

Pair-wise AttentionNet による歩行者候補領域の高精度化

小芝 駿王†, 福井 宏†, 山下 隆義†, 藤吉 弘亘†, 村瀬 洋‡

†: 中部大学, {koto@mprg.cs, fhiro@mprg.cs, takayoshi@isc, fujiyoshi@isc}.chubu.ac.jp

‡: 名古屋大学

概要: Faster R-CNNベースの歩行者検出は, Region Proposal Network(RPN)で歩行者候補領域を検出し, 歩行者候補領域に対して歩行者と背景の2クラス識別を行う. しかし, RPNで検出した領域にオクルージョンが発生していると, 位置ずれが生じやすく, 背景と誤識別することがある. そこで本研究では, 繰り返し処理により矩形の補正と歩行者判定を同時に行うAttentionNetを活用し, 誤識別を抑制する. 一方で, 従来のAttentionNetは矩形の左上と右下の座標位置の移動方向を独立して推定するため, オクルージョンが発生した歩行者領域に対して, 正しい移動方向の推定ができない. そのため, 本研究では, 左上と右下の座標位置の移動方向を連動させたPair-wise AttentionNetを提案する. これにより, オクルージョンが発生した歩行者領域に対しても矩形の補正が可能となる. 評価実験により提案手法は, 従来手法では補正が困難な歩行者領域に対しても正確な矩形の検出を可能とし, オクルージョンが発生した歩行者領域に対する精度向上を確認した.

1. はじめに

歩行者検出は画像中の歩行者の位置と大きさを推定する技術で, 自動運転支援や経路予測, 姿勢推定など幅広く活用されている. 従来の歩行者検出は, HOG特徴量 [1]が用いられており, 歩行者の勾配方向ヒストグラムを特徴量として扱うことで, 歩行者の向きや姿勢, 服装などの見えの変化に頑健な検出を実現している. DalalらがHOG特徴量を提案後, HOG特徴量をベースにした手法が数多く提案されている[2] [3] [4]. Dollorらは, 色情報や勾配情報等を用いたチャンネル特徴量とBoosted treeを組み合わせたICFが提案している[5]. 複数の特徴抽出手法を組み合わせたチャンネル特徴量を用いることで検出性能が大幅に向上し, ICFをベースにした歩行者検出法が提案されている[6][7][8]. 2012年に開催された, 大規模な一般物体認識コンテスト[9]においてDeep Convolutional Neural Network(CNN)を用いた方法が優勝して以降[10], CNNを歩行者検出に応用した手法が数多く提案されている[11][12][13][14].

本研究ではその中でも, 歩行者検出のベンチマークで一般的に用いられるCityPersons Dataset[15]において, 高い精度を実現しているFaster R-CNN[11]をベースに用いる. この手法は, Region Proposal

Network (RPN)で歩行者候補領域を検出し, 歩行者候補領域に対して歩行者と背景を識別する. しかし, Faster R-CNNは, RPNで検出した候補領域にオクルージョンが発生している場合や歩行者のスケールが小さい場合, 正解位置に対してずれが生じやすい. 正解領域からずれた領域は歩行者全体を正しく囲むことができず, 歩行者を誤って背景と識別する原因となる. そこで, 本研究では, 矩形を繰り返し処理により補正するAttentionNet[16]に着目し, 位置ずれが生じた歩行者領域に対して補正および歩行者判定を同時に行う. AttentionNetは, 与えられた歩行者領域に対して矩形の左上と右下の座標位置の移動方向を推定して補正を行う. この処理を繰り返し行い, 矩形を正解領域に近づけていくことで正確な矩形の検出を可能とする. しかし, AttentionNetは矩形の左上と右下の座標位置の動きを個別で推定するため, 歩行者にオクルージョンが生じている場合, 左上または右下の座標を正しく移動できず背景と判定し, その領域を誤って棄却してしまう.

