Vol.2020-CVIM-221 No.35
2020/3/17

複数の深層学習法に基づいた歩行者再識別

張継偉 [1]　呉海元 [2]

概要：本稿では，２種類のディープ・ラーニングを融合した歩行者の再識別方法を提案する．監視カメラで撮影された歩行者の映像から，ディープ・ラーニング（instance segmentation）によって背景を除去し，歩行者のグローバル特徴を抽出する．得られた特徴マップに基づいて畳み込みネットワーク（bag of tricks and strong baseline）よって歩行者の再識別を実現する．

キーワード：ディープ・ラーニング，歩行者の再識別，グローバル特徴

Person Re-identification based on multiple deep learning methods

JIWEI ZHANG †1　　HAIYUAN WU †2

Abstract: In this paper, we propose a Person re-identification method fused the two types of deep learning algorithm. First, by using the instance segmentation algorithm, we remove the background and extract Person global features of a Person in a video taken by a surveillance camera. Then, based on the obtained feature map, Person re-identification is realized by a convolution network (bag of tricks and strong baseline).

Keywords: deep learning,　MS-Word,　Person re-identification,　global features

1. はじめに

近年，安全安心な社会を構築するために，監視カメラを急速に普及し，防犯の同時に重大犯罪が発生する時，観測データに基づいて犯人の同定に貢献している．その関連で，人の身体的な特徴を用いて個人を識別する手法の研究が盛んに行われている．一方，カメラ映像中の対象人物は空間解像度が低く，鮮明な顔情報が得られない場合が多いため，実際の監視環境では顔識別を行えないケースが多い．

歩行者の再識別（Person Re-identification，略して ReID とも呼ばれる）は，特定の歩行者が画像シーケンスの中に存在するかどうかを判断するするテクノロジーである．歩行者は顔が映っていなく，または空間解像度の低い映像から歩行者歩き方や外観などの情報を抽出できれば，その歩行者に関する個人識別を行えると考える．これは，画像検索のサブ問題として一般化考えると同じ，監視対象の歩行者画像を指定すると，異なるカメラで撮影された映像から歩行者画像を探し出す．例えば，図１に示すような,ビデオシーケンスがキャプチャされたエリアには複数のカメラがある．ReID では，関心のある歩行者のすべての写真を取得する必要がある．

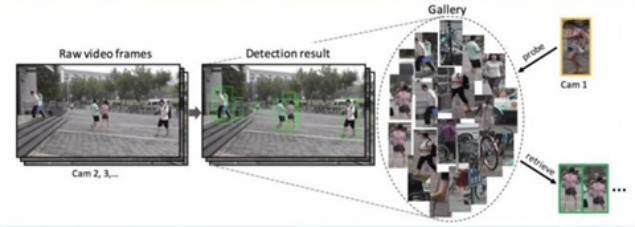

図 1　ReID のフロー[4]

Figure　1　ReID Flow

歩行者の再識別は，背景変化，姿勢，照明，カメラの視点の変化などの原因で，頑健な識別用特徴量を抽出することは困難になり，伝統的なコンピュータビジョンの手法だけで識別率を保証できるアルゴリズムは少なかった．

近年，ディープニューラルネットワークに基づく歩行者認識（ReID）テクノロジーが大きな進歩しているため，身体の部位からそれぞれの特徴を自動的に抽出できるようになる．

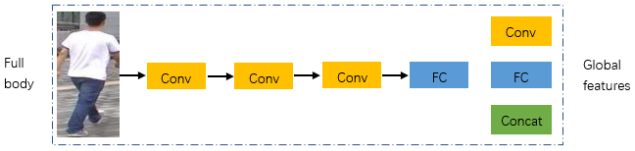

図 2　グローバル特徴

Figure　2　global features

1 和歌山大学
Wakayama University.
2 和歌山大学
Wakayama University

従来の ReID の研究は，画像全体を使用して画像検索用の特徴ベクトルを取得し，グローバルな特徴[1,2,3,4]（図 2）に焦点を当てていた．しかし，グローバル特徴がボトルネックに残っていたため，ローカル特徴の研究を開始した．ローカル特徴[5]を抽出するために，画像のセグメンテーション，スケルトンキーポイント（Skeleton key points）の配置，姿勢修正などが一般的に使用される．

