Vol.2015-CVIM-195 No.59
2015/1/23

Deep Convolutional Neural Network による全方位画像からの人検出

浅沼 仁[1,a)]　川本 一彦[2,b)]　岡本 一志[3,c)]

概要：全方位画像では位置による物体の見えの変化が大きく，従来の見えに基づく特徴による分類では人の識別が難しい．本研究では，Deep Convolutional Neural Network を全方位画像からの人検出に応用する．Deep Learning では学習に大量の学習サンプルが必要となるが，手作業で作成することは時間的な問題から現実的でない．少ない学習サンプルから大量の学習サンプルを生成し，生成された学習サンプルでも学習が行えることを示す．また，実環境下で HOG 画像特徴量と Real AdaBoost を用いた人検出法と比較し，識別率が向上することを示す．

キーワード：Deep Learning，人検出，全方位画像

1. はじめに

本研究では，Deep Convolutional Neural Network（ConvNets）[5] を用いた全方位画像からの高精度な人検出手法を提案する．画像からの人検出は，監視や行動分析など幅広い応用をもち，とくに全方位カメラは周囲 360° を一度に撮影できるため広く利用されている [6] [10] [11]．しかし，全方位画像は大きな歪みを持ち，さらに広い視野角を持つため，位置による物体の見えの変化が大きく，自動的な人検出には必ずしも適していない．そのため，従来手法は，カメラ校正 [4] により通常のカメラと同様な歪みのない画像に変換したうえで人検出を適用することが多い [12]．一方，Deep Learning では，識別に有効な特徴量を学習により獲得することができるため，歪みや位置による物体の見えの変化も表現する画像特徴量を獲得することが期待できる．すなわち，画像の見えに基づいて設計された従来の特徴量では識別が難しい状況であっても，事前カメラ校正などが不要で，獲得した特徴量により識別を行うことができる．実際に Deep Learning の 1 つである ConvNets は歩行者検出 [9]，手形状領域の抽出 [13] など多くの場面で利用されている．

一般に，Deep Learning では学習を十分に行うために大量の学習サンプルが必要となるが，手作業で収集することは時間的な問題から現実的でない．本研究では，少量の学習サンプルから大量の学習サンプルを自動的に生成し，生成された学習サンプルでも学習が行えることを示す．実際に監視目的のための設置されている全方位カメラを用いて，人検出のための標準的な手法である HOG 画像特徴量 [2] と Real AdaBoost [8] を組み合わせた結果と比較し，提案手法の識別率が向上することを示す．

2. 全方位画像からの人検出

全方位画像は大きな歪みを持ち，位置による物体の見えの変化が大きい．そこで，比較的向き，歪みによる見えの変化が小さい頭部に着目する．頭部を含む画像をポジティブクラス，含まない背景画像をネガティブクラスとした 2 クラス分類器としてネットワークを構成する．少量のサンプルの変形と背景画像の合成により大量の学習サンプルを作成する．実画像から生成した学習サンプルを用いてネットワークの学習を行う．ネットワークの学習には確率的勾配降下法を用いる．学習したネットワークを用いて画像分類を行い，頭部の検出を行う．

2.1 学習サンプルの作成

Deep Learning では学習を行うために大量の学習サンプルが必要となる．そこで，表 1 に示す並進，スケーリング，

[1] 千葉大学大学院融合科学研究科
Graduate School of Advanced Integration Science, Chiba University
[2] 千葉大学統合情報センター
Institute of Management and Information Technologies, Chiba University
[3] 千葉大学アカデミック・リンク・センター
Academic Link Center, Chiba University
[a)] asanuma0369@chiba-u.jp
[b)] kawa@faculty.chiba-u.jp
[c)] okamoto.kazushi@chiba-u.jp

表 1　サンプルの変形パラメータ

変形	変形範囲
並進	±5
スケーリング	±20%
回転	±180°
輝度変化	±20

表 2　ネットワークの構成

層	種類	カーネルサイズ，出力数
入力	画像	40 × 40
1	畳み込み	5 × 5，32
2	プーリング	2 × 2
3	maxout	4
4	畳み込み	5 × 5，32
5	プーリング	2 × 2
6	maxout	4
出力	Classify	2

回転といった変形を適用し，少量のサンプルから大量の学習サンプルを生成する．変形パラメータは表 1 の範囲からランダムに与える．全方位カメラは通常のカメラと異なり空間の中心に設置されることが多いため，全方位画像では対象が最大で ±180° 回転して写る．そのため，回転の変形範囲も ±180° と大きな値となる．ネガティブサンプルは実環境の背景画像から切り出して作成する．ポジティブサンプルの作成手順を以下に示す．

(1) 透過情報を持つサンプル（前景画像）を作成
(2) 前景画像に並進，回転，スケーリングを適用
(3) 前景画像とネガティブサンプルを透過情報を元に合成
(4) 画像全体に輝度変化を適用