そのため, 本研究では左上と右下の座標位置の動きを連動させたPair-wise AttentionNetを提案する. Pair-wise AttentionNetでは, 矩形の左上と右下の移動方向を連動させた16種類の動きと, 背景クラスの全17クラスの推定を行う. 左上と右下の移動方向を連動

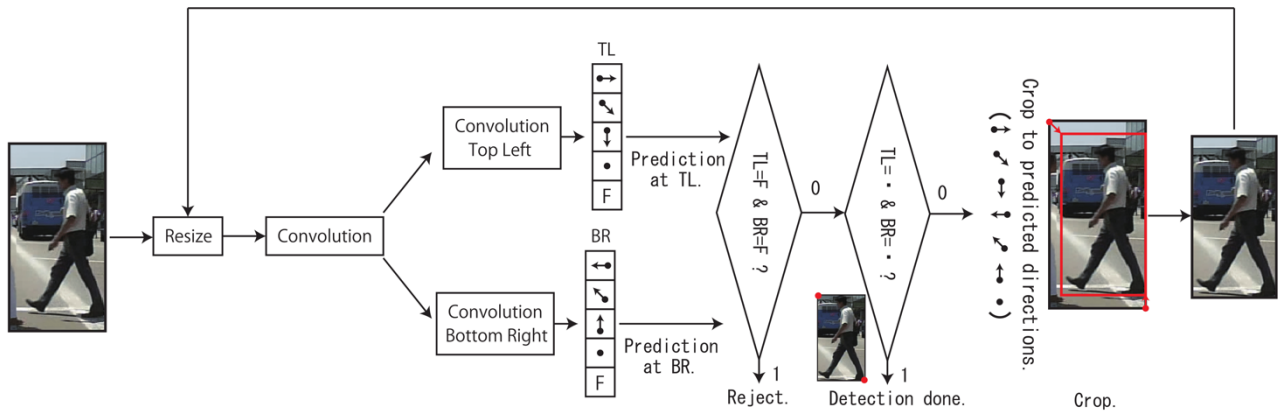


図 1: AttentionNet の構造

させることで、従来のAttentionNetの場合、左上または右下の推定結果が背景となり、誤って棄却していたオクルージョンの生じている歩行者に対する検出精度を向上させることが可能となる。

2. 関連手法

歩行者検出は、入力画像から直接歩行者位置を検出する1段階の手法と、歩行者候補領域の検出後に歩行者か背景かを判定する2段階の手法がある。2段階の手法では歩行者以外の一般物体を対象とした検出手法があり、Faster R-CNNやAttentionNetが代表的な手法である。ここでは、これらの関連手法について述べる。

2.1. 1段階の検出手法

Dalalらは歩行者の勾配に着目するHistogram of Oriented Gradient (HOG)特徴量を提案し、Support Vector Machine (SVM)と組み合わせることで、歩行者検出の性能を大幅に向上させている[1]。そして、DeformablePart Model (DPM)は、解像度の異なるHOG特徴量から、歩行者の全身と身体のパーツを捉えて検出するため、姿勢の変化に頑健な歩行者検出を実現している。Dollarらは、HOG特徴量の勾配方向ヒストグラムだけでなく、色情報等を特徴量に加えるチャンネル特徴量を用いたICFを提案している[5]。チャンネル特徴量は、勾配情報や色情報などを複数の画像のチャンネルのように扱う特徴量である。また、識別器にBoosted treeを用いることで、膨大な特徴量群から歩行者検出に有効な特徴量を選択することができる。ICFにチャンネル特徴量のAggregate(集約)処理を加えたAggregate Channel Feature(ACF)は、各チャンネル特徴量に対して特徴ピラミッドを導入することで、歩行者のスケール変化に頑健な歩行者検出

を実現している[7]。その後、歩行者検出においてCNNをベースとした手法が数多く提案されている[12][13][14]。近年ではFast R-CNN[17]やFaster R-CNN[12]、Single Shot Detector[18]をベースにした歩行者検出法が高精度かつ高速な検出を実現している。

2.2. 2段階の検出手法

Faster R-CNNはRPNにおいて、入力画像に対して畳み込み処理を行い、特徴マップを得る。その後、特徴マップの各位置について、アンカーを用いて物体領域かどうか判定する。アンカーは、スケールとアスペクト比の異なる矩形のパターンであり、アンカーごとに物体かどうか判定を行う。それと同時に物体の位置や大きさについても出力する。その位置の特徴をRoI poolingにより一定サイズにプーリングし、全結合層のネットワークに入力する。そして歩行者かどうか判定する。

