Drop and Median Inference による歩行者検出の高精度化

High Accurate Pedestrian Detection by Drop and Median Inference

福井宏†, 山下隆義†, 綿末太郎‡, 山内悠嗣†, 藤吉弘亘†, 村瀬洋††
Hiroshi Fukui† , Takayoshi Yamashita† , Taro Watasue‡ , Yuji Yamauchi†,

Hironobu Fujiyoshi†, Hiroshi Murase††

†: 中部大学, {fhiro@vision., yamashita@, hf@}cs.chubu.ac.jp
‡: とめ研究所
††:名古屋大学

概要:Deep Convolutional Neural Network(CNN)は, 重みフィルタによる畳み込み層を持つ特徴抽出部と全結合層の識別部からなる多層ニューラルネットワークである. CNN は, 高い汎化能力を有しているため, 様々な画像認識問題に利用されている. CNN の汎化性能を向上させるテクニックとして, Dropout がある. Dropout は, 学習時にランダムに選択したユニットの出力を 0 にすることで更新する結合重みを限定し, 汎化性能を向上させることができる. 従来の Dropout は, 学習時におけるランダム性を利用した汎化性能向上に留まっており, 識別処理時に同様の処理は行われていない. そこで, 本研究では, 識別処理において Dropout により結合重みを取り除く手法を導入した Drop and Median Inference(Dn'MI)を提案する. Dn'MI は, 結合重みを Dropout により取り除いた複数のネットワークを構築し, 各ネットワークの応答値の中から中央値を出力する. 歩行者検出の評価実験により, state-of-the-art な CNN による歩行者検出法である, 複雑な構造を持つ Joint Deep と比べ, 提案手法はシンプルな構造で同等の検出精度であることを確認した.

1. はじめに

従来の画像認識は, 研究者や開発者が設計した画像局所特徴量と統計的学習法の組み合わせにより実現されている. 顔検出では, 画像局所特徴量に Haar-like 特徴量[1], 統計的学習法には AdaBoost[2] が用いられている. 歩行者検出では, Histogram of Oriented Gradient(HOG)特徴量[3]と SVM の組み合わせが用いられ, 車載カメラからの歩行者検知による自動緊急ブレーキに利用されている. また, 画像分類の問題では, Scale-Invariant Feature Transform(SIFT)特徴量[4]と Bag-of-Features[5]が用いられてきた. 一方, 2012 年に行われた 1,000 クラス物体認識のコンテスト(ILSVRC)では, Deep Convolutional Neural Network(CNN)[6]を用いた手法が, 従来の画像認識のアプローチより大幅に性能が向上することが報告され, 注目されている[7]. CNN の特長は, 学習過程において識別処理に適した特徴量を自動獲得することができる点である. また, Hinton らは, CNN の結合重みを限定する Dropout や, 高速に学習ができる ReLU を提案し, CNN の学習に利用されている[8]. Dropout は, 学習の際に識別部の結合重みをランダムに取り除いて学習することにより更新する結合重みを限定し, 汎化性能の向上を実現している. 我々は, 従来の Dropout は学習時におけるランダム性を利用した汎化性能向上に留まっており, 識別処理時に同様の処理は行われていない点に着目する.

本稿では, 識別処理において汎化性能を向上するアプローチとして, Dropout の結合重みを取り除くアルゴリズムを導入した Drop and Median Inference(Dn'MI)を提案する. Dn'MI では, Dropout の結合重みを取り除くアルゴリズムを識別処理に導入する. そして, 識別部の結合重みを取り除いた複数のネットワークを構築し, 各ネットワークの応答値から中央値を出力する. これにより, 識別能力の向上が期待できる. 本稿では, 歩行者検出問題を対象とし, 提案手法により従来の CNN より識別能力が向上することを示す.

