A のカメラ，Gallery を C のカメラの場合，表 3 に，Query を B のカメラ，Gallery を C のカメラの場合の結果を示す．それぞれ視覚情報のみの場合と比較して，時間情報と方角情報を使用してフィルタリング処理を行なった場合の方が精度が高い．また，撮影された歩行者の向きが似ているカメラ (Query:A，Gallery:C) と異なるカメラ (Query:B，Gallery:C) で精度に大きく差が見られた．

5.3 考察

ある Query に対し，視覚情報のみの場合と，時間情報と方角情報を使用してフィルタリング処理を行なった場合の実験結果について述べる．視覚情報のみで行なった場合の実験結果を図 5 に示す．また，視覚情報+時間情報で行なった場合の実験結果を図 6 に示し，視覚情報+時間情報+方角情報で行なった場合の実験結果を図 7 に示す．

図 5 と図 6 より，時間情報を使用することで，無関係と考えられる Gallery 内の歩行者画像を削除し，効果的に予測することができている．また，視覚的特徴表現の ReID モデルは，誤った人物でも，人物の向きが同じ場合に，同一人物であると判断する傾向にある．しかし，方角情報を使用することで，異なった向きの場合でも予測精度が向上する (図 7)．すなわち，人物の向き，似た服の人の存在で視覚情報のみでの照合が困難な場合でも，時間情報と方角情報を使用することが人物同定の助けとなる．

6 結論

本論文では，熊本城データセットを新たに作成し，撮影される歩行者の向きの影響を検討するとともに，視覚情報のみの人物同定と，時間情報や方角情報を使用した人物同定を比較し，有効性の検証を行った. 結論は以下の通りである．

1. 一本道において時間情報と方角情報を使用したフィルタリング処理が有効である．本実験では，データセットの時間が 1 時間と短い期間であったが，期間が長ければ長いほど，本手法は効果を発揮することが予想される．
2. 歩行者を撮影する向きが精度に大きな影響を与える．視覚的特徴表現では，歩行者の向きを大きく考慮してしまう傾向にある．方角情報を使用したフィルタリング処理は，これらを解決する有効的な手段である．本研究では，方角情報を既知の情報として扱ったが，方角を自動的に検知することが今後の課題として挙げられる．

図 5: 視覚情報のみ

図 6: 視覚情報+時間情報

図 7: 視覚情報+時間情報+方角情報

参考文献

[1] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang, “Beyond part models: Person retrieval with refined part pooling,” Proc. of ECCV, pages 480–496, 2018.

[2] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang, “Omni-scale feature learning for person re-identification,” Proc. of ICCV, pages 3702–3712, 2019.

[3] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, “Scalable Person Re-identification: A Benchmark,” Proc. of ICCV, pages 1116–1124, 2015.

[4] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie, “Spatial-temporal person re-identification,” Proc. of AAAI, pages8933-8940, 2019.

[5] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang, “Bytetrack: Multi-object tracking by associating every detection box,” arXiv preprint arXiv:2110.06864, 2021.

[6] 熊本城「観覧案内」, https://castle.kumamoto-guide.jp/info/

[7] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou, “Learning discriminative features with multiple granularities for person re-identification,” Proc. of MM, pages 274–282, 2018.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” Proc. of CVPR, pages 248–255, 2009.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” Proc. of CVPR, pages 770–778, 2016.

[10] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu, “Shape and appearance context modeling,” Proc. of ICCV, pages 1–8, 2007.