特徴学習による全方位画像からの人検出†

浅沼 仁 #*1・岡本 一志 #*2・川本 一彦 #*3

カメラパラメータを推定することなく、全方位画像という歪みや位置による見えの変化のある画像に対する人検出器を提案する。Deep Convolutional Neural Networkをベースに実現し、さらに、少量の学習サンプルに対して並進・スケーリング・回転・輝度変化の変形を適用することで大量の学習サンプルを生成する手法も提案している。実際に設置されている全方位カメラの画像を用いた人検出実験では、HOGとReal AdaBoostの組み合わせによる人検出器が誤検出率0.001で未検出率77.5%であるのに対し、提案手法が誤検出率0.001における未検出率が28.2%となることを確認している。提案する学習サンプル生成法が精度向上に寄与することや可視化した特徴マップの検出精度との関連性も検証している。これによりカメラパラメータの推定が難しい状況でも全方位画像からの人検出の精度向上を実現する。

キーワード：全方位画像，カメラ校正，特徴学習，ニューラルネットワーク，人検出

1. はじめに

360度の視野を持つ全方位カメラは、通常のカメラに比べ少ない台数で空間全体を記録できることから、防犯・防災などを目的とした定点観測カメラシステムにおいて広く活用されている。定点での記録用途以外にも、移動型ロボットへの搭載や個人向けの持ち運び可能な製品の研究開発も進められており、360度の視野を記録した全方位画像を扱う機会の増加が今後もさらに見込まれる。このようなシステムや製品において、画像中の人物の位置を特定する人検出は、人物の数え上げや人物追跡、行動認識などの応用に欠かせない要素技術である。全方位カメラは撮影原理の違いにより魚眼レンズ型やミラー型、マルチカメラ型に分類され、得られる全方位画像には大きな歪みやエリアによって解像度が不均一といった画像認識を行うにあたっての課題がある。一般に、人検出は、画像特徴量の抽出と特徴空間上での識別処理により実現されるが、全方位画像を対象とした場合、画像中の歪みや位置による見えの変化にどのように対応するかが重要と

なる。

SIFT(Scale-Invariant Feature Transform)[1]やHOG(Histogram of Oriented Gradients)[2]は、人検出における代表的な画像特徴量であり、全方位画像からの画像特徴量の抽出に関する既存研究では、これらをベースとした手法が提案されている。これまで、球面カメラモデル上で勾配を計算する手法[3, 4, 5, 6, 7, 8, 9]や歪み耐性を持ったSIFT[10]などが提案されている。これらの手法では、カメラモデルによる全方位画像の球面座標系への投影や歪み補正などを考慮することから、センサの中心座標や焦点距離などのカメラ固有の情報が欠かせない。したがって、このようなカメラパラメータは、格子パターン画像などを用いたカメラ校正[11, 12, 13]により推定できていることが前提となる。一方で、校正対象のカメラが多い場合やカメラ自体が遠隔地にある場合、撮影したカメラそのものを特定できない場合など校正作業によりカメラパラメータを推定することが現実的でない状況も存在する。

本研究は、カメラパラメータの推定を行うことなく、全方位画像からの人検出を高精度化することを目的とする。全方位画像は通常の画像に比べて、歪みや解像度の不均一さ、見えの大きな変化が起こるため、これらを踏まえた特徴量を、カメラパラメータ無しにユーザがどのように設計すればよいかの明確な方法論はまだ定まっていない。本研究では、学習データを用いて入力画像を特徴量に変換する関数を機械的に発見する特徴学習の枠組みに注目し、全方位画像からの人検出の高精度化を図る。具体的には、画像認識での特徴学習において、有力な手法のひとつとして考えられているDeep Convolutional Neural Network(ConvNets)

† Feature Learning Based Human Detection for Omnidirectional Images

Hitoshi ASANUMA, Kazushi OKAMOTO and Kazuhiko KAWAMOTO

These authors contributed equally to this work.

*1 大日本印刷株式会社

Dai Nippon Printing Co., Ltd.

*2 電気通信大学 大学院情報理工学研究科 総合情報学専攻
Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications

*3 千葉大学統合情報センター

Institute of Management and Information Technology, Chiba University

[14]をベースに実現する。ConvNetsはネオコグニトロン[15]をルーツに、学習則に誤差逆伝搬法を取り入れたモデルとなっている。ネオコグニトロンには変形にロバストな特徴があることが知られており、この特徴は本研究が対象とする全方位画像にとって欠かせない要素である。ConvNetsも同様の特徴を有していることが期待でき、さらに、pylearn2[16]などの実装環境が多く研究者により整備されている点は実用上重要と考えたため、本研究ではConvNetsに着目する。

ConvNetsでは畳み込み処理とプーリング処理をどのように行うかによってネットワークの構造が決まり、その構造により計算の効率や識別結果に差が出ることになる[17]。ネットワークの構造を決めるハイパーパラメータには、中間層の数や各層での特徴マップの数、カーネルのサイズ、プーリングの比率などがある。本研究では、全方位画像からの人検出にConvNetsを適用するにあたって、これらのパラメータをどのように定めれば良いかを明らかにする。また、ConvNets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り、特に、ポジティブサンプル(正例)を手手で集めることはコストが高い作業である。本研究では、位置ずれた学習サンプルを生成する手法[18]や回転、スケーリングを適用する手法[19]などの組み合わせにより、人手で取得した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する手法についても提案する。