AttentionNetは物体候補領域の左上と右下の座標位置に対する移動方向を予測するネットワークである。AttentionNetの構造を図1に示す。AttentionNetはネットワークが途中で2つに分岐しており、それぞれ矩形の左上、右下の座標位置を補正する移動方向を推定する。各分岐で推定した移動方向に矩形を修正し、画像を切り出し、再びAttentionNetへ入力する。加えて、AttentionNetは補正終了と背景クラスの判定も行う。両方の座標で補正終了と推定された場合は、繰り返し処理を終了させ、背景と推定された場合は、その領域を棄却する。本研究では繰り返し処理が20回を超えても両方の座標で終了と推定されない場合、その領域に対する補正を終了し背景として棄却する。補正終了と推定されるまで何度も処理を繰り返すことで正確な歩行者矩形を得ることができる。しかし、AttentionNetに図2のような下半身にオクルージョンが

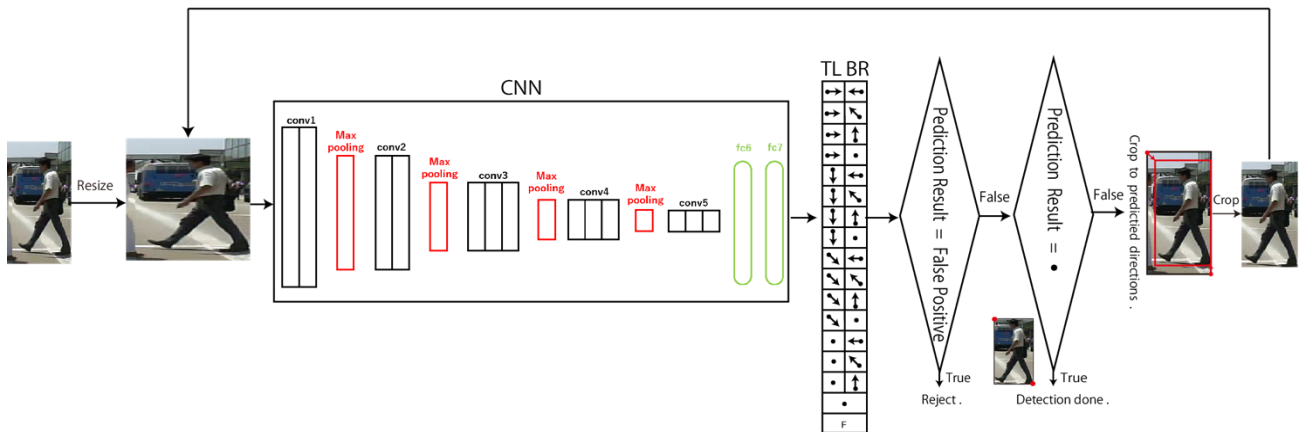


図 5: Pair-wise AttentionNet の構造

発生した歩行者候補領域を入力した場合、右下の座標位置の推定に必要な足元の特徴が獲得できず、背景と推定する。AttentionNetは矩形の左上と右下の座標位置の動きを独立して推定するため、図2のような領域を誤って棄却してしまう。



図 2: オクルージョンした歩行者領域の例

3. 提案手法

本研究では、オクルージョンに頑健なPair-wise AttentionNetを提案する。Pair-wise AttentionNetは、Faster R-CNNのRPNで検出した歩行者候補領域に対して、矩形の補正を繰り返し行う。その際、矩形の左上と右下の移動方向を連動させることで、従来のAttentionNetの問題点であるオクルージョンの生じた歩行者領域に対する棄却を抑制することができる。

3.1. RPN の検出領域の拡大

Pair-wise AttentionNetに入力する歩行者候補領域の位置及び大きさはRPNの検出結果に依存する。歩行者の正解領域の左上及び右下の座標と、RPNの検出領域の座標の差を図3に示す。ここで、左上の座