2. Deep Convolutional Neural Network

CNN の構造は，図 1 のように特徴抽出部と識別部から構成される．特徴抽出部は，重みフィルタを畳み込む層と，得られた特徴マップに対して Pooling をする Pooling 層から構成される．また，識別部では，全てのユニットを全結合する Fully connection 層，最終的な認識結果を出力する出力層から構成される．以下では，CNN の構造の詳細と Dropout を用いた学習について述べる．

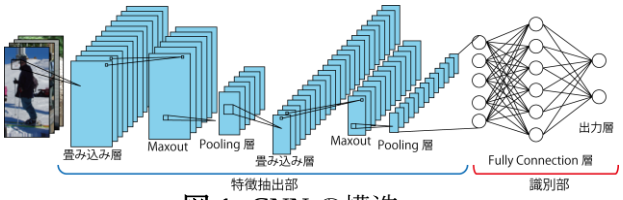


図 1 CNN の構造

2.1. CNN の構造

畳み込み層は，重みフィルタを入力画像に，または特徴マップに畳み込み処理をする層である．畳み込みにより得られた値は，活性化関数に入力し，次の層の特徴マップに入力される．活性化関数には，通常シグモイド関数や ReLU，Maxout が用いられる．畳み込み層で用いる重みフィルタは，従来のニューラルネットワークの学習法である誤差逆伝播法により学習する．

Pooling 層では，入力される特徴マップの小領域から値を出力して新たな特徴マップに変換する処理である．例えば，Max Pooling の場合，特徴マップの小領域に対して最大値を選択し，新たな特徴マップに変換する．Max Pooling の効果として 2 つあり，1 つ目は，Pooling によりユニット数を減らし，更新する結合重みやバイアスを減らすことができる点である．2 つ目は，小領域から応答値を出力するため，位置ずれに対する不変性を獲得することができる点である．

Fully connection 層は，前層のユニットに結合重みがすべて結合され，Fully connection 層の最後の層が出力層となる．出力層の各ユニットの出力は，活性化関数に Softmax 関数を用いる．

2.2. CNN の学習と Dropout

ニューラルネットワークの汎化性能を向上させる学習法の1つとして Dropout がある．Dropout は，ニューラルネットワークの学習において，ランダムに選択したユニットの応答値を 0 にすることで，更新する結合重みを取り除く方法である．ここで，応答値を 0 にするユニットは，各更新処理で異なる．Dropout は一般的に 0.5 の割合でユニットの出力を 0 にする．各更新処理に結合重みを変化させることで近似的なアンサンブル学習となる．

3. 提案手法

本稿では，学習時に Random Dropout，識別時に Dn'MI を導入した手法を提案する．以下に 2 つの提案手法について詳細に述べる．

3.1. Random Dropout による学習

従来の Dropout では，各層の応答値を 0 にするユニットの割合は，各更新処理で一定である．提案する Random Dropout は，応答値を 0 にするユニットの割合を各更新処理にランダムで変化させる．図 2 を例としたとき，更新 1 回目のとき，各層のユニットの削減率は 60%と 30%となっている．更新 2 回目では，乱数を用いて各層のユニットの削減率を更新し，図 2 では，40%と 70%となる．このように各更新処理において，各層のユニットの削減率をランダムに指定することで，汎化性能の向上を実現する．

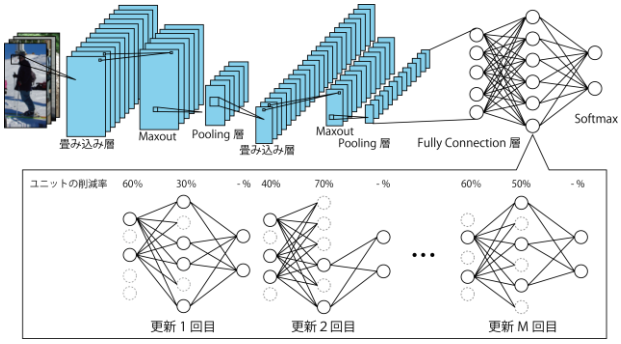


図 2 Random Dropout による学習法

3.2. Drop and Median Inference による識別

Dn'MI は，識別処理において Dropout により結合重みを取り除くアルゴリズムである．層間の結合重みをランダムに取り除いた複数のネットワークを構築し，各ネットワークの応答値から中央値を求めて出力する構造となる．中央値を用いることで，アウトライアとなるユニットの応答値を識別結果として用いないことができる．以下に Dn'MI を用いた識別過程の各ステップについて述べる．