全方位画像からの人検出用に本研究で提案するConvNetsのネットワーク構成と学習サンプルの生成法について、その性能評価を目的に、実際に設置されている全方位カメラの画像を用いた人検出実験を行う。人検出実験では、頭部を含む画像をポジティブクラス、それ以外の画像をネガティブクラスとした2クラス識別器としてConvNetsのネットワークを構築する。ネットワークの性能は、畳み込むカーネルサイズや特徴マップ数などに依存するため、カーネルサイズ・特徴マップ数・maxoutの有無をパラメータとし、ネットワークの学習時間・検出窓あたりの誤検出率・1枚の全方位画像にいる人の数に対する未検出率の観点で分析する。その際、物体検出の標準的な手法であるHOG画像特徴量とReal AdaBoost[20]の組み合わせも比較する。

本論文の構成は次の通りである。2では、本研究で用いるConvNetsについて説明する。3では、全方位画像からの人検出用に本研究で提案するConvNetsのネットワーク構成と学習サンプルの生成法について述べる。4では、実際に設置している全方位カメラから

の画像を用いた人検出実験を行い、得られた人検出結果について、ネットワークの学習時間・検出窓あたりの誤検出率・1枚の全方位画像にいる人の数に対する未検出率の観点で分析する。

2. Deep Convolutional Neural Network

Deep Learningとは多層ニューラルネットワークによる学習器の総称であり[21]、その本質は、学習データから識別に有効な特徴量を機械的に発見するところにある。これまで、Deep Belief Network[22]などの枠組みが提案されているが、本研究では、その中でも有力な手法のひとつとして考えられているDeep Convolutional Neural Network(ConvNets)に注目する。特に、近年の物体検出や物体認識に関するコンペティションを通じてConvNetsが優れた認識性能を有していることが報告されている[23]。一般に、ニューラルネットワークを多層にすると過学習が起こりやすく、誤差情報が分散し下層まで伝播されにくいことが知られている。その一方で、ConvNetsでは各層間の結合を局所的にし、過疎性をもたせることで、多層のネットワークでも学習を実現している。

ConvNetsは、2次元の画像を初期入力とし、重みカーネルの畳み込み処理とその結果に対するプーリング処理を繰り返す多層ニューラルネットワークである。ConvNetsによる識別は、(1)重みカーネルの畳み込みによる特徴マップの生成、(2)プーリングによる特徴マップの縮小、(3)畳み込みとプーリングの繰り返しによる特徴抽出、(4)出力層での識別といった流れで処理される。これらの処理の概要を図1に示す。

図1 ConvNetsの処理の流れ

2.1 畳み込み層

畳み込み層では、前層からの入力に対して重みカーネルの畳み込み処理を行う。このとき、前層からの入力を入力マップと呼び、初期入力は画像そのものとなる。入力マップ全体にラスタ走査で重みカーネルの畳み込みを行うことにより特徴マップを得る。サイズ $n_x \times n_y$ の入力マップ h_c に対する、サイズ $n_w \times n_w$ のカーネル w の畳み込みを

$$f_c = h_c * w \quad (1)$$

と記述すると、畳み込み層の特徴マップ f_c のサイズは、 $(n_x - n_w + 1) \times (n_y - n_w + 1)$ となる。畳み込み層では L 個の重みカーネル $w^1, w^2, \dots, w^L$ を持ち、カーネル毎に独立の特徴マップ $f_c^1, f_c^2, \dots, f_c^L$ を出力する。ConvNets では、カーネルの範囲毎に局所的に2つの層が結合されており、この畳み込み処理により各層間の結合を局所的にし、ネットワークの過疎性を実現している。

2.2 プーリング層

プーリングとは、幾何変化や照明条件の変化などに対する不変性を獲得することを目的に、抽出した特徴マップから識別に不要な情報を省略し、識別に有効な情報を保ったまま新たな表現に変換することである[24]。情報を削減するため、学習するパラメータ数も減らすことにもつながる。プーリング層では、入力マップ毎に独立にプーリング処理を行い、各マップを平均や最大演算などにより縮小する。具体的には、サイズ $n_x \times n_y$ の入力マップ h_p に対して、サイズ $n_k \times n_k$ の小領域 K 毎に特徴量(代表値)を得る。小領域を入力マップ内で互いに重複しないようにとれば、プーリング層の特徴マップ f_p のサイズ $n'_x \times n'_y$ は、

$$n'_x = \frac{n_x}{n_k}, \quad n'_y = \frac{n_y}{n_k} \quad (2)$$

となる。

小領域毎の特徴量の計算方法には、average pooling, max pooling, L_p pooling などが存在する。average poolingは小領域 K の応答の平均を、max poolingは小領域 K での応答の最大値を特徴量とする方法である。また、 L_p poolingはaverage poolingとmax poolingの中間の方法となる。本研究ではmax poolingの適用を行うが、その定義は、

$$\max_{k \in K} k \quad (3)$$

である。

2.3 出力層

ConvNets では、重みカーネルの畳み込み処理とプーリング処理を繰り返すことにより特徴を抽出し、最終的に得られた特徴マップを入力に出力層での識別を行う。出力層の前に、全マップを重み付きで結合した全結合層を扶むこともある。出力層のノード数は識別を行いたいクラス数となり、 M クラスの識別の場合、出力層のノードからの出力を $y_1, y_2, \dots, y_M$ とすると、 i 番目のクラスの確率 $P(y_i)$ はsoftmax関数