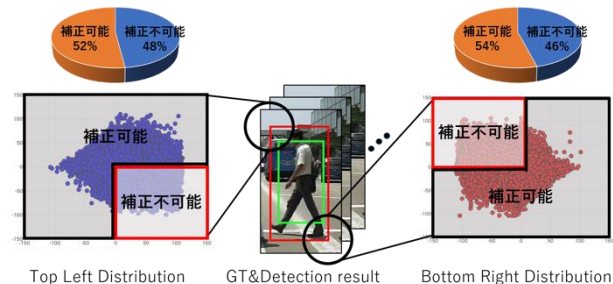


図 3: 正解と検出領域との座標の差

標の差は青点、右下の座標の差は赤点でそれぞれ示している。また、AttentionNetによる補正可能である領域と補正不可能である領域の割合を円グラフで示す。

図3から、全候補領域の45%以上で左上と右上の座標が正解の座標よりも内側に位置している。これは図4(a)に示すような緑色の領域で、従来のAttentionNetでは補正ができずに正しく検出できない。そのため本研究では、RPNの検出領域を拡大して入力することで、補正不可能に分布している領域に対してもAttentionNetで矩形を補正できるようにする。一方、図4(c)のように領域を拡大しすぎた場合は、同じ領域中に二人以上の歩行者を含んでしまうため、AttentionNetでは補正ができない。そのため、領域中に二人以上の歩行者を含まず、AttentionNetで矩形を補正できるようにする必要がある。本研究では、評価実験においてRPNの領域の拡大率ごとの平均Miss Rateを示し、最適な拡大率を確認する。

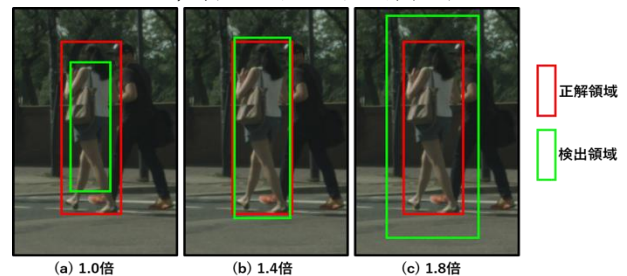


図 4: 領域の拡大

3.2. Pair-wise AttentionNet の構造

Pair-wise AttentionNetの構造を図5に示す。Pair-wise AttentionNetは、ネットワークを途中で分岐させず、矩形の左上、右下の座標の移動方向を連動させている。本研究では、RPNで検出した歩行者候補領

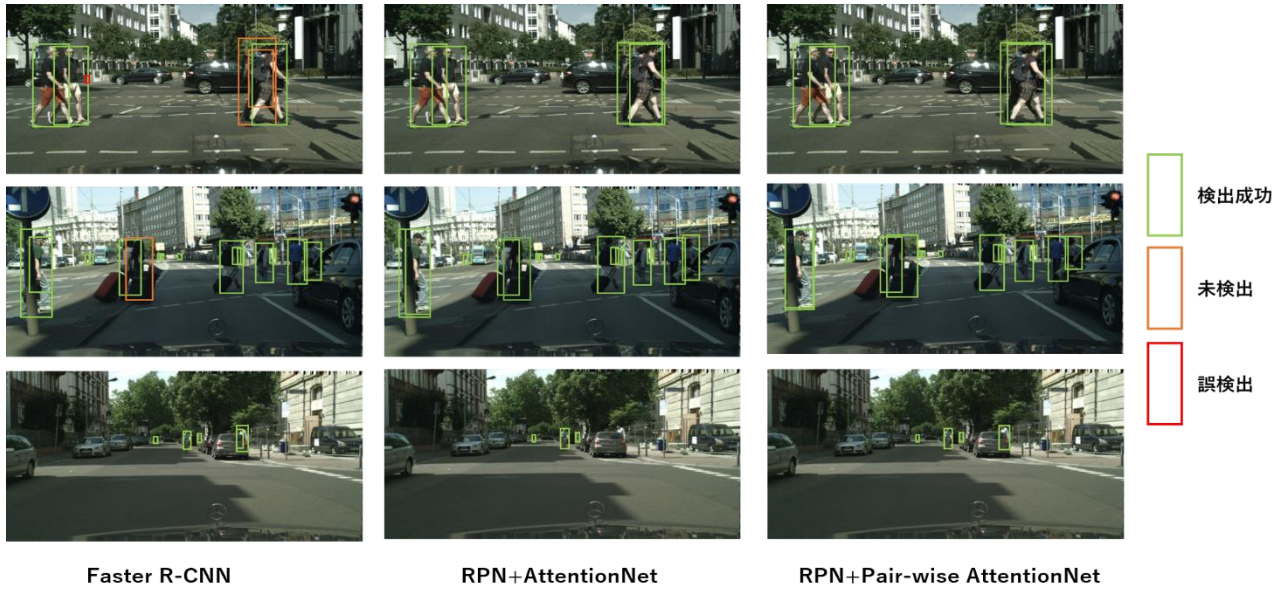


図 6: 検出結果の例

域を入力として用いる. Pair-wise AttentionNetは, 図5のように左上と右下の座標位置の動きを連動させ, (右向き, 左向き), (右向き, 上向き)などの16種類の移動方向のペアと背景を合わせた17種類いずれかを推定する. 左上と右下の座標の動きを連動させ, 背景を1クラスとすることで従来のAttentionNetの場合に片方の座標で背景と推定して棄却されていたオクルージョンが生じた歩行者に対する検出精度の向上が期待できる. 本研究では, 畳み込み層が13層, 全結合層が3層から構成されるVGG16をベースとしたネットワーク構造とする. また, Faster R-CNNを歩行者検出に応用するため, アンカーのアスペクト比を0.41, 1.0, 2.0に固定する. また, CityPersons Datasetの評価指標に合わせるため, 学習サンプルを50pixel以上の領域に限定する.

