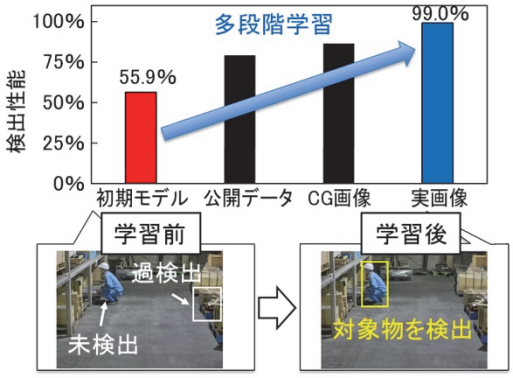
深層学習を用いた画像認識における多段階学習による少数の実画像で構築可能な物体検出技術の開発

Object Detection Developed with a Small Number of Acquired Images by Multi-step Training Method in Image Recognition Using Deep Neural Network



小林 周 *1 Amane Kobayashi
松本 知浩 *2 Tomohiro Matsumoto
杉本 喜一 *3 Kiichi Sugimoto
岩田 健司 *4 Kenji Iwata

産業車両の安全支援のための人検出システムなどでは，画像上の物体を検出する画像認識機能が不可欠である。深層学習を用いた画像認識では，実運用環境に映る検出対象物の画像を大量に学習させる必要があるが，画像取得の作業に多大な労力を要していた。そこで，三菱重工業株式会社(以下，当社)は，国立研究開発法人産業技術総合研究所(以下，産総研)と共同で，実運用環境の条件に近い公開データやCG(コンピュータグラフィックス)を用いて学習データを拡充し，これらを段階的に効率良く学習させる多段階学習手法を開発した。本手法により，従来法で実画像およそ5000枚の学習で得られる検出性能と同等の性能を，実画像280枚のみの学習で達成できることを確認した。当開発技術は，深層学習を用いた画像認識機能を有する当社製品に広く適用する予定である。

1. はじめに

画像上の人物や車両など特定の物体の位置や種別を推定する物体検出は，深層学習を用いた画像認識技術により実現されることが多い[(1)]。近年では，公開データセットを用いて学習した物体検出器が容易に入手できるようになったが，これをそのまま当社製品に組み込んでも，お客様の実運用環境で要求する検出性能を満足することはできない。これは，公開データセットと実運用環境では，背景に映る物体やカメラ機材など撮影条件が異なることに起因しており，この違いを解消するためには，運用環境ごとに取得した実画像を相当枚数用いて，再学習させる必要がある。しかしながら，画像取得の作業には多大な労力を要する上に，お客様の環境に簡単に立ち入れない場合や一般には目にしない特殊な物体を検出対象物とする場合など，画像取得機会が制限されている場合では，高い画像認識機能を実現することが困難となる。そこで，実運用環境の条件に近い公開データやCGを用いて学習データを拡充し，これらを段階的に学習することで高い検出性能を実現する多段階学習手法を開発した。

2. 多段階学習

本研究で開発した多段階学習手法は，ニューラルネットワークの学習で利用されてきた“転移学習”に基づいた手法である[(2)]。本研究では，転移学習を段階的に適用するよう拡張し，(ステップ0)公開データセット，(ステップ1)最適化した公開データセット，(ステップ2)背景画像に対するCG 重畳画像，(ステップ3)少数の対象物の実画像，の順に学習する方法を開発した。このプロ

*1 デジタルイノベーション本部 CIS部 博士(理学)
*2 デジタルイノベーション本部 CIS部
*3 デジタルイノベーション本部 CIS部 主幹技師 技術士(情報工学部門)
*4 産業技術総合研究所 情報・人間工学領域 主任研究員 博士(工学)

セスの狙いは，ステップ3に至るまでの学習を通じて，物体検出器を汎化的なものから実運用環境に特化したものに順次チューニングすることで，必要な対象物の実画像を少数に抑えることである。以降では，公開データセット学習済のモデルをステップ0として，ステップ1以降の詳細を説明する。