Step1： 特徴マップの生成

まず，入力画像 $\mathbf{I}$ に対して式(1)のようにフィルタ $\mathbf{V}$ を畳み込む．

$$\mathbf{h} = \mathbf{V}^T\mathbf{I} + \mathbf{b} \tag{1}$$

ここで，$\mathbf{b}$ はバイアスを示す．そして，畳み込み後に活性化関数に入力する．活性化関数には Maxout を用いる．Maxout は，式(2) のように K 枚の特徴マッ

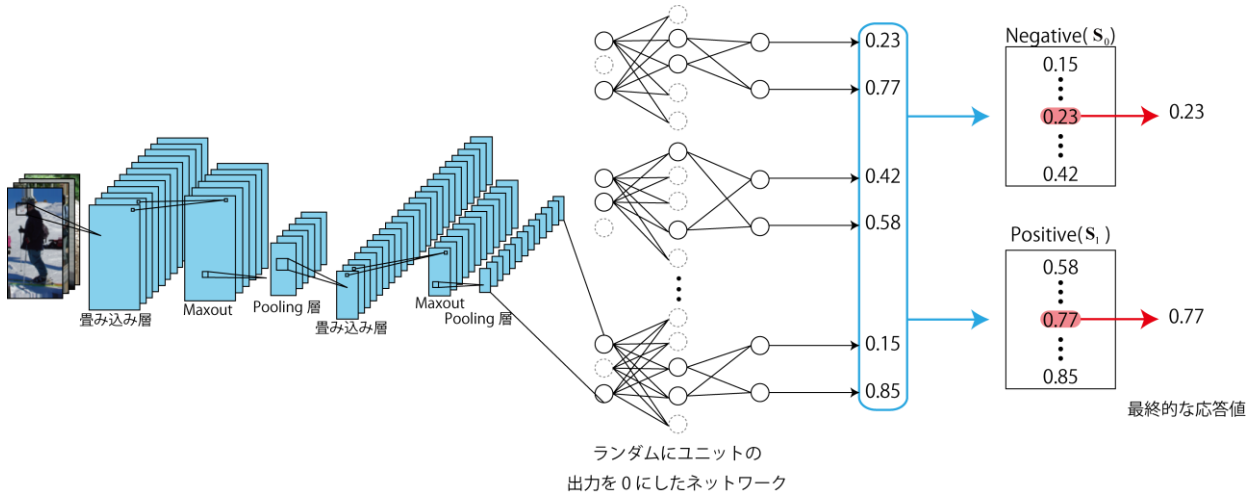


図 3 Dn'MI のアルゴリズム

プのあるユニット i から最大値を選択する.

$$h_i' = \max_{k\in[1,K]} h_{ik} \qquad (2)$$

Maxout により, 特徴マップを統合した後に Pooling する. ここで, Pooling には Max Pooling を用いる. Max Pooling は, 式(3)のように特徴マップの領域 P_i の中から最大値を出力する手法である.

$$h_i'' = \max_{p\in Pi} h_p' \qquad (3)$$

特徴マップは, 畳み込みと Maxout, Pooling を繰り返すことで生成する.

Step2 : 複数のネットワークの構築

識別層では, L 層の Fully connection 層を持つ CNN に対して, ランダムに選択したユニットの出力を 0 にすることで結合重みを取り除く. ここで, Step1 で得られた特徴マップを $\mathbf{x}$ と定義する. ランダムに選択した l 層のあるユニット j の応答値 h_j^l を式(4)のようにして 0 にする.

$$h_j^l = f\left(\mathbf{W}^l\mathbf{x} + b^l\right)\cdot m_j^l \qquad (4)$$

ここで $\mathbf{W}^l$, b^l は l 層目の結合重みとバイアスを示している. m は応答値 h_j^l の出力を 0 と 1 に制御する変数である. 応答値ユニットの応答値を 0 にする場合は m を 0, ユニットの出力を伝播する場合は m を 1 とする. そして, ランダムに選択したユニットの応答値を 0 にした N 個のネットワークを構築する. そして, 各ネットワークの各クラス c の応答値 O_{nc} を式(5)の Softmax 関数により求める.