4. 評価実験

評価実験では, Pair-wise AttentionNetの有効性を示すため, 従来のAttentionNetと検出精度の比較を行う.

4.1. 実験概要

データセットにはCityPersons Datasetを使用する[11]. AttentionNetの学習には歩行者1,687,950枚, 背景21,842枚, 評価にFaster R-CNNのRPNで検出した歩行者5,826枚, 背景3,886枚用いる. 入力画像サイズは224×224とし, 学習回数は200エポック, ミニバッチサイズを12とする. 領域の検出の成否は, 正解領域との重なり率IoUで判定する. ここで, 正解領域の面積をT, 出力される領域の面積をSとする.

$$\text{IoU} = \frac{S \cap T}{S \cup T} \quad (1)$$

一般的な歩行者検出では, 正解と検出結果のIoUが50%以上のとき正解として精度を比較する. 本実験では, より正確な歩行者の矩形の検出に関する精度の変化を確認するため, IoUのしきい値を50%から90%まで変化させて各手法の精度を比較する.

4.2. RPNの検出領域の拡大率ごとの精度

RPNの検出領域に対する拡大率がAttentionNetの精度に影響するか確認する. 表1に, RPNの検出領域を1.2倍から1.7倍まで0.1刻みで拡大した際の検出精度を示す. 表1より, RPNの拡大率を大きくすると平均Miss Rateは低くなるが, 大きくしすぎると平均Miss Rateは高くなる. RPNの拡大率を1.4倍にした時にAttentionNetのMiss Rateが最も低い. そこで, 本研究では, 最適な拡大率を1.4倍として利用する.

4.3. 全歩行者に対する精度の比較

表 1: 拡大率ごとの平均 Miss Rate[%]

拡大率[倍]	1.2	1.3	1.4	1.5	1.6	1.7
RPN+						
AttentionNet	35.54	22.43	18.30	18.40	19.24	42.37

従来手法と提案手法の平均Miss Rateと, 矩形の補正処理の繰り返し回数を表2に示す. 表2より, IoUのしきい値が50%の場合, 提案手法は平均Miss RateをFaster R-CNNより約1.2%, RPN+AttentionNetより約1.08%低下させることができた. また, IoUのしきい値を80%とした場合, 平均Miss RateをFaster R-CNNより約9.54%, RPN+AttentionNetより約7.24%低下させる

ことができた. RPN+Pair-wise AttentionNetはIoUのしきい値を80%とした際に, 大きく精度が向上している. この結果から, 提案手法であるRPN+Pair-wise AttentionNetは, より正解に近い検出が可能であることがわかる. しかし, IoUのしきい値を90%とした場合, 提案手法と従来手法の精度には大きな差はなかった. これは矩形補正の繰り返し処理に関する終了条件をIoUが85%以上としており, 90%を満たす矩形の補正が対象外であるためである. 一方, 補正の繰り返し回数は, 従来のAttentionNetよりも平均0.52回増加しただけである. 以上のことから, 提案手法であるPair-wise AttentionNetによる検出領域の補正が有効であることがわかる.

表 2: 全歩行者に対する平均 Miss Rate[%]

IoU	50%	60%	70%	80%	90%
Faster R-CNN	18.42	21.89	34.06	60.78	91.42
RPN+ AttentionNet (補正回数)	18.30 (3.2 回)	21.34 (4.2 回)	33.02 (4.9 回)	58.48 (6.2 回)	90.34 (6.9 回)
RPN+Pair-wise AttentionNet (補正回数)	17.22 (3.9 回)	19.51 (5.0 回)	29.92 (5.4 回)	51.24 (6.6 回)	89.32 (7.1 回)

検出領域の例を図6に示す. 図6より, RPN+Pair-wise AttentionNetとRPN+AttentionNetは, Fater R-CNNより正解領域に近い歩行者領域を検出できている. また, 2段階の検出構造を用いたRPN+Pair-wise AttentionNetとRPN+AttentionNetは後段の手法により誤検出や未検出の抑制が確認できる.