$$P(y_i) = \frac{\exp(y_i)}{\sum_{j=1}^M \exp(y_j)} \quad (4)$$

によって求められる。

2.4 ConvNets の学習

ConvNets の学習では、1つの学習サンプルを用いてパラメータを更新する確率的勾配降下法が用いられている[25]。確率的勾配降下法により更新されるパラメータは、各層の結合重みとバイアスである。ConvNets では、結合重みは畳み込み層での重みカーネルの要素に相当する。確率的勾配降下法では、誤差逆伝播法[25]で求められた誤差の勾配に基づき、誤差が最小となるようなパラメータの更新を行う。誤差逆伝播法により誤差の勾配を求める際、確率的勾配降下法では学習サンプルを1つずつ与えてパラメータの更新を行うことから、学習サンプルの数が多いと繰り返し回数が膨大となる。そのため本研究では、学習サンプルの数が多い場合でも効率的にパラメータの更新が可能なMini-batch学習法を用いることとする。Mini-batch学習法では、学習サンプルを小さな部分集合に分割し、その部分集合毎にパラメータを更新する。

3. Deep Convolutional Neural Network による全方位画像用人検出器の設計

本研究では、カメラパラメータを推定することなく、全方位画像という歪みや位置による見えの変化のある画像に対する人検出器を提案する。ここでは全方位画像からの人検出用に本研究で提案するConvNetsのネットワーク構成と学習サンプルとしての正例および負例画像の生成法について述べる。

3.1 ConvNets のネットワーク構成

ConvNets では畳み込み処理およびプーリング処理をどのように行うかによってネットワークの構造が決

表1 全方位画像からの人検出用に本研究で提案する ConvNets のネットワーク構成

層	処理	ハイパーパラメータ
入力層	画像	40×40 ピクセル
中間層 第1層	畳み込み プーリング	カーネルサイズ7×7, 特徴マップ数 16 小領域サイズ2×2
中間層 第2層	maxout 畳み込み プーリング	4 特徴マップ毎 カーネルサイズ7×7, 特徴マップ数 16 小領域サイズ2×2
出力層	maxout 識別	4 特徴マップ毎 ノード数 2

まり、計算の効率や認識精度はその構造に依存する[17]。ネットワークの構造を決めるハイパーパラメータには、畳み込み層とプーリング層の層数や各層での特徴マップの数、カーネルのサイズ、プーリングの比率などがある。全方位画像からの人検出用に本研究で提案するConvNets のネットワーク構成を表1に示す。ネットワークは入力層・中間層第1層・中間層第2層・出力層から成り、畳み込みとプーリング、maxout の処理を組にした中間層を2層重ねることで画像特徴量を特徴マップとして抽出する。

3.1.1 畳み込み処理の設計

畳み込みは近接画素とのみ線形結合を行うことで、局所的な応答を表現する。本研究では全方位画像を対象とするため、歪みなどの変形が大きく、想定される様々な背景も同様に変形の影響を受ける。したがって、その影響により本来は検出すべきではない領域を検出対象として認識してしまう誤認識について考慮する必要がある。カーネルサイズは局所応答をどの範囲で評価するかを意味しており、カーネルサイズが小さすぎると、変形によって多発する小さな歪みに対して応答が出てしまい背景の誤認識に繋がると考える。そこで、背景の誤認識を避けるためカーネルサイズをできるだけ大きくとることとする。その一方で、Xeon プロセッサを2基搭載した高性能な計算機を用いても7×7のカーネルサイズで学習に3日かかり、9×9のカーネルサイズではより多くの学習時間が必要となることから、本研究では学習にかかる計算コストを考慮しカーネルサイズを7×7としている。また、畳み込み処理の活性化関数には、

$$g(x) = \max(0, x) \quad (5)$$

で表現される打ち切り線形関数[26]を用いる。正負それぞれの範囲で線形であり、非線形関数であるシグモイド関数や双曲線正接関数と比較して学習が早いこと

が特徴である。また、学習において誤差勾配が消失しないため、画像認識において広く利用されている[18]。

3.1.2 プーリング処理の設計

プーリングはパターンの幾何変化や照明条件の変化などに対して不変な特徴を取り出すことを目的としている。全方位画像による歪みや位置による変形は畳み込み層の重みカーネルによって表現できることから、プーリング処理では全方位画像特有の要因を考える必要がないと考える。そのため本研究では、通常の画像に対する場合と同様に、微小な移動などの幾何変化を考慮し、プーリング対象の小領域を2×2、プーリング比率を4:1とする。これは、小領域の4つの値から1つの代表値を抽出することに相当する。また、ネットワークの汎化性向上のため、プーリング手法として小領域内の最大値を応答とする max pooling を用いる[24]。

3.1.3 maxout の適用

標準的なConvNets では、プーリング処理の出力を全て次の中間層の入力としている。その一方で、maxout[27]と呼ばれる処理をプーリング処理の後に適用することにより、次の中間層に入力される特徴マップ数の削減ができ、学習の高速化を図ることが期待できる。そこで本研究ではmaxout を適用する。maxout は、

$$f'_{ij} = \max_{l \in [1, k]} f_{ij}^l \quad (6)$$

のように k 個の特徴マップの座標 (i, j) の値 $f_{ij}^1, f_{ij}^2, \dots, f_{ij}^k$ の最大値をとり1つの特徴マップに集約する手法であり、表1のネットワーク構成ではプーリング層の16の特徴マップを1/4に削減している。