2.2　ネットワークの構成

ネットワークの構成を表 2 に示す．ネットワークは入力層，畳み込み層，プーリング層，出力層からなる．本研究では入力に 2.1. の手法で作成した 40 × 40 pixel の学習サンプルを用いる．

2.2.1　畳み込み層

畳み込みは近接画素とのみ結合を行うことで，局所的な応答を表現する．本研究では，畳み込み層の活性化関数として maxout [3] を用いる．maxout は，

$$h'_i = \max_{j \in [1,k]} h_{ij} \tag{1}$$

のように k 個の特徴マップから各画素の最大値を出力マップの画素値とする手法であり，シグモイド関数や打ち切り線形関数 [7] より高い表現力をもつ．

2.2.2　プーリング層

プーリングはパターンの幾何学的変動，照明条件などに対して不変な特徴を取り出すことを目的とする．本研究では，プーリングとして max pooling を用いる．max pooling とは，

図 1　アクティブ・ラーニング・スペース

$$h'_i = \max_{j \in K_i} h_j \tag{2}$$

のようにプーリングの対象領域 K に含まれる各画素 h_i の最大値を出力とする手法である．max pooling を用いることで汎化性を向上させることができる [1] .

2.2.3　出力層

出力層では，ポジティブ，ネガティブそれぞれのクラスの確率を求める．クラス y^i の確率 $P(y^i)$ は， softmax によって

$$P(y^i) = \frac{\exp(h_i)}{\sum_j^2 \exp(h_j)} \tag{3}$$

のように求める．

3.　実環境下での人検出

実際に監視目的のための設置されている全方位カメラの画像を用いて学習サンプルを作成し，ネットワークの学習を行う．ネットワークの構成，学習には Deep Learning ライブラリ Pylearn2 [*1] を用いる．学習には Minibatch-SGD を用い，更新回数は 1 万回とする．学習結果を用いて人検出を行い，HOG 画像特徴量と Real AdaBoost を組み合わせた結果と比較し，提案手法の識別率が向上することを示す．

3.1　実験環境

実験には千葉大学附属図書館 2 階のアクティブ・ラーニング・スペース [*2] をに設置されている監視カメラの画像を用いる．アクティブ・ラーニング・スペースの全方位画像の例を図 1 に示す．実験データの詳細を表 3 に示す．

[*1] http://deeplearning.net/software/pylearn2/
[*2] http://alc.chiba-u.jp

表 3　実験データの詳細

内容	学習サンプル作成用	評価用
撮影期間	2013 年 7 月	2013 年 8 月
画像枚数	200	50

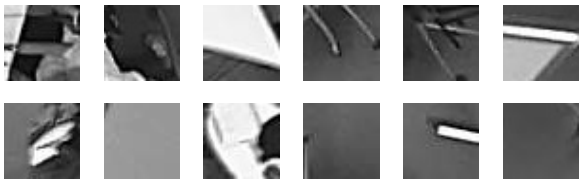

図 2　ネガティブサンプル

図 3　前景画像

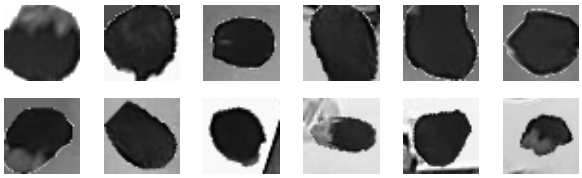

図 4　ポジティブサンプル

3.2　学習サンプルの作成

表 3 の学習サンプル作成用画像からランダムな位置を切り出し，ネガティブサンプル（背景画像）とする．ネガティブサンプルの例を図 2 に示す．画像から頭部を切り出し，透過情報を付加した画像を前景画像とする．前景画像は異なる頭部から作成するため，それぞれ個人差がある．前景画像は 50 枚作成する．前景画像の例を図 3 に示す．

前景画像と背景画像を元に，2.1. の手法でポジティブサンプルを作成する．ポジティブサンプルの例を図 4 に示す．ポジティブサンプル，ネガティブサンプルはそれぞれ 2 万枚作成し，学習に用いる．

3.3　人検出

表 3 の評価用画像から検出窓で画像の切り出しを行う．切り出した画像に対して ConvNets で分類を行うことで，人検出処理を行う．検出窓と，検出窓と同サイズの頭部の正解領域の重なりが 60% 以上のとき，正しく検出できていると判定する．

3.4　実験結果

提案手法と，HOG 画像特徴量と Real AdaBoost による手法で，評価用画像からそれぞれ人検出を行う．評価用の 50 枚の画像には，合計で 549 人が写っている．それぞれの手法で人検出を行った結果の DET 曲線を図 5 に示す．提