$$O_{nc} = \frac{\exp\left(\mathbf{W}_j^L\mathbf{h}_j^L + b_j^L\right)}{\sum_{c=0}^{C}\exp\left(\mathbf{W}_c^L\mathbf{h}_c^L + b_c^L\right)} \qquad (5)$$

Step3: 最終的な応答値の算出

Step2 で求めた各ネットワークと各クラスに対する応答値 O_{nc} を用いて最終的な応答値を求める. まず, 各ネットワークの応答値をクラスごとで格納する. このとき, 各ネットワークの各クラスに対する応答値の集合を $\mathbf{S}_c$ とする. Dn'MI の各クラスに対する最終的な応答値は $\mathbf{S}_c$ の中央値 $\mathbf{S}_c^{Median}$ を用いる.

4. 提案手法による歩行者検出

歩行者検出では, 入力画像を網羅的にラスタスキャンし, 得られた検出ウィンドウを識別する. そのため, ラスタスキャンにより発生した膨大な検出ウィンドウを識別処理する必要がある. CNN の場合, 畳み込み層の畳み込み演算に多大な計算コストを要するため, ラスタスキャンにより入力画像 1 枚から多くの検出ウィンドウを対象とすると, リアルタイム処理が不可能となる. そこで, 本研究では, 図 4 のように 2 段階の識別処理によりこの問題を解決する.

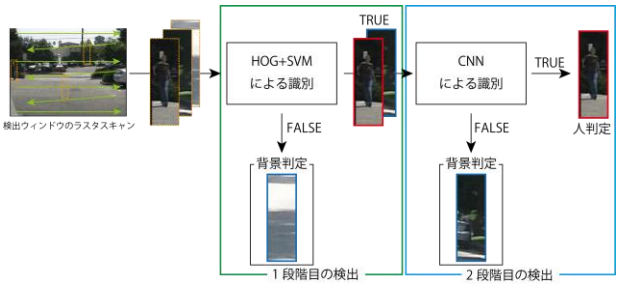


図 4 2 段階処理による歩行者検出

1 段階目で HOG+SVM を用いて歩行者の候補領域の絞り込みをする. そして, 絞り込んだ歩行者領域に対して CNN により最終的に判定する. このように, 2 段階処理により識別することで, CNN を用いた際の識別処理の効率化を実現する.

5. 評価実験

提案手法の有効性を調査するために，評価実験により検出精度を比較する．

5.1. 実験概要

Random Dropout と Dn'MI の効果について評価実験を行う．Random Dropout の評価では，従来の CNN と提案手法の検出精度を比較する．Dn'MI の評価では，各ネットワークの応答値の統合方法と，統合するネットワーク数の変化に対する検出精度を確認する．

比較に用いる手法は，CNN，HOG[3]，HogLbp[9]，LatSvm-V2[10]，VJ[11]，DBN-Isol[12]，ACF[13]，ACF-Caltech[13]，Pls[14]，FPDW[15]，ChuFtrs[16]，CrossTalk[17]，RandomForest[18]，MultiResC[19]，Roerei[20]，MOCO[21]，Joint Deep[22]である．

比較実験で使用する CNN の構造を表 1 に示す．今回の実験では，畳み込み層が 3 層，Fully connection 層が 3 層の計 8 層の CNN を用いる．入力層は，108×36 画素の RGB 画像を入力するため、ユニット数は 11,664 となる．また，出力層には，人と非人の識別問題として，2 クラスのソフトマックス関数を用いる．CNN の学習パラメータを表 2 に示す．ここで，学習係数は結合重みを更新する際に用いる確率的勾配降下法の係数を示している．データセットは，Caltech Pedestrian Dataset を使用する．学習サンプルには，ポジティブサンプル 4,000 枚とネガティブサンプル 10,000 枚のサンプルを Data Augmentation により，ポジティブサンプル 101,808 枚，ネガティブサンプル 200,000 枚に生成したものを使用する．評価には，8,273 枚の評価サンプルを用いる．