（ステップ1） 最適化した公開データセットの学習

ステップ0の学習データである公開データセットは，一般的に幅広いバリエーションの画像を網羅しているので，実運用環境に近い画像データセットとなるよう最適化し，再学習することで実運用環境に特化した物体検出器を構築する。最適化する条件としては，対象物の色彩や背景に写る物体，撮影時刻など物体検出を適用するアプリケーションごとに様々なものが考えられるが，膨大量の公開データに対して，これらを判定することは簡単ではない。ここでは，アノテーション時のBounding BOX（BBOX）の情報から容易に判定できる画像上の対象物のサイズについて述べる。

実運用環境で撮影される画像上の対象物のサイズは，使用するカメラの取付け位置や画角といった特性や，カメラと対象物との相対位置関係から，画像上に映る対象物のサイズが限定される。 そこで，画像上に映る対象物の高さ・幅の画素数の全体の画像サイズとの割合が，実運用環境で撮影される対象物と同じ条件の画像を公開データセットから選別し，これを学習データとする。全体の画像サイズとの割合を条件としたのは，物体検出器に画像が入力される際には，どのサイズの画像でも決まった画素サイズにリサイズされるためである。

（ステップ2） 背景画像に対するCGの重畳画像の学習

本ステップでは，実運用環境の背景画像に対象物の CG 画像を重畳することにより，不足している学習用画像を作成し，学習に用いる。CGモデルは，色彩や形状，姿勢を運用条件に合わせたものを用意する。CG を画像へレンダリングする際には，使用するカメラの取付け位置・角度，視野に合致する投影パラメータを設定する。投影グリッドに対する CG モデルの配置，カメラ視点に対する向きをバリエーションとして，多数の学習用画像を作成する。

（ステップ3） 少数の実画像の学習

実運用環境で特定のエリアに限定して，対象物が映る実画像を少数取得し，学習する。このとき，学習する画像が少数であるために，過学習が生じる恐れがある。そのため，定点カメラを設置するなどしてお客様の負担とならない方法で，評価用画像を用意し，学習した物体検出器の評価を実施することが必要となる。実画像を取得するエリアは，次章で説明する背景分析手法で特定する，未検出・過検出となりやすい条件のエリアとする。

3. 多段階学習を用いた物体検出器構築方法

多段階学習は過去に学習した物体検出器のネットワークパラメータを検出対象物に適合させるための方法論であり，これだけでは高い性能を達成するために必要な対象物の実画像を削減し，画像収集コストを低減することができない。そこで，多段階学習を有効に活用するための物体検出器構築方法を開発した。この方法の作業フローを図1に示す。多段階学習の前に実運用環境の背景撮影と背景分析，少数の対象物が映った実画像の撮影を実施する。多段階学習によって得られる物体検出器を評価し，目標性能の達成を確認した後，製品に組み込む。

背景撮影では，運用エリア全域で背景画像を取得する。運用エリア全域が撮影対象となるが，対象物を配置して姿勢や向き，カメラとの相対位置といった様々な条件で撮影する作業と比較して画像収集コストは格段に低い。ここで得られる背景画像は，ステップ2において CG を重畳する画像として用いる。さらに，背景画像に対して，次に述べる2つの背景分析手法を適用し，未検出・過検出が生じやすいエリアを特定する。未検出・過検出エリアで対象物の実画像を取得することで，少数であっても，検出が難しい条件，即ち，本質的に学習に必要なデータを，多段階学習の最終ステップに適用することができる。なお，ここでは，未検出は，画像上に検出対象物が映っているのに位置・種別情報が検出されない事象を，過検出は，画像上の検出対象物ではない