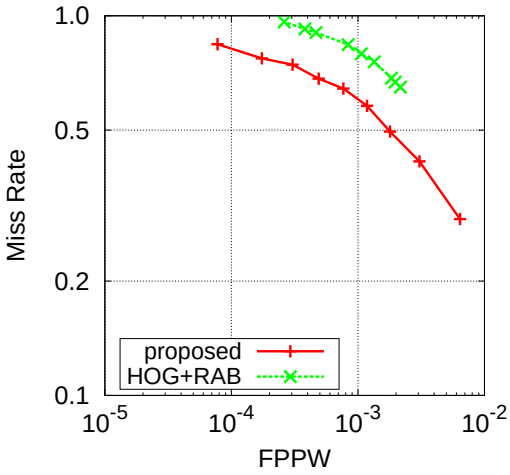

図 5　DET 曲線

図 6　提案手法による人検出結果の例：矩形が検出位置

案手法による検出結果例を図 6 に示す．

3.5　考察

図 5 より，提案手法は HOG 画像特徴量と Real AdaBoost による手法に比べて性能が優れていることがわかる．HOG+RAB では誤検出率が 0.18% の際の未検出率が 68.4% であるのに対し，提案手法では誤検出率が 0.18% の際の未検出率は 49.5% である．これは，HOG 画像特徴量では対象の見えの大きな変化を考えていないため，見えの変化が大きい全方位画像では識別が難しいためである．一方で提案手法では，学習によって獲得した特徴量により，HOG 画像特徴量より高精度な識別を行うことができている．また，ネットワークの学習に基本となるサンプル（前景画像）を 50 枚しか用いていないにもかかわらず，図 6 のように全方位画像からの人検出が行えている．

提案手法では重なりなどで隠れが生じている頭部は未検出となっている．2.1. では頭部の隠れを考慮していないた

め，隠れを含む頭部の検出を行うためには学習サンプルの作成に隠れの影響を加える必要がある．

4. おわりに

本研究では，ConvNets を用いた全方位画像からの高精度な人検出手法を提案している．50 枚の学習サンプルから 2 万枚の学習サンプルを自動的に生成し，生成した学習サンプルで ConvNets の学習を行った．実環境下の見えの変化が大きい全方位画像から，ConvNets によって自動的に特徴を獲得し，カメラ校正などを無しに人検出を行った．人検出のための標準的な手法である HOG 画像特徴量と Real AdaBoost を組み合わせた手法による結果と比較し，提案手法の識別率が向上することを示した．

謝辞

本研究は JSPS 科研費 25330186 の助成を受けたものです．

参考文献

[1] Boureau, Y.-L., Ponce, J. and LeCun, Y.: A theoretical analysis of feature pooling in visual recognition, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 111–118 (2010).
[2] Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection, *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1, IEEE, pp. 886–893 (2005).
[3] Goodfellow, I., Warde-farley, D., Mirza, M., Courville, A. and Bengio, Y.: Maxout Networks, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1319–1327 (2013).
[4] Kanatani, K.: Calibration of Ultra-Wide Fisheye Lens Cameras by Eigenvalue Minimization, *Pattern Analysis and Machine Intelligence*, Vol. 35, No. 4, pp. 813–822 (2013).
[5] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (1998).
[6] Miaou, S.-G., Sung, P.-H. and Huang, C.-Y.: A customized human fall detection system using omni-camera images and personal information, *Distributed Diagnosis and Home Healthcare, 2006. D2H2. 1st Transdisciplinary Conference on*, IEEE, pp. 39–42 (2006).
[7] Nair, V. and Hinton, G. E.: Rectified linear units improve restricted boltzmann machines, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814 (2010).
[8] Schapire, R. E. and Singer, Y.: Improved boosting algorithms using confidence-rated predictions, *Machine learning*, Vol. 37, No. 3, pp. 297–336 (1999).
[9] Sermanet, P., Kavukcuoglu, K., Chintala, S. and LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning, *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, IEEE, pp. 3626–3633 (2013).
[10] Yuan, P.-H., Yang, K.-F. and Tsai, W.-H.: Real-time security monitoring around a video surveillance vehicle with a pair of two-camera omni-imaging devices, *Vehicular Technology, IEEE Transactions on*, Vol. 60, No. 8, pp. 3603–3614 (2011).
[11] 森田真司，山澤一誠，寺沢征彦，横矢直和：全方位画像センサを用いたネットワーク対応型遠隔監視システム，電子情報通信学会論文誌 D， Vol. 88, No. 5, pp. 864–875 (2005).
[12] 窪田　進，丸山昌之，伊久美智則：複数の全方位カメラによる人物動線計測システム，東芝レビュー，Vol. 63, No. 10, pp. 44–47 (2008).
[13] 山下隆義，綿末太郎，山内悠嗣，藤吉弘亘：Deep Convolutional Neural Network による手形状領域の抽出，第 20 回画像センシングシンポジウム， Vol. 20, No. IS2, pp. 1–6 (2014).