表 1 CNN の構造

層数		
1 層目	重みフィルタ	20×9×3
	Max pooling	2
	Maxout	2
2 層目	重みフィルタ	64×5×4
	Max pooling	2
	Maxout	2
3 層目	重みフィルタ	32×6×4
	Max pooling	2
	Maxout	2
4 層目	ユニット数	1,000
	Dropout	あり
5 層目	ユニット数	500
	Dropout	あり
6 層目	ユニット数	100
	Dropout	あり
7 層目	Softmax	2

表 2 CNN の学習パラメータ

学習係数	0.01
バッチサイズ	10
更新回数	100,000 回
学習誤差関数	クロスエントロピー誤差関数

5.2. Random Dropout による精度の変化

Random Dropout による学習の効果について評価する．従来の Dropout を用いて学習した CNN と，Random Dropout を用いて学習した CNN を用いた際の Detection Error Tradeoff(DET)カーブを図 5 に示す．図 5 より，通常の Dropout より Random Dropout の方が，False Positive per Image(FPPI)が 0.1 のとき約 8%向上していることがわかる．

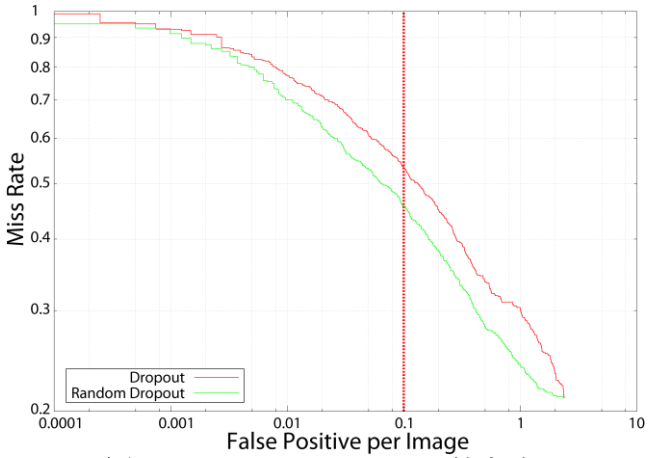


図 5 Random Dropout の比較実験

5.3. ネットワーク数の変化による精度の比較

Dn'MI のネットワーク数を変化させたときの Miss rate を図 6 に示す．図 6 のグラフは，横軸に Dn'MI のネットワーク数、縦軸に False positive per Image(FPPI)が 0.1 のときの Miss rate を示している．今回の実験では，Dn'MI の各ネットワークの応答値の統合方法を中央値・平均値・最大値の 3 パターンの結果に加え，Random Dropout あり，なしを比較する．実験結果から，Random Dropout を導入し統合方法に中央値を使用し，ネットワーク数が 61 のとき，Miss rate が 39.94%で最も精度が良い．

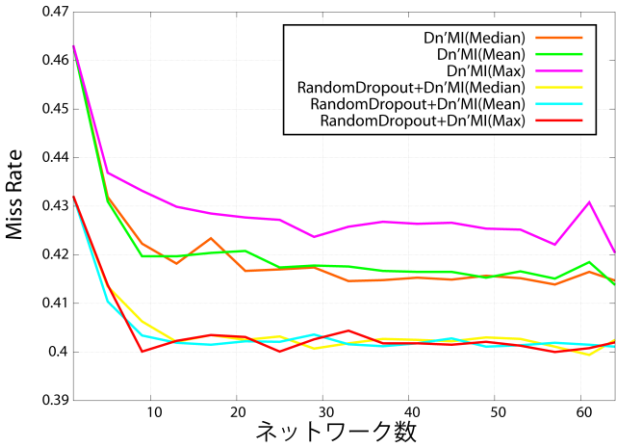


図 6 ネットワーク数と応答値の変化による Miss Rate の比較

5.4. 従来法との精度の比較

従来の歩行者検出法との比較結果を図 7 に示す. FPPI が 0.1 のときに従来の CNN に比べ 10.5%精度が向上した. また, Caltech Pedestrian Dataset でトップの性能を出している Joint Deep と比較する. Joint Deep は, CNN に従来のパーツベースの歩行者検出の考え方を導入しており, 1 段階目の CNN により各パーツのスコアを推定し, その結果を用いて 2 段階目のネットワークにより歩行者と背景を識別する手法である. 図 7 の比較結果より, Joint Deep とほぼ同等の性能であることが確認できる.