表2 前景画像の変形パラメータ

変形	変形範囲
並進	$[-5, 5]$
スケーリング	$[0.8, 1.2]$
回転	$[-180^\circ, 180^\circ]$
輝度変化	$[-20, 20]$

3.2 学習サンプルの生成法

ConvNets は、学習サンプルの数が少ない場合、過学習に陥りやすいことが知られており、その対応のひとつとして大量の学習サンプルを用意すること重要である。一方で、ポジティブサンプルを人手で集めることはコストが高い作業といえる。そこで、手作業で用意した前景画像(ポジティブサンプル作成用画像)に対して、スケーリング、回転などの変形を適用し、実環境の背景画像と合成することで、少量のポジティブサンプルから大量のポジティブサンプルを生成する。

前景画像の変形パラメータは表2の範囲でランダムに与える。これらのパラメータは、頭部の輪郭の一部が隠れることや空間中の人の位置により頭部の大きさが変わってくる、人物の向き、照明条件の変化にそれぞれ対応しており、全方位カメラに写る人物に起こりうる見えの変化を表現している。また、元々歪んだ画像から前景画像を取得するため、歪みに関する情報は既に学習データに含まれていると本研究では仮定している。全方位カメラは通常のカメラと異なり真上からの視点となるように設置されることが多く、その場合、対象が最大で $\pm 180^\circ$ 回転して写ることから、本研究では回転の変形範囲を $[-180^\circ, 180^\circ]$ と大きくしている。ポジティブサンプルの作成は、

1. 透過情報を持つ前景画像を作成
2. 前景画像に回転、スケーリングを適用
3. 前景画像と背景画像を透過情報を元に合成
4. 前景画像全体に輝度変化を適用

の流れで行う。ネガティブサンプルおよび前景画像と合成するための背景画像は、実環境の背景画像からランダムな位置で切り出して作成する。

4. 全方位画像からの人検出実験

全方位画像からの人検出用に本研究で提案するConvNetsのネットワーク構成と学習サンプルの生成法について、その性能や適用すべきパラメータを実際に設置されている全方位カメラの画像を用いた人検出実験により明らかにする。本実験ではA大学附属図書

図2 本実験で用いる全方位画像の例

館に設置されている全方位カメラの画像を用いる。本実験で用いる全方位画像の例を図2に示す。図2は天井設置型の全方位カメラであるため俯瞰画像となっている。

図2のような全方位カメラの利用は、屋内での利用形態としては代表的なものであり、真上から見下ろす形で撮影するため、安定的に人を検出するには顔や体の検出よりはむしろ頭部を検出が最も効果的と考える。また、頭部検出であれば着席などの影響もほとんど受けないといった利点もある。そこで本研究では、提案手法の性能評価のため、全方位画像における人の頭部の識別による人検出器を実装する。具体的には、頭部を含む画像をポジティブクラス(正例)、ポジティブクラスに属さない画像をネガティブクラス(負例)とした2クラス識別器としてConvNetsのネットワークを構築する。このとき、学習に用いるデータ全てにラベルが付与されており、学習モデルとしては教師あり学習となる。

4.1 実験環境

本実験で用いる全方位カメラは 1280×980 ピクセルの解像度のカラー画像を出力するが、本実験ではこれをグレースケールに変換して扱う。また、全方位画像の上端・下端は吹き抜けであり人が居ない空間であること、左端・右端は解像度が極めて低く人がいるかどうか目視で判断できず、正解データとして人の位置を人手により求められないことから、全方位画像の中心から 768×672 ピクセルの領域を本実験で用いる画像として切り出している。実験用画像は、2013年7月に撮影した200枚の画像を学習サンプル作成用に、

2013年7月および8月に撮影した50枚の画像を評価用としている。評価用の50枚の画像には、合計で549人が写っている。

4.1.1 学習サンプルの作成

200枚の全方位画像から20,000件のポジティブサンプルと20,000件のネガティブサンプルを作成する。200枚の全方位画像から頭部の画像を手で切り出し、頭部以外の領域に透過情報を付加した50枚のポジティブサンプル作成用画像である前景画像と20,000枚の背景画像を合成することでポジティブサンプルを生成する。前景画像の変形パラメータは表2に従うものとする。背景画像およびネガティブサンプルは、200枚の全方位画像からランダムな位置で画像を切り出すことで生成する。前景画像・ポジティブサンプル・ネガティブサンプルの例をそれぞれ図3に示す。これらの画像は、提案するConvNetsのネットワークに入力できるよう40×40ピクセルの大きさに規格化されているものとする。図3より、生成したポジティブサンプルであるが、頭部画像が背景画像の半分以上を占めていることもあり、背景画像の合成は頭部の周囲にランダムな摂動を与えるものに相当すると考える。