物体を，検出対象物として誤って検出してしまう事象を，それぞれ指すものとする。

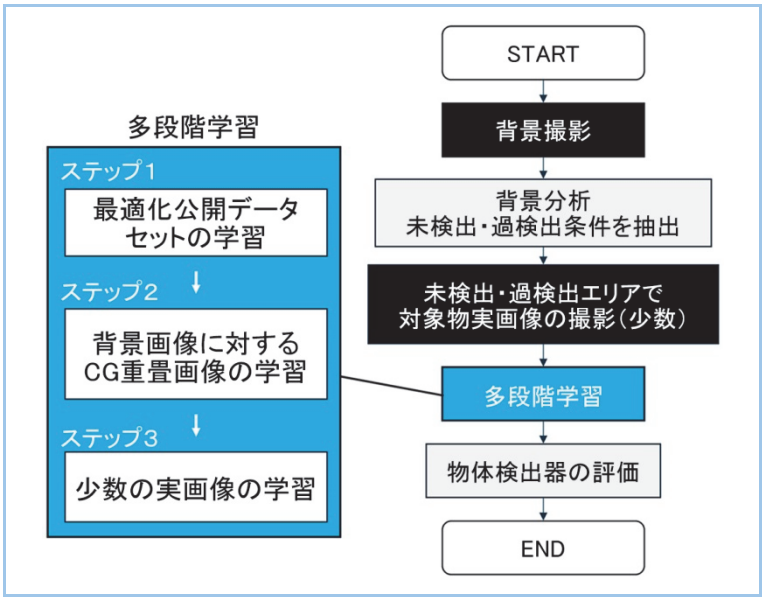


図1　物体検出器構築フロー

未検出は，物体検出器が算出する，画像に映る物体が対象物である確度を示す“信頼度スコア”が低いために生じる事象である。信頼度スコアが低下する一因は，対象物付近の背景領域のテクスチャによって対象物の特徴抽出が弱められてしまうことにある。従って，どのようなテクスチャを持つ背景領域で未検出が生じやすいか判定できれば，未検出エリアを特定することが可能となる。本研究では，未検出エリアを特定するため，図2(a)に示すように背景画像に対して検出難易度をヒートマップとして可視化した。検出難易度マップの可視化には，畳込みネットワークをベースとした特徴抽出器を使用した。この特徴抽出器には，様々なテクスチャのバリエーションを網羅した背景画像データセット Places205[(3)]の画像に，対象物画像を全ての配置に対して順次重畳し，都度，検出処理をした際の信頼度スコアを重畳位置の検出難易度として学習している。このようにして獲得した特徴抽出器を実運用環境の背景画像に適用し，検出難易度マップを分析することで，高い難易度を示す物体が映るエリアを未検出エリアとして特定することが可能となる。

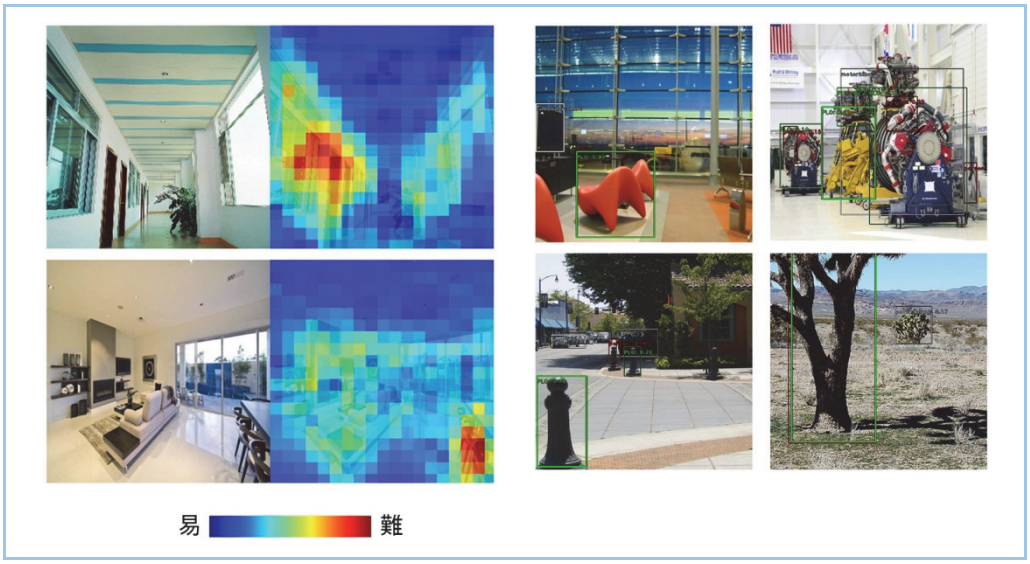


図2　背景分析による未検出・過検出条件の抽出例
(a)　背景画像（左）と検出難易度マップ（右）　(b)　背景画像における過検出物体（緑枠）

一方で，過検出は，特定のテクスチャを持つ背景領域の信頼度スコアが高いことで生じる事象である。そこで，本研究では，対象物が映っていない背景画像データセット Places205 に対して物