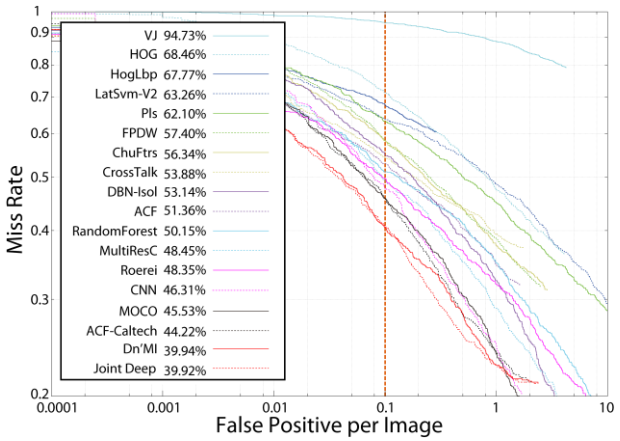
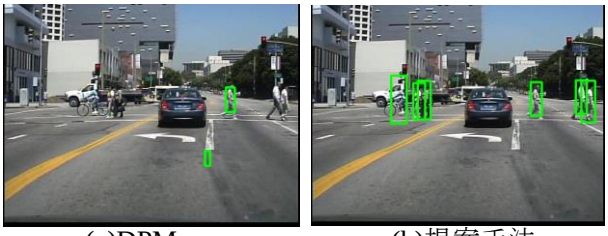


図 7 他手法との比較

図 8(a)に従来の DPM の歩行者の検出結果を示しており, 図 8(b)には提案手法である, Random Dropout と Dn'MI を導入したときの歩行者の検出結果を示している. 検出例より, 提案手法は, 一般的に利用されている DPM と比べ, 歩行者検出能力が高いことがわかる. また, このとき 1 検出ウィンドウあたりの検出時間は約 50 ミリ秒であった.



(a)DPM (b)提案手法

図 8 検出結果の比較

6. おわりに

本稿では, CNN における Dropout のアルゴリズムをベースとした汎化性能向上を目的とした手法を提案した. Dn'MI による識別処理では, 検出過程でランダムに選択したユニットの出力を 0 にした複数のネットワークを用いることで検出精度を向上させた. また, Dropout の割合を学習の更新回数ごとにランダムで決定する Random Dropout を学習に導入することで, 汎化性能を向上させた. 今後の課題として, リアルタイムで歩行者検出を実現するために CNN の高速化が挙げられる.

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] P.Viola, and M.Jones : Rapid object detection using a boosted cascade of simple features, Computer Vision and Pattern Recognition, 2001.

[2] Y.Freund, and R.E.Schapire : A decisiontheoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, No1, Vol.55, pp.119-139, 1997.

[3] N.Dalal, and B.Triggs : Histgram of oriented gradients for human detection, Computer Vision and Pattern Recognition, 2005.

[4] D.G.Lowe:Distinctive image features from scale-invariant keypoints, IJCV,60, 2, pp. 91-110, 2004

[5] G.Csurka, C. R. Dance, L. Fan, J. Willamowski and C. Bray : Visual categorization with bags of keypoints, ECCV International Workshop on Statistical Learning in Computer Vision, pp.1-22, 2004.

[6] Y.Lecun, B.Boser, J.S.Denker, D.Henderson, R.E.Howard, W.Hubbard, and L.D.Jackel: Backpropagation applied to handwritten zip code recognition, Neural Computation, vol.1, pp.541-551, 1989.