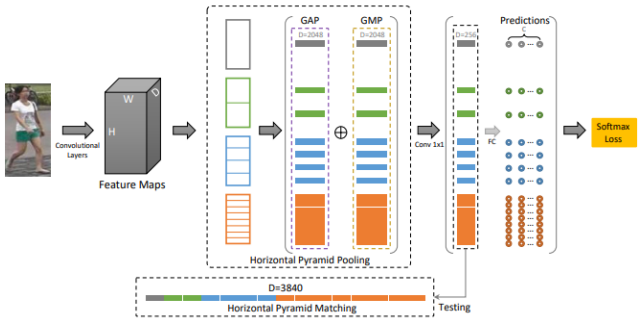

図 3　ローカル特徴の研究例[4]

Figure 3　Examples of local features

図 3 に示すような，歩行者再識別用のローカル特徴では，画像はいくつかのブロックで垂直に分割されている．そのセグメント化された画像ブロックは順に短期記憶ネットワーク（Long Short Term Memory Network, LSTM）[6]に送信され，最終的な特徴がすべての画像ブロックのローカル特徴を融合する．しかし，この手法の問題点は，画像同士の位置合わせの精度が非常に高いことを求めている．例えば，2 枚の画像内の人体位置はズレがあれば，頭と上半身のローカル特徴が融合されてしまい，構成されたモデルに精度の問題がある．

2. liu らの歩行者の再識別法

Liu ら[7]は，SSD（Single Shot MultiBox Detector）（図 4）と ReID（図 5）の 2 つのネットワーク構造を利用して歩行者の再識別を実現している，最初に SSD を使用して画像内の歩行者を検出する．次に監視対象の歩行者を元の画像から切り取り，検出された歩行者たちの画像とともにバックエンドに送信する．ネットワーク ReID は，それらが監視対象の歩行者であるかどうかを判定する．

2.1 SSD 構造の概要

SSD アルゴリズムは，ターゲットカテゴリ（Target category）とバンテックボックス（bounding box）を直接予測するマルチターゲット検出アルゴリズムである．Faster RCNN と比較し，SSD アルゴリズムはプロセスレイヤーを生成しないため，検出速度が大幅に向上することができる．

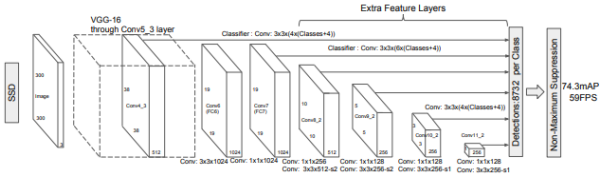

図 4　SSD ネットワーク構造

Figure 4　SSD Network Structure

サイズが異なるターゲットを検出するために，従来の NMS（Non-Maximum Suppression）法では，最初に画像をピラミッドに変換し，次にそれらをそれぞれ検出し，最後に結果を統合する．このように，SSD アルゴリズムは，異なる畳み込みの特徴マップを使用して異なるターゲットの検出を実現することができる．図 4 に示すような，SSD アルゴリズムの主なネットワーク構造は VGG16[14]で，最後の 2 つの完全接続層は畳み込み層に変更され，そして，ネットワーク構造を構築するために 4 つの畳み込み層を追加した．

2.2 VGG16 構成の概要

オックスフォード大学のコンピュータービジョングループ（Visual Geometry Group）と Google DeepMind の研究者は,新しい深層畳み込みニューラルネットワーク VGGNet を開発した．VggNet には合計 6 つの異なるネットワーク構造があり,各構造には 5 つの畳み込みグループが含まれる．畳み込みの各グループは 3×3 の畳み込みカーネルを使用し，グループコンボリューションの後，2x2 の最大プーリングを実行し,続いて 3 つの完全に接続されたレイヤーが実行される．