(a) 前景画像の画像例

(b) ポジティブサンプルの画像例

(c) ネガティブサンプルの画像例

図3 前景画像・ポジティブサンプル・ネガティブサンプルの画像例

4.1.2 人検出処理および検出精度の評価指標

人検出を行う全方位画像に対して、学習サンプルと同サイズの検出窓(40×40ピクセル)を5ピクセルのステップサイズでラスタ走査し画像の切り出しを行う。768×672ピクセルの解像度の画像に対してこのような走査を行うことで、18,542枚の画像を切り出している。切り出した画像を学習済みのConvNetsに入力することで、入力がどのクラスに属するか識別し、識別結果のクラス確率が閾値以上のとき検出結果を人と判定する。閾値は誤検出率が必要な精度である $[10^{-5}, 10^{-2}]$ となるよう実験的に決定する。また、検出窓と頭部の正解領域(検出窓と同サイズ)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。検出精度の評価指標には、検出窓あたりの誤検出率(FFPW: False Positive Per Window)と1枚の全方位画像にいる人の数に対する未検出率(Miss Rate)、横軸に誤検出率・縦軸に未検出率をとったDET曲線、DET曲線の曲線下面積(AUC: Area Under Curve)を用いる。畳み込むカーネルサイズや畳み込み層の特徴マップ数などによって、ネットワークの性能が影響を受けることが想定できるため、カーネルサイズ・特徴マップ数・maxoutの有無をパラメータとし、ネットワークの学習時間・FFPW・Miss Rateの観点で分析する。

4.2 ConvNets ネットワークの学習時間および検出処理時間の計測

ConvNetsはDeep Learningの公開ライブラリのひとつであるPylearn2[16]を用いて実装する。Pylearn2は2014年10月時点のものを使用している。また、学習にはMini-batch 確率的勾配降下法を適用し、更新回数を10,000、バッチサイズを100とする。ネットワークの学習時間および人検出時間の計測には表3に示すスペックの計算機を用いる。

表1に示すConvNetsのネットワーク構成において学習処理に必要な時間および人検出処理に必要な時間を計測する。表3の計算機および4.1.1で作成した学習サンプルを用いた学習処理速度の計測より、表1に

表3 人検出器を実装した計算機の仕様

構成	内容
OS	Scientific Linux 6.4
CPU	Intel® Xeon® E5645 2.4GHz×6コア×2基
GPU	NVIDIA® GeForce® GTX580 1.5GHz×512コア
Memory	47.1GB

示すネットワーク構成では892分, maxout を省略したネットワーク構成では4,350分で学習が終了することを確認している. 計測結果から, maxout の適用は, maxout を適用しない場合に比べ学習処理が4.9倍高速であり, 学習サンプルの量を増やす場合など学習速度がより求められる場合に有効といえる. また, 画像1枚あたりの人検出処理時間は, 表1に示すネットワーク構成で76秒(1検出窓あたり0.0041秒), maxout を省略したネットワーク構成で80秒(1検出窓あたり0.0043秒)となっている. 検出時間の観点では, 5%程度のオーバーヘッドが許容できれば, maxout の省略は可能といえる.

4.3 ネットワークの構成によるConvNets の人検出精度の分析

表1に示すConvNets のネットワーク構成について, カーネルサイズ・特徴マップ数・maxout の有無をパラメータとして, 表4の範囲で変化させた結果のFFPW 10^{-3} におけるMiss Rate および DET 曲線のFFPWの範囲 $[10^{-4}, 10^{-3}]$ でのAUCを表5に示す. 表5より, カーネルのサイズは, maxout 有りの場合 5×5 が, maxout 無しの場合 7×7 の構成がそれぞれ

れ最も高い検出精度となっている. また, maxout 有りの場合, カーネルサイズ 7×7 のMiss Rate およびAUCの値はカーネルサイズ 5×5 の値に近く, ほぼ同等の検出精度であると考ええる. このことは特徴マップ数が16のDET曲線を図4からも確認できる. 特徴マップの数は16のものがmaxoutの有無を問わず最も良い.

図5に中間層第1層における畳み込み処理後の特徴マップを示す. 図5は, ポジティブサンプルおよびネガティブサンプルをそれぞれ20,000枚入力し, その畳み込み処理後の特徴マップの平均を可視化したものである. 各特徴マップの値はマップ毎に0~255に正規化しており, 打ち切り線形関数では負の値が全て0となるため, 黒のところは特徴マップの値が0または

表4 比較対象のConvNets のネットワーク構成パラメータ

パラメータ	値
maxout	有, 無
特徴マップ数	16, 32, 48
カーネルサイズ	3×3 , 5×5 , 7×7

表5 FFPW 10^{-3} におけるMiss Rate [%] および FFPW $[10^{-4}, 10^{-3}]$ におけるAUC (曲線下面積) [%]

maxout	カーネルサイズ	特徴マップ数	Miss Rate@FFPW 10^{-3}	AUC@FFPW $[10^{-4}, 10^{-3}]$
有	3×3	16	37.1%	80.5%
有	5×5	16	<u>33.1%</u>	<u>75.6%</u>
有	7×7	16	35.1%	78.3%
有	3×3	32	59%	90.3%
有	5×5	32	41.6%	82.8%
有	7×7	32	34.6%	76.5%
有	3×3	48	85.4%	97.6%
有	5×5	48	36.2%	80.2%
有	7×7	48	38.1%	79%
無	3×3	16	84.2%	97.6%
無	5×5	16	74.4%	94.7%
無	7×7	16	<u>28.2%</u>	<u>73.4%</u>
無	3×3	32	94.2%	99.1%
無	5×5	32	73.1%	94%
無	7×7	32	37.2%	83%
無	3×3	48	94.8%	99.3%
無	5×5	48	82%	96.9%
無	7×7	48	32.4%	78.6%