[7] A. Krizhevsky, S. Ilva, and G. E. Hinton : ImageNet Classification with Deep Convolutional Neural Network, Advances in Neural Information Processing System 25,

pp.1097-1105, 2012.
[8] G.E.Hinton, N.Srivastava, A.Krizhevsky, I.Sutskever, and R.Salakhutdinov：Improving neural networks by preventing co-adaptation of feature detectors,Clinical Orthopaedics and Related Research, 2012.
[9] X.Wang, T.X.Han, and S.Yan：An HOG-LBP Human Detection with Partial Occlusion, International Conference on Computer Vision, 2009.
[10] P.Felzenzwalb, R.Girshick, D.McAllester, and D.Ramanan：Object detection with discriminatively trained part based models, Pattern Analysis and Machine Intelligence, Vol.32, pp.1627-1645, 2010.
[11] P.Viola, and M.Jones：Robust Real-Time Face Detection, Computer Vision and Pattern Recognition, 2004.
[12] W.Ouyang, and X.Wang：A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling , Computer Vision and Pattern Recognition, 2012.
[13] P.Dallár, R.Appel, S.Belongic, and P.Perona：Fast Feature Pyramids for Object Detection, Pattern Analysis and Machine Intelligence, 2014.
[14] W.R.Schwartz, A.Kembhavi, D.Harwood, and L.S.Davis：Human Detection Using Partial Least Squares Analysis, International Conference on Computer Vision, 2009.
[15] P.Dollar, S.Belongie, and P.Perona：The Fastest Pedestrian Detectio in the West, British Machine Vision Conference, 2010.
[16] P.Dollár, Z.Tu, P.Perona, and S.Belongie：Integral Channel Feature, British Machine Vision Conference, 2009.
[17] P.Dollár, R.Appel, and W.Kienzle：Crosstalk Cascades for Frame-Rate Pedestrian Detection, European Conference on Computer Vision, 2012.
[18] J.Marin, D.Vazquez, A.Lopez, J.Amores, and B.Leibe：Random Forests of Local Experts for Pedestrian Detection, International Conference on Computer Vision, 2013.
[19] D.Park, D.Ramanan, and C.Fowlkes：Multi Resolution models for Object Detection, European Conference on Computer Vision, 2010.
[20] R.Benenson, M.Mathias, T.Tuytelaars, and L.V.Gool：Seeking the Stroungest Rigid Detector, Computer Vision and Pattern Recognition, 2013.
[21] G.Chen, Y.Ding, J.Xiao, and T.Han：Detection Evolution with Multi-order Contextural Co-occurrence,Computer Vision and Pattern Recognition, 2013.
[22] W.Ouyang, and X.Wang：Joint Deep Learning for Pedestrian Detection, The IEEE International Conference on Computer Vision, 2013.

福井宏：2014 年中部大学工学部情報工学科卒, 現在同大学大学院工学研究科情報工学専攻博士前期課程在学中, 画像を用いた歩行者検出の研究に従事.

山下隆義：2002 年奈良先端科学技術大学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, 2009 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, 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, 2009 年画像センシングシンポジウム高木賞, 2013 年電子情報通信学会情報・システムソサエティ賞, 2013 年電子情報通信学会 PRMU 研究会研究推奨賞, 2014 年画像センシングシンポジウム オーディエンス賞.

綿末太郎：2002 年大阪大学大学院博士前期課程修了, 2004 年 NPO 国際レスキューシステム研究機構勤務, 2005 年神戸大学大学院経済学助教, 2007 年(株)とめ研究所

山内悠嗣：2012 年中部大学大学院博士後期課程修了. 2010 年独立行政法人日本学術振興会特別研究員 DC. 2012 年中部大学院博士研究員. 2014 年中部大学助手. コンピュータビジョン, パターン認識の研究に従事.

藤吉弘亘：1997 年中部大学大学院博士後期課程修了. 1997～2000 年米カーネギーメロン大学ロボット工学研究所 Postdoctoral Fellow. 2000 年中部大学講師. 2004 年より同大学教授. 2005～2006 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, 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 コンピュータビジョンとイメージメディア 優秀論文賞, 2009 年山下記念研究賞, 2010・2013 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. 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.