体検出処理を実施し，信頼度スコアが高いテクスチャ領域を過検出物体として学習した物体検出器を用意した。図2(b)に示すように，この物体検出器を実運用環境の背景画像に適用し，過検出を起こす確度の高いテクスチャ領域を検出することによって，過検出エリアとして特定することが可能となる。

4. 実運用を想定した評価

図1の作業フローに従って，産業車両の安全支援のための人検出システムを例題に，倉庫を試験エリアとして開発手法を検証した。

4.1 背景撮影と未検出・過検出エリアでの撮影

まず，運用条件と同じカメラ高さ・視野角で，試験エリア全域を移動しながら背景を撮影した。背景分析の結果，今回対象とした試験エリアのうち，6箇所のカメラ配置で未検出・過検出エリアが特定された。取得した学習データの概要を表1に示す。背景画像は，自然光の変化を考慮し，朝・昼・夜の3シーンを取得した。未検出・過検出エリアでは，短時間で撮影が完了するように，対象物の姿勢バリエーションは歩行姿勢のみとした。

学習データの全取得作業は，背景分析を含めて2日で完了した。背景画像では一回の撮影で数千枚もの画像を取得するが，収集時間は移動時間のみであるため，画像収集コストは低い。また，対象物を試験エリア全域で撮影することがないため，従来であれば背景画像と同じ千枚オーダの画像を収集していたところが，280 枚と少数の画像収集に留まった。

表1 実画像の学習データと評価データ

取得データ	学習データ		評価データ	
	背景画像	未検出・過検出エリアの対象物画像	未検出・過検出エリアの対象物画像	未検出・過検出エリア“以外”の対象物画像
撮影方法	試験場所全域を移動しながら連続撮影	歩行する人を撮影	直立・中腰・しゃがみ姿勢，前後左右の向きを撮影	歩行する人を撮影
データ量	朝:4437 枚， 昼:5448 枚， 夕:5131 枚	画像:280 枚， 対象物数:280 人	画像:1245 枚， 対象物数:1245 人	画像:193 枚， 対象物数:193 人

4.2 多段階学習の適用

表2に多段階学習に用いた学習データを示す。初期モデルは，MS-COCO[4]を事前学習した物体検出器を採用した。まず，運用条件と同じサイズの対象物のみとなるように選定した公開データセットを学習した。先に取得した実運用環境の画像に映る対象物のサイズは，全体が幅1150×高さ 870 画素の画像に対して，幅 30～290×高さ 100～690 画素であった。公開データセットFCDB（Fashion Culture Data Base）[5]から，このサイズの対象物のみ映っている画像を MS-COCO と同程度の 20 万枚抽出して学習した。

表2 多段階学習の学習データ

ステップ	学習データ	枚数
0	MS-COCO	12 万枚
1	最適化 FCDB	20 万枚
2	CG 重畳画像＋MS-COCO	5 千＋1 万枚
3	未検出・過検出エリアの人画像	280 枚

次に，取得した背景画像に対して人の CG を重畳し，学習した。人物画像として，体型や服装の異なる人物 CG モデルを作成し，これを歩行などの動作から様々な体勢として，カメラの画角や視点，レンズ曲率を運用場所のカメラ条件に合わせて調整し，レンダリングした。このようにして作成した CG を背景画像に重畳すると図3のようになった。CG 重畳画像の枚数は，従来手法で必要であった実画像の枚数分を設定した。学習データにはこれらの重畳画像に加えて，過学習防止のため，MS-COCO のデータを混合した。最後に未検出・過検出エリアの人の歩行画像を学習

した。なお，ステップ1では大量の画像を学習することで，計算コストが高くなるため，産総研の AI 橋渡しクラウド ABCI を利用した。

図3 背景画像と CG の重畳画像例

4.3 物体検出器の評価

多段階学習で構築した物体検出器の評価データは，未検出・過検出エリアで，種々の姿勢，向きの人画像とした（表1）。このデータは，検出難易度が高い条件であり，これがクリアできれば，実運用環境のその他の条件でも検出可能と評価することとした。ただし，多段階学習で過学習が生じている場合も想定して，未検出・過検出エリア“以外”の人画像を加えて評価した。