※ 下線部は maxout の有無それぞれでの最良値

図4 本研究で提案する ConvNets のネットワーク構成 (特徴マップ数 [16]) における DET 曲線

負であることを意味している。図5より、カーネルサイズが小さいほど、特徴マップの反応箇所が弱く、スパースな特徴マップになっている。カーネルサイズが大きくなるにつれ、反応箇所が多く滑らかな特徴マップになっていることが確認できる。また、ポジティブサンプルとネガティブサンプルを比較すると、ポジティブサンプルでは入力である頭部の円形に強く反応しているのに対し、ネガティブサンプルでは背景のテクスチャを学習していることがうかがえる。本研究では maxout により4枚の特徴マップを1枚の特徴マップに集約しているが、図5では、4特徴マップ毎に1の特徴マップを作成していることに相当する。

人検出精度と特徴マップとの関係においては、ポジティブサンプル・ネガティブサンプル問わず特徴マップの非スパースさに検出精度は依存していると考えており、入力される特徴マップのサイズにもよるができるだけカーネルサイズは大きくしておくことが望ましい。一方で、特徴マップ数を増やしたとしても、図6のように類似した特徴マップが生成されてしまう傾向にあり、計算時間の増加に見合うだけの精度向上にはつながらないといえる。人検出精度には、特徴マップの数よりも、カーネルサイズのハイパーパラメータが影響しているといえる。また、maxout の適用による 5×5 のカーネルが最も良かった理由であるが、図5(c)より、maxout によるラスタ走査方向での4特徴マップの集約の結果、頭部を学習するために必要な円形に強く反応する特徴マップの重みカーネルが学習されたためと考える。

本実験で最も良い検出精度となっている ConvNets のネットワーク構成は表1で提案する 7×7 の構成であるため、以降の実験では全て表1の構成を用いる。maxout 用いた構成は、Miss Rate が30% 台に抑えられており、maxout を用いない構成と比較して4.9倍高速に学習を行うことができることから、学習速度・検出精度の観点でバランスがとれていると判断できる。

4.4 人検出精度における既存手法との比較

表1のネットワーク構成の提案手法の人検出器および HOG 画像特徴量[2]・RAB (Real AdaBoost) [20] の組み合わせによる人検出器の有用性の比較のため、評価用画像からそれぞれ人検出を行う。HOG は全身画像を入力とした人検出のために文献[2]で提案されたものであるが、現在では人検出に限らず車両などの検出にも適用されており、物体検出での画像特徴量としてまず適用を検討される代表的な手法である。そのため、本研究でも頭部検出において HOG を適用する。また、HOG を物体検出の画像特徴量として適用する際、学習器には RAB が用いられることが多く、画像処理ライブラリ OpenCV での実装によりその組み合わせが広く活用されていることから、本研究では HOG と RAB を比較対象として用いる。カメラモデルを考慮した球面モデル上での HOG [8] も提案されているが、本研究では、カメラパラメータが未知という前提での検証を目的としているためオリジナル HOG [2] を用いる。HOG と RAB の学習には OpenCV で配布されている opencv_traincascade プログラムを用いる。用いた OpenCV のバージョンは 2.4.7 である。ConvNets と同じ学習サンプルを使用し、表6に示すパラメータ

図5 中間層第1層における畳み込み処理後の特徴マップ (特徴マップ数16, maxout 無し)

(a) ポジティブサンプル, カーネルサイズ 7×7 (b) ネガティブサンプル, カーネルサイズ 7×7

図6 中間層第1層における畳み込み処理後の特徴マップ (特徴マップ数48, maxout 無し)

表6 opencv traincascade の学習パラメータ

パラメータ	値
-numStages	16
-featureType	HOG
-bt	RAB
-w, -h	40

により学習を行う。表6では、表記が無いパラメータはデフォルト値を使用している。

提案手法と既存手法(HOG + RAB)のDET 曲線を図7に示す。本実験の環境では、FPPW が 10^{-3} を下回るとき、画像1枚あたりの誤検出数がおおよそ1桁におさまる。そのため、FPPW が 10^{-3} を下回るときの検出精度を実用的な値とする。図7より、FPPW 10^{-3} でのMiss Rate は、ConvNets を用いた提案手法が28.2%であるのに対し、HOG とRAB の組み合わせによる既存手法では77.5% となっている。このことから、提案手法は学習によって獲得した特徴量により、既存手法に比べ全方位画像での頭部検出において高い検出精度を有していることを確認している。

4.5 背景合成による人検出精度の検証

提案手法では、学習サンプルの自動生成において文献[18, 19] の変形に加え、実環境の背景画像を合成することでポジティブサンプルのバリエーションを増し

図7 提案手法およびHOGとRABの組み合わせによる手法のDET曲線

図8 背景合成の有無によるDET曲線

ている。そこで、背景合成を行わなかった場合のポジティブサンプルと背景合成を行った場合のポジティブサンプル、これら2種類の学習サンプルを用いて学習を行った結果をDET曲線を用いて比較し、背景合成による効果を確認する。ネガティブサンプルについては、背景からランダムに切り出したものを共通して使用する。

図8に背景合成の有無によるDET曲線を示す。図8より、背景合成を行った場合と比較して、行わなかった場合は未検出率が増加しており、背景合成によって検出精度が向上していることが確認できる。これは、識別器から見た場合にポジティブサンプルおよび検出すべき人画像は、同じとは言えないまでも、かなり類似したモデルから生成されていることを示唆している。ただし、生成したポジティブサンプルが自然画像と視覚的に遜色ないかどうかまでは明らかではない。機械学習において精度の向上にはできるだけ学習データを増やすことが重要であり、本研究で行っているランダムな摂動を加えた前景画像とランダムに切り出した背景画像を合成し学習データとするアプローチは、全方位画像における人検出において有用と考える。