表 1　ConvNet 構成[14]

Table 1　ConvNet configurations

A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					

conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4s096					
FC-1000					
soft-max					

2.3 ReID 構造の概要

ReID は，歩行者の再識別の問題を分類として表現するディープニューラルネットワークアーキテクチャ（図 5）である． 入力画像のペアが与えられた場合，タスクは 2 つの画像が同じ人物を映っているかどうかを判定でき，同じ人物に属するかどうかの最終推定値を生成する. ReID ネットワークは，複数の異なるレイヤーで構成され, 2 つ層バインドコンボリューション（Bind convolution），最大プール，クロス入力近隣差（Cross input neighborhood differencs），クロスパッチフィーチャ（Cross patch feature），完全接続レイヤー（Fully connected layers），および softmax 関数を含む.

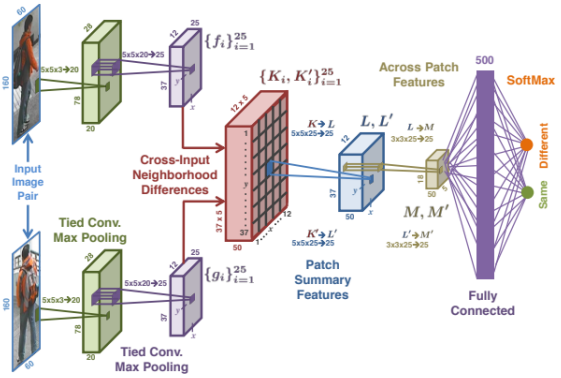

図 5　ReID ネットワーク構造
Figure 5　ReID Network Structure

3. 提案手法

3.1 歩行者検出の改良

今までの歩行者検出法では，背景を含める特徴マップを作成し利用するので，認識精度に影響を与える欠点が存在している．それで，認識精度を向上するために，本研究では, Bolya ら[9]のアルゴリズムに基づいて背景を除去し，歩行者の特徴マップを作成する．Bolya ら[9]のアルゴリズムでは，以下の 3 つの利点を含む：

- ネットワークへの変形可能な畳み込みを導入することによって，ネットワーク表現能力を強化し，より効率的な検出器とマスクプロトタイプを提供できるようになる.
- ターゲットのより良いアンカーサイズとアスペクト比を設定できるようになる.
- 高速マスクリスコアリングブランチがネットワークに導入された，マスクの品質評価を導入するためにマスクスコアリング RCNN を使用する.

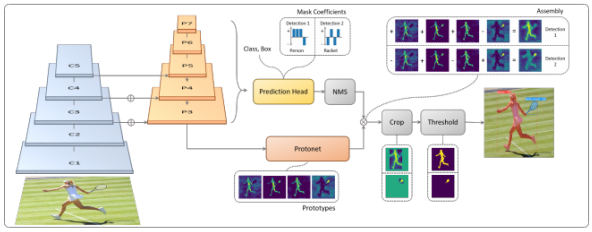

図 6　YOLACT++アーキテクチャ[9]
Figure 6　YOLACT++ Architecture

3.2 歩行者の再識別

本研究では, Luo ら[10]のオープンソースコードに基づいて歩行者の再識別を行う．画像内の歩行者は他の物体に見え隠れされる可能性があるので，歩行者の再識別モデルの汎化能力を高めるために, Luo らは遮蔽問題を解決し，ランダム消去増強（図 7）(Random erasing ,REA) [11]というアルゴリズムを導入した．この方法を用いることは，遮蔽問題による影響を減少でき，モデルのロバスト性も高めることができる．しかし，実際の監視データセットを用いた実験結果より，未学習のデータセットに対して検出率が低下した（クロスドメインパフォーマンス低下）問題が残っている.

そこで，本稿では，クロスドメインパフォーマンスを向上させるために，以下の 3 点を改良する：

- 複数のデータセットが共同でトレーニングする
- クロスドメイン（Cross-domain）効果を改善するために，より良いクロスドメイン IBN-Net を入れ替える.
- トレーニング中にランダム消去の操作部分は削除し，ネットワークはより多くの背景情報を学習する可能性がある.