評価指標は下記の再現率と過検出率とした。

$$\text{再現率} = \frac{\text{正検出数}}{\text{対象物の数}}, \quad \text{過検出率} = \frac{\text{過検出数}}{\text{検出数}}$$

正検出は，正解となる BBOX A と識別結果の予測箇所を示す BBOX B の重なり具合を評価する，IoU＝A∩B/A∪B が 0.25 以上となり，なおかつ，信頼度スコアが任意の閾値以上となる検出結果とし，それ以外は過検出とした。本研究では，過検出率が 0.5%以下となるスコアの閾値を採用し，そのときの再現率を物体検出器の検出性能として採用した。

図4に示すように，多段階学習によって，再現率は 55.9%から 99.0%へ改善された。初期モデルでは，未検出・過検出となっていた事象について，本学習によって正確に検出されることを確認した（図5）。図6に示すように，画像上の対象物サイズが幅 200～290×高さ 500～690 画素（画像全体の高さの 57～79%）以上の大きく映る画像と，これより小さく映る画像，対象物が物陰に隠れたオクルージョン(*1)条件の画像，対象物を分類し，各ステップのモデルの未検出数をカウントした結果，表3となった。この傾向を分析すると，次のことがわかった。

表3 各ステップの未検出数

ステップ	モデル	未検出数			合計
		小さく映る対象物	大きく映る対象物	オクルージョン	
0	初期モデル ステップ1	209	368	58	635
1	最適化公開データセットの学習	114 55%	148	52	314
2	CG 重畳画像の学習	109 ステップ2	51 66%	33	193
3	少数の実画像の学習	1	2	11	14

（ステップ1）公開データセットの最適化によって，全てのサイズの対象物に対して検出性能を大幅に改善した。

（ステップ2）CG 重畳画像の学習では，大きく映るサイズの対象物の検出性を改善した。一方で，これより小さく映る対象物に対する効果はほとんどなかった。

（ステップ3）少数の実画像の学習では，すべての条件の対象物の未検出を抑えることができた。しかし，オクルージョン条件の対象物を中心に，一定数未検出を残す結果となった。