提案手法による検出結果例を図9に示す。矩形領域は頭部として識別された検出窓である。ネットワークの学習に基本となる前景画像を50枚しか用いていないにもかかわらず、図9のように全方位画像からの人検出が行えていることが確認できる。一方で、カバンや椅子の背もたれなどの黒っぽいぼろ丸に見える物体は誤検出されやすいことを確認している。これは、図5(e)や図6(a)のように、ポジティブサンプルの学

図9 提案するConvNetsのネットワーク構成による人検出結果の例

習において重みカーネルが円形を抽出するように設定されることに起因している。本研究では頭部画像かどうかの2クラス分類を行っているが、椅子のように事前にその空間にあることがわかっているような物体については、他クラス分類へと本手法を拡張することで誤認識を低くすることができると考える。

5. おわりに

本研究では、カメラパラメータを推定することなく、全方位画像という歪みや位置による見えの変化のある画像に対する人検出器を提案している。Deep Convolutional Neural Networkをベースにその実現を

図っており、さらに、少量の学習サンプルに対してスケリング・回転・輝度変化の変形を適用することで、大量の学習サンプルを生成する手法も提案している。

全方位画像からの人検出用に本研究で提案するConvNetsのネットワーク構成と学習サンプルの生成法について、その性能評価を目的に、実際に設置されている全方位カメラの画像を用いた人検出実験を行っている。検出精度の評価実験では、maxout 無し・カーネルサイズ 7×7 ・特徴マップ数 16 のConvNets ネットワーク構成で、誤検出率 10^{-3} における最良の未検出率 28.2% を達成している。特に、カーネルサイズは局所応答をどの範囲で評価するかを意味しており、中間層第 1 層の可視化により、カーネルサイズが小さすぎると得られる特徴マップがスパースとなることを確認している。人検出精度と特徴マップとの関係においては特徴マップの非スパースさに検出精度は依存していると考えており、特徴マップがスパースにならないようにできるだけカーネルサイズは大きくしておくことが望ましい。一方で、maxout 用いたConvNets ネットワーク構成は未検出率が 30% 台に抑えられており、学習処理が maxout を用いない場合に比べ 4.9 倍高速であるため、学習速度・検出精度の観点でバランスがとれていると判断している。

学習サンプル生成時の背景合成の有無による検出精度の比較を DET 曲線を用いて行っており、本研究で提案する背景合成を行うことにより検出精度が向上することも示している。これは、識別器から見た場合にポジティブサンプルおよび検出すべき人画像是かなり類似したモデルから生成されていることを示唆しており、提案法は検出精度の向上に必要な学習データを増やすアプローチとして有用と考える。

HOG 画像特徴量と Real AdaBoost を用いた既存手法との比較では、HOG と Real AdaBoost の組み合わせによる人検出器が誤検出率 10^{-3} で未検出率 77.5% であるのに対し、提案手法では未検出率が 28.2% となっている。全方位画像からの人検出用に本研究で提案するConvNetsのネットワーク構成が、代表的な物体検出手法である HOG と Real AdaBoost を組み合わせた手法に比べ検出精度が向上しているといえる。頭部画像は円形であることを検出器が適切に捉えられていることがその要因と考えている。

今後の展望として、特徴マップの非スパース性が検出精度に影響を与えられられることから、local contrast normalization [17] などの特徴マップの正規化手法について適用を検討する。また、本研究では全身画像は学習の対象とはしていないため、頭部画像だけでなく全身画像も学習させた際の検出精度の検証も

必要と考えている。本研究は静止画からの人検出を扱っているが、精度の更なる向上や検出処理時間の削減には時系列性は重要なファクターであり、時系列性を考慮した人検出器の開発へと展開する。

謝辞

本研究は JSPS 科研費 24650546, 25330186 の助成を受けたものです。

参考文献

- [1] David G. Lowe. Object recognition from local scaleinvariant features. In *7th IEEE International Conference on Computer Vision*, Vol.2, pp.1150-1157, September 1999.
- [2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.1, pp.886-893, June 2005.
- [3] Peter Hansen, Peter Corke, Wageeh Boles, and Kostas Daniilidis. Scale-invariant features on the sphere. In *IEEE 11th International Conference on Computer Vision (ICCV)*, pp.1-8, October 2007.
- [4] Zafer Arican and Pascal Frossard. Omnisift: Scale invariant features in omnidirectional images. In *2010 17th IEEE International Conference on Image Processing (ICIP)*, pp.3505-3508, September 2010.
- [5] Luis Puig and José. J. Guerrero. Scale space for central catadioptric systems: Towards a generic camera feature extractor. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pp.1599-1606, November 2011.
- [6] Zafer Arican and Pascal Frossard. Scale-invariant features and polar descriptors in omnidirectional imaging. *IEEE Transactions on Image Processing*, Vol.21, No.5, pp.2412-2423, May 2012.
- [7] Javier Cruz-Mota, Iva Bogdanova, Benoît Paquier, Michel Bierlaire, and Jean-Philippe Thiran. Scale invariant feature transform on the sphere: Theory and applications. *International Journal of Computer Vision*, Vol.98, No.2, pp.217-241, June 2012.
- [8] Ibrahim Cinaroglu and Yalin Bastanlar. A direct approach for human detection with catadioptric omnidirectional cameras. In *2014 IEEE 22nd Signal Processing and Communications Applications Conference (SIU)*, pp.2275-2279, April 2014.
- [9] Luis Puig, José. J. Guerrero, and Kostas Daniilidis. Scale space for camera invariant features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.36, No.9, pp.1832-1846, September 2014.
- [10] Miguel Lourenço, Joao P. Barreto, and Francisco Vasconcelos. sRD-SIFT: Keypoint Detection and Matching in Images With Radial Distortion. *IEEE Transactions on Robotics*, Vol.28, No.3, pp.752-760, June 2012.
- [11] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis*