4.4. オクルージョンした歩行者に対する精度比較

Pair-wise AttentionNetによるオクルージョンが発生した歩行者に対する検出精度の向上を確認するため, 検出対象をオクルージョンの生じている歩行者に限定したときの各手法の検出精度を比較する. 検出対象の身体が35%までオクルージョンした歩行者に限定した時(Partial occlusion)の時の各手法の平均Miss Rateを表3, 35%から75%までオクルージョンした歩行者に限定した時(Heavy occlusion)の各手法の平均Miss Rateを表4に示す. 表3より, Faster R-CNNとRPN+AttentionNetを比較すると, すべての重なり率でRPN+AttentionNetのMiss Rateが低いことが分かる. これにより, 身体のオクルージョンが35%ほどの場合, 従来のAttentionNetを用いても問題なく補正が可能であることが分かる. しかし, 表4より, Faster R-CNNとRPN+AttentionNetを比較すると, すべての重なり率

でRPN+AttentionNetのMiss Rateが高いことが分かる. このことより, 従来のAttentionNetはオクルージョンが大きい歩行者領域に対して補正ができていないことが分かる. また, RPN+Pair-wise AttentionNetは, Faster R-CNN, RPN+AttentionNetと比較すると, どちらの手法よりもMiss Rateを低下させることができており, Pair-wise AttentionNetによる補正とオクルージョン

表 3: Partical occlusion における平均 Miss Rate [%]

IoU	50%	60%	70%	80%	90%
Faster R-CNN	18.69	22.36	35.17	58.74	90.40
RPN+ AttentionNet	18.44	21.82	33.52	56.90	89.98
RPN+Pair-wise AttentionNet	17.52	20.63	30.74	50.72	88.54

表 4: Heavy occlusion における平均 Miss Rate [%]

IoU	50%	60%	70%	80%	90%
Faster R-CNN	48.75	61.69	75.85	89.28	98.41
RPN+ AttentionNet	64.42	72.35	83.25	92.52	99.32
RPN+Pair-wise AttentionNet	45.25	60.64	74.73	86.54	97.79

が発生した歩行者領域に対する精度向上が確認できる.

オクルージョンが生じた歩行者の検出領域の例を図7に示す. 図7(b)から, RPN+AttentionNetはオクルージョンが発生した歩行者を棄却しているのに対し, 提案手法は棄却せずに検出できていることがわかる. しかし, 図7(a)の低解像度の歩行者に対してRPN+AttentionNetとRPN+提案手法のいずれでも最終の検出領域が得られる前に背景として推定されて棄却していることがわかる.

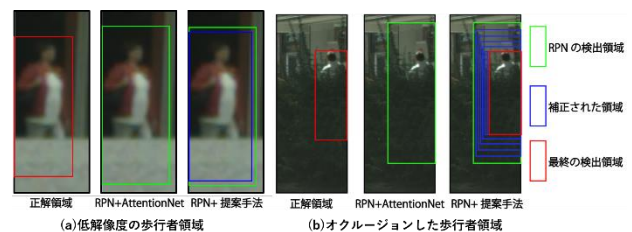


図 7: オクルージョンした歩行者の検出領域の例

5. おわりに

本研究では, 歩行者候補領域の右上と左下の座標位置を連動して補正可能なPair-wise AttentionNetを提案した. 正解領域のIoUのしきい値を80%以上と