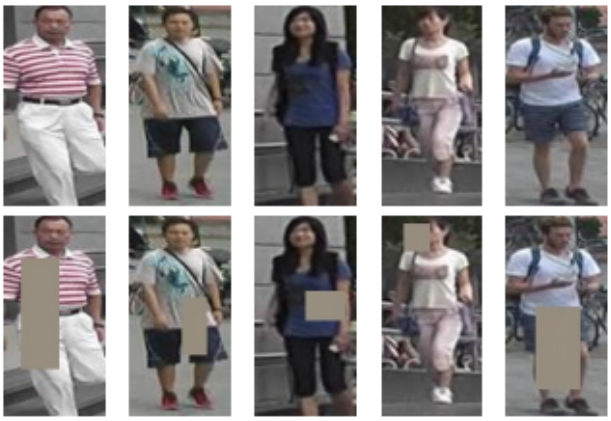

図 7　ランダム消去
Figure 7　Random erasing

4. 実験

本節では, 提案手法の有効性と安定性を示すために,複数のデータセットを用いて行者の再識別の験を行い,評価を行った.

4.1 歩行者検出の実験（Detection experiment）

歩行者の検出の実験では,MS COCO データベース（Microsoft Common Objects in Context）の中で人を含める画像（約 6 万枚の画像）を用い, 1 つの GPU のみを使用し 350,000 回にトレーニングを行った.

表 2　歩行者の検出のトレーニング
Table 2　Person detection training.

Training datasets	MS COCO
Weights	yolact_coco_person_43_350000
Number of pictures	64115
Training tools	CPU, GPU
Lable	Person
Training times	350,000
Test datasets	Market1501

図 8 に示すような,レーニングモデルに基づく Market1501 データトで画像の背景を除去するテストを行った．画像の背景を除去するため、歩行者のグローバルな特徴マップを作成し．歩行者の検出の認識精度を向上することができる.

図 8　背景除去
Figure 8　Background remove

4.2 歩行者の再識別を用したデータベース

クロスドメインパフォーマンス（cross-domain performance）を改善するために, 歩行者の再識別の実験では, Market1501, CUHK03, MSMT17 の 3 つのデータセットを同時に用い,約 17 万枚の画像でトレーニングを行った. また，DukeMTMC-reID をテストデータベースとして利用した.

表 3　データベース
Table 3　Database

Dataset	MSMT17	Duke	Market	CUHK03
BBoxes	126441	36411	32668	28192
Identities	4101	1812	1501	1467
Cameras	15	8	6	2
Detector	Faster RCNN	hand	DPM	DPM,
Scene	outdoor,indoor	outdoor	outdoor	indoor

4.3 歩行者の再識別の評価方法

提案モデルを学習するとき，標準ベースラインに優れたクロスドメイン IBN-Net を追加し，トレーニングの設定を変更しない.

評価指標

学習された提案モデルを評価するために,ランク 1 の精度（Rank-1 Accuracy）と平均精度（Mean Average Precision, mAP ）という二つ評価指標を用いた．さらに,提案手法が過剰適合による誤識別率を軽減できていることを示すために，クロスドメイン(cross-domain)の実験も行った.

4.4 歩行者の再識別のブレーション実験

上記のデータセットを同時に用いたトレーニングを行い, トレーニングされていない Duke データセットでクロスドメインテストを行った．次に, 各トリック（Trick）でアブレーション（Ablation）実験を比較した. 本実験は, Duke データセットが評価データセットとしてテストを行った.

表 4　ブレーション実験

Table 4　Ablation experiment.

Model	DukeMTMC mAP	DukeMTMC Rank-1
ResNet50 IBN-a 256x128	40.7%	58.2%
ResNet50 256x128	49.2%	66.0%
+Weight Decay	50.1%	68.0%
+Label Smooth	47.5%	66.0%
+No Bias Weight Decay	53.7%	70.3%
+Tuning	55.6%	71.6%

4.5 実験結果

DukeMTMC データセットのテストセットをテストした結果は, Rank-1 が 71.9%, mAP が 56.2%であった．従来手法より優れたクロスドメイン効果を達成できていることを確認できた.