- and Machine Intelligence*, Vol.22, No.11, pp.1330–1334, November 2000.
- [12] Richard Hartley and Sing Bing Kang. Parameter-free radial distortion correction with center of distortion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.29, No.8, pp.1309–1321, August 2007.
- [13] Kenichi Kanatani. Calibration of ultrawide fisheye lens cameras by eigenvalue minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.35, No.4, pp.813–822, April 2013.
- [14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Vol.86, No.11, pp.2278–2324, 1998.
- [15] Kuniyuki Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, Vol.36, No.4, pp.193–202, April 1980.
- [16] Ian J. Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, and Yoshua Bengio. Pylearn2: a machine learning research library. *arXiv preprint arXiv: 1308.4214*, August 2013.
- [17] Kevin Jarrett, Koray Kavukcuoglu, Marc' Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pp.2146–2153, September 2009.
- [18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pp.1097–1105, 2012.
- [19] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. *Journal of Machine Learning Research*, Vol.28, No.3, pp.1058–1066, June 2013.
- [20] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, Vol.37, No.3, pp.297–336, December 1999.
- [21] Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, Vol.2, No.1, pp.1–127, 2009.
- [22] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems 19*, pp.153–160, 2007.
- [23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *arXiv:1409.0575*, 2014.
- [24] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *27th International Conference on Machine Learning (ICML)*, pp.111–118, June 2010.
- [25] Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp.9–48. Springer, 2012.
- [26] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *27th International Conference on Machine Learning (ICML)*, pp.807–814, June 2010.
- [27] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *Journal of Machine Learning Research*, Vol.28, No.3, pp.1319–1327, June 2013.

(2015年3月4日 受付)

(2015年7月26日 採録)

[問い合わせ先]

〒182-8585 東京都調布市調布ヶ丘1-5-1 西5号館707 室
 電気通信大学 大学院情報理工学研究科 総合情報学専攻
 岡本 一志
 TEL : 042-443-5280
 E-mail : kazushi@uec.ac.jp

著者紹介

あさぬま ひとし
浅沼 仁 [非会員]

2013年千葉大・工・情報画像卒、2015年同大学院融合科学研究科博士前期課程修了。同年、大日本印刷株式会社入社。現在に至る。修士(工学)。

おかもと かずし
岡本 一志 [正会員]

2006年高知工科大学工学部情報システム工学科卒業。2008年同大学院工学研究科基盤工学専攻修士課程修了。2011年年東京工業大学大学院総合理工学研究科知能システム科学専攻博士後期課程修了。同年、千葉大学アカデミック・リンク・センター特任助教。2015年電気通信大学大学院情報理工学研究科総合情報学専攻助教。現在に至る。博士(工学)。データマイニングの研究に従事。IEEE、情報処理学会、日本知能情報ファジィ学会各会員。

かわもと かずひこ
川本 一彦 [正会員]

1997年千葉大・工・情報工卒。2002年同大学院自然科学研究科博士後期課程了。同年、東工大学院総合理工学研究科知能システム科学専攻助手。2005年 九工大工学部数情報基盤講座助教授。2008年同大学院工学研究科基礎科学研究系准教授。2009年千葉大学総合メディア基盤センター(2013年より統合情報センター)および同大学院融合科学研究科情報科学専攻准教授。現在に至る。博士(工学)。コンピュータビジョン研究に従事。IEEE、情報処理学会、日本知能情報ファジィ学会各会員。

Feature Learning Based Human Detection for Omnidirectional Images by

Hitoshi ASANUMA, Kazushi OKAMOTO and Kazuhiko KAWAMOTO

Abstract :

A human detection method for omnidirectional images with distortion and appearance change is proposed in order to improve detection accuracy without camera calibration. The method is based on deep convolutional neural network, and a training data increase method is also proposed with applying translation, scaling, rotation, and brightness change operators to manually prepared training data. In the evaluation experiment with actual omnidirectional images, under the condition at false positive per detection window 0.001, the proposal achieves miss rate 28.2% and miss rate of a combination of HOG and Real AdaBoost, which is standard method in object detection, is 77.5%. In addition, it is validated that the proposed training data increase method improves detection accuracy, and relationships between visualized feature maps and detection accuracy are discussed. The proposal provides improvement of human detection accuracy without estimating camera parameters, where camera calibration is difficult in situations such as large number of cameras, cameras on distant places, and unknown camera models.

Keywords : omnidirectional image, camera calibration, feature learning, neural network, human detection

Contact Address : **Kazushi OKAMOTO**

Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications

1-5-1, Chofugaoka, Chofu, Tokyo, 182-8585, Japan

TEL : +81-42-443-5280

E-mail : kazushi@uec.ac.jp