した場合に、最大で約9.54%の検出精度の向上を実現した。今後は拡大方向にも移動できるような出力方法を検討し、RPNの出力する矩形のばらつきに対応する。

謝辞 本研究の一部は独立行政法人科学技術振興機構(JST)の研究成果展開事業「センター・オブ・イノベーション(COI)プログラム」の支援によって行われた。

参考文献

- [1] N.Dalal : Histograms of Oriented Gradients for Human Detection. CVPR, 2005.
- [2] P.Felzenszwalb : A Discriminatively Trained, Multiscaled, Deformable Part Model. CVPR, 2008.
- [3] Z.Cai : An HOG-LBP human detector with partial occlusion handling, ICCV, 2009.
- [4] W.Stefan : New Features and Insights for Pedestrian Detection. CVPR, 2010.
- [5] P.Dollar : Integral Channel Features. British Machine Vision Conference, 2009.
- [6] R.Benenson : Pedestrian detection at 100 frames per second. CVPR, 2012.
- [7] W.Nam : Local Decorrelation For Improved Pedestrian Detection. Neural Information Processing Systems, 2014.
- [8] P.Dollar : Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.
- [9] R.Olga : ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 2015.
- [10] A.Krizhevsky : Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25, eds. by F. Pereira, et al., pp.1097-1105, Curran Associates, Inc., 2012.
- [11] S.Ren : Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Neural Information Processing Systems, 2015.
- [12] W.Ouyang : Joint deep learning for pedestrian detection. ICCV, 2013.
- [13] P. Luo : Switchable Deep Network for Pedestrian Detection, 2014.
- [14] Z.Cai : Learning Complexity-Aware Cascades for Deep Pedestrian Detection, ICCV, 2015.
- [15] S.Zhang : CityPersons: A Diverse Dataset for Pedestrian Detection. CVPR, 2017.
- [16] D.Yoo : Attentionnet:Aggregating weak directions for accurate object detection. ICCV, 2009.
- [17] R.Girshick : Fast R-CNN. ICCV, 2015.
- [18] W.Liu : SSD: Single Shot MultiBox Detector. ECCV, 2016.

小芝駿王:現在中部大学大学院修士課程在学中,画像を用いた歩行者検出の研究に従事

福井宏:2016 年中部大学大学院博士前期課程修了,現在同大学大学院博士後期課程在学中,画像を用いた物体検出の研究に従事.

山下隆義:2002年 奈良先端科学技術大学院大学博士前期課程修了, 2002年 オムロン株式会社入社, 2011年 中部大学大学院博士後期課程修了(社会人ドクター), 2014年 中部大学講師, 2017年 中部大学准教授, 人の理解に向けた動画像処理, パターン認識・機械学習の研究に従事, 2015年画像センシングシンポジウム オーディエンス賞, 2016年画像電子情報通信学会 査読功労賞, 2016年画像センシングシンポジウム 優秀学術論文賞, 2016年画像センシングシンポジウム 最優秀学術論文賞, 2017年 IWAIT2017 Best Paper Award.

藤吉弘亘:1997年 中部大学大学院博士後期課程修了, 1997年 米カーネギーメロン大学ロボット工学研究所Postdoctoral Fellow, 2000年 中部大学工学部情報工学科講師, 2004年 中部大学准教授, 2005年 米カーネギーメロン大学ロボット工学研究所客員研究員(～2006年), 2010年 中部大学教授, 2014年名古屋大学客員教授, 計算機視覚, 動画像処理, パターン認識・理解の研究に従事, 2005年 ロボカップ研究賞, 2009年 情報処理学会論文誌CVIM優秀論文賞, 2009年 情報処理学会山下記念研究賞, 2010, 2013, 2014年画像センシングシンポジウム優秀学術賞, 2013年 電子情報通信学会 情報・システムソサイエティ論文賞.

村瀬洋:1980年名古屋大学大学院工学部電気電子工学専攻修士課程卒業. 1980年 NTT 入社. 1987年名古屋大学大学院情報工学専攻工学博士取得. 1992年米国コロンビア大学コンピュータ科学部客員研究員. 2003年名古屋大学大学院情報科学研究科メディア専攻教授, 2017年名古屋大学大学院 情報学研究科 研究科長・教授, 文字・図形認識, コンピュータビジョン, マルチメディア認識の研究に従事, 1985年篠原学術推奨賞. 1992年テレコムシステム技術賞. 1994年IEEE Best Paper Award : CVPR. 1995 年山下記念研究賞. 1996年IEEE Best Video Award : ICRA. 2001年高柳記念推奨賞. 2001年システムソサイエティ論文賞. 2002年電子情報通信学会業績賞. 2003年文部科学大臣賞. 2004年 IEEE 論文賞. 2004年画像認識理解シンポジウム MIRU2004 優秀論文賞. 2005年テレコムシステム技術推奨賞. 2006年 IEEE フェロー. 2006年 Best Industry Related Paper Award. 2007年FIT2007論文賞. 2007年 Most Influential Paper over the Decade Award : MVA. 2007年電子情報通信学会フェロー称号授与. 2009 年 MMM2009 Best Paper Award. 2010 年前島密賞. 2012 年紫綬褒章.