アブレーション（ablation）研究の結果は, 各トリック（trick）によるパフォーマンス（performance）の向上を示している． 過剰適合による誤識別を防ぐために, クロスドメイン(cross-domain)実験の結果も示す.

実験配置:

本実験では, パソコンを利用し, python3.6 Windows10 のプラットフォームでトレーニングしたモデルに基づいて歩行者の再識別のテストを行った.

表 5　実験環境の構成

Table 5 Experimental environment configuration

Training datasets	Market1501, CUHK03, MSMT17
Weights	yr50_ibn_a.pth
Number of pictures	175000
CPU	intel (R) corel7-4790@3.6GHZ
GPU	RTX 2080TI
python packages	opencv-python, tb-nightly , torch >= 1.0
Test datasets	DukeMTMC-reID

上記の実験環境でパフォーマンスのテストを実施し,改良した歩行者再識別のアルゴリズムに基づいて歩行者を検出し,歩行者のグローバル特徴マップを作成した．実験の処理速度は約 33FPS のリアルタイム(30 FPS 以上)に達した.

5. あとがき

本論文では 2 つの先行研究を融合して改良した歩行者再識別法を提案し，３つの共通データセットを用いて学習することによって歩行者再識別法モデルを構築した．未学習データセットを用いたテスト結果より, IBN-Ne を用いてクロスドメイン効果を得られた．しかし, 歩行者検出と歩行者再識別には, まだ様々な不足が残っている．したがって, 歩行者再識別技術の利用範囲が大幅に拡大するために, 今後の研究ではさらに以下のような部分に対して改善を行う必要があると考える.

- 検出プロセス

実際の検出プロセスでは, 外部環境が複雑であり（照明, オクルージョンなど）, カメラが揺れるなど, 検出結果のみに依存し, 検出フレームは不安定な可能性がある．それで, 検出と追跡を組み合わせる研究を行う.

- 識別プロセス

推定を高速化するために, 歩行者再識別モデルは剪定（Pruning）[12]や知識の蒸留（Knowledge Distillation）[13]などトリックを行う.

参考文献

[1] Hermans, A., Lucas B., and Bastian L. In defense of the triplet loss for person re-identification, 2017,arXiv preprintarXiv:1703.07737.
[2] Ding, Shengyong, et al. Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition 48.10, 2015: 2993-3003.
[3] Li, Dangwei, et al. Learning deep context-aware features over body and latent parts for person re-identification, 2017, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[4] Ahmed E, Jones M, Marks T K. An improved deep learning architecture for person re-identification, Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3908-3916.
[5] Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Huang, and T. Horizontal pyramid matching for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligenc,2019, Vol. 33, pp. 8295-8302.
[6] Hochreiter, S., Schmidhuber, and J. Long short-term memory. Neural computation, 1997, 9(8), 1735-1780.
[7] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, and A. C. Ssd: Single shot multibox detector. In European conference on computer vision , 2016, pp. 21-37.
[8] Ahmed, E., Jones, M., Marks,and T. K. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015 , pp. 3908-3916.
[9] Bolya, D., Zhou, C., Xiao, F., Lee, and Y. J. YOLACT++: Better Real-time Instance Segmentation, 2019, arXiv preprint arXiv:1912.06218.
[10] Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, and W. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops , 2019, pp. 0-0.
[11] Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation,2017,arXiv preprint arXiv:1708.04896.

[12] He, Y., Kang, G., Dong, X., Fu, Y., and Yang, Y. Soft filter pruning for accelerating deep convolutional neural networks, 2018, arXiv preprint arXiv:1808.06866.
[13] Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531.
[14] Simonyan, K., and Andrew, Z .Very deep convolutional networks for large-scale image recognition , 2014, arXiv preprint arXiv:1409.1556.