(*1) オクルージョン(Occlusion)：カメラに対して手前にある物体が後ろにある物体を隠す状態

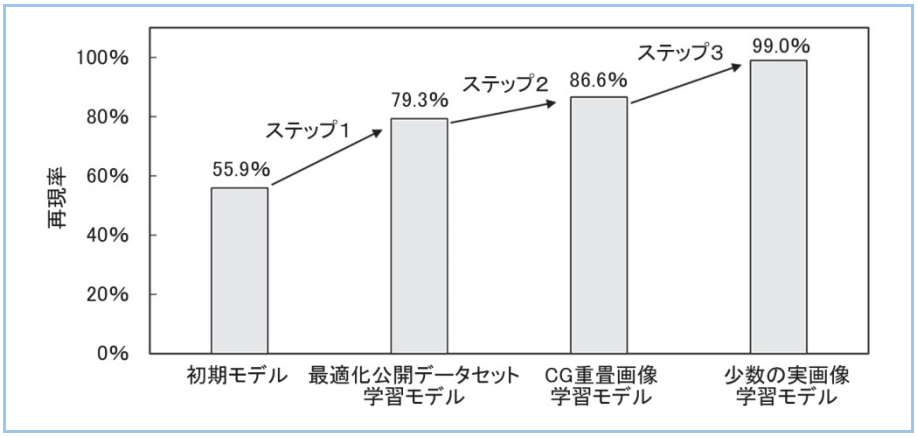


図4 多段額学習のステップごとの検出性能

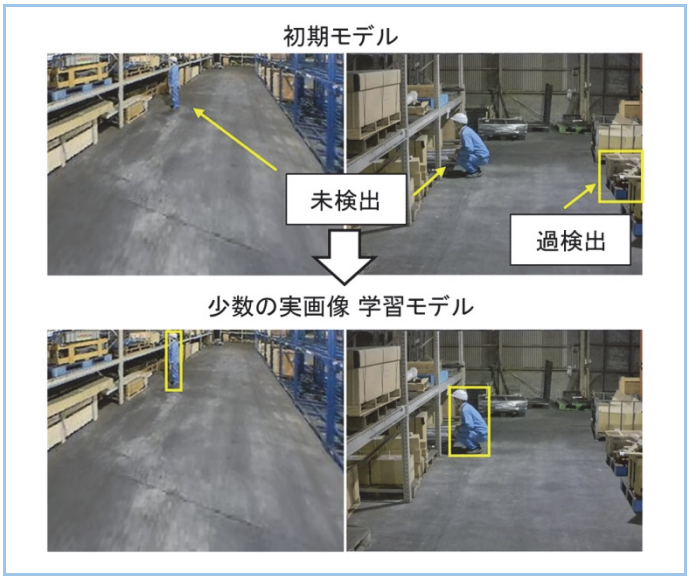


図5 学習前後の検出処理例図

図6 評価データの代表例

5. 考察

4.3 節の結果から，公開データセットの最適化をすることで，該当するサイズの対象物に対する検出性能を改善し，CG 重畳画像の学習によって CG で表現可能な姿勢の対象物の未検出を抑える効果があると推察することができる。一方で，小さく映る対象物は同条件の CG が多数重畳されていたにも関わらず，CG 重畳画像の学習では大きな改善は見られなかったことから，小さい対象物の CG 表現が実画像の特徴と大きくかけ離れたものであったと考えられる。また，オクルージョン条件は，学習データに反映させなければ検出させることは困難であることを確認した。この条件は，近年精度が向上している深層学習を用いた単眼奥行推定[(6)]などを用いて，背景に映る物体と重畳するCGの奥行方向の前後の位置関係を決定し，オクルージョンを再現した重畳画像生成ロジックを導入することで改善が期待できる。

多段階学習を導入した作業フロー(図1)を用いることで，対象物の実画像 280 枚のみで，再現率 99%と高い検出性能を有する物体検出器を構築可能であることを確認した。表4に示すように，データ取得作業時間を従来の 1/4 に短縮する効果があることがわかった。さらに，アノテーション[(*2)]に要する時間も大幅に短縮できることから，全体の作業を効率化するフローであると評価した。

(*2) アノテーション(Annotation)：画像上の検出対象の物体が映っている場所を矩形で囲うことで，学習時に必要な情報を画像データに付与する作業

表4 従来との比較

項目	従来	本研究
対象物の実画像の枚数	4647 枚	280 枚
データ取得作業時間	8 日間	2 日間
アノテーション時間	1 ヶ月	10 時間

6. まとめ

本研究では，公開データセットや CG を用いて学習データを拡充し，少数の実画像だけでも極めて高い検出性能を有する物体検出器を構築できる技術を開発した。本成果は，多段階学習を用いて戦略的に学習するフローを確立したことが大きく寄与しており，何度もデータ取得を繰り返すことなく，計画的に画像取得作業を物体検出器の構築工程に組み込むことができるようになった。これにより，作業工数を削減すると共に，お客様の現場生産作業への影響を最小限に抑えることができる。今後は，様々なアプリケーションに展開し，課題点を洗い出すことで，物体検出器構築手法のさらなる高度化を目指す。

参考文献

(1) L. Jiao et al., A Survey of Deep Learning-Based Object Detection, IEEE Access, vol. 7, pp. 128837-128868, 2019.
(2) 池田ら, AI を利用した膀胱癌の内視鏡診断, Precision Medicine, 2, 230-233, (2019).
(3) B. Zhou et al., Learning Deep Features for Scene Recognition using Places Database, NIPS 2014, Vol. 1, pp. 487-495, 2014
(4) Lin, et al., Microsoft COCO: Common Objects in Context, arXiv:1405.0312 (2015).
(5) Abe et al., Fashion Culture Database: Construction of Database for World-wide Fashion Analysis., ICARCV , page 1721-1726. IEEE, (2018).
(6) René Ranftl et al., Vision Transformers for Dense Prediction, Proc. ICCV 2021, pp. 12179-12188.