平成 29 年度卒業論文

深層学習と画像変換を用いた
複雑背景における物体検出

宮崎大学　工学部　情報システム工学科

大岐　建次郎

指導教員　椋木雅之

目次

1. はじめに

近年、画像認識の分野において深層学習が大きな成果を挙げ、注目を集めている。その活用例は多岐にわたり、顔識別の他、インターネット販売における類似商品の画像検索、SNS に投稿された画像において不適切なものを自動判定するサービスなど、様々な活用が進められている。中でも物体検出は 2012 年の ILSVRC 画像分類コンテストで優勝して以来、盛んに研究されている。

しかし、一般背景に対する物体検出の研究例は多いが、複雑背景に対して物体検出を行う研究は少ない。ここで複雑背景とは、検出したい物体に対してまぎらわしい背景であり、一方、一般背景とは、検出したい物体と区別がつきやすい背景のことを指す。

現在、実生活において複雑背景における物体検出の技術が求められている。例えば海外からの輸入品に虫やごみがついていないか検品する場面や、森林や草原の中で周囲の自然に溶け込みうまく身を隠す動物、昆虫、迷彩服を着用した人物などを検出することは複雑背景における物体検出となる。複雑背景での物体検出は一般背景に比べて困難と考えられる。

本研究では、深層学習を用いた最先端の物体検出手法と画像変換を組み合わせることで、複雑背景にも適用できる物体検出を目指す。

2. 複雑背景における物体検出

2.1.　複雑背景とは

本研究で扱う複雑背景とは、検出したい物体と類似した物体が多く含まれるような背景である。例えば、迷彩服着用人物を検出する際、服の模様とよく似た木々や枯葉などの背景である。

物体が多数存在し、そこから目的の物体を検出するという意味での複雑背景からの物体検出ではなく、検出したい物体とまぎらわしい背景が、本研究における複雑背景の定義である（図 1)。

図 1　複雑背景の例

2.2. 深層学習を用いた物体検出

複雑背景における物体検出では背景と対象物体の判別がつきにくいため、人手によって対象物らしさを与えることは難しい。計算機自体が対象物らしさを学習する深層学習であれば複雑背景においても有効である可能性がある。

従来から深層学習を用いた物体検出の研究が多く行われている。[1]では物体候補を抽出し、畳み込みニューラルネットワークにより特徴量を抽出して物体位置を推定している。[2]ではあらかじめ画像全体をグリッド分割し、領域ごとに物体のクラスと物体位置を推定することで提案生成、特徴量の再サンプリングが不要になり、実行時間の短縮に成功している。[3]では[2]の検出精度、実行時間の短縮を図っている。

このように近年、深層学習を用いた物体検出の研究が行われているが、それらは一般背景を想定しており、複雑背景において手法を評価、検討しているものはない。

2.3. 画像変換による物体検出の改善

複雑背景画像では均等化、メディアン処理、スムージングなどの画像変換を施すことで、背景と物体との僅かな差異が強調され、人が目視したときに対象物体の位置が分かりやすくなることがある。計算機による物体検出器においても、入力画像に画像変換を施すことで検出性能が向上する可能性がある。

画像変換を施すことで物体検出の精度を向上させる研究として[4]は、検出する画像に対しコントラスト強調、スムージング処理、階調変化など、様々な画像変換を行い、LBP 特徴量と Ada Boost を用いて多数の候補領域を抽出することで検出漏れを少なくする手法を提案し、最終的な検出率の向上に成功している。ただし、検出器自体の性能が低く誤検出が非常に多くなるため、画像内には必ず対象物体が 1 つのみ存在するという前提条件をおいており、多数の候補領域の中から最も対象物らしいもののみを検出結果としている。

3.　深層学習と画像変換を用いた物体検出手法

本研究では深層学習と画像変換を組み合わせた手法を提案する。

3.1.　SSD

本研究では物体検出器として[3]の Single Shot Multi Box Detector(SSD)を用いる。SSD は画像の特徴量を示すフィーチャマップを複数の階層で求め、検出用畳み込みフィルタでフィーチャマップ毎に、画像内の物体の種類(物体クラス)と物体位置(検出ボックス)とその物体である確率(スコア)を予測、算出することで物体検出を行う。図 2 に SSD のモデルイメージを、図 3 に各フィーチャマップの畳み込みフィルタのイメージを示す。異なるスケールのフィーチャマップを用いることで小さな物体の検出も可能であり、速度、精度ともに[1][2]など、他の手法を上回る性能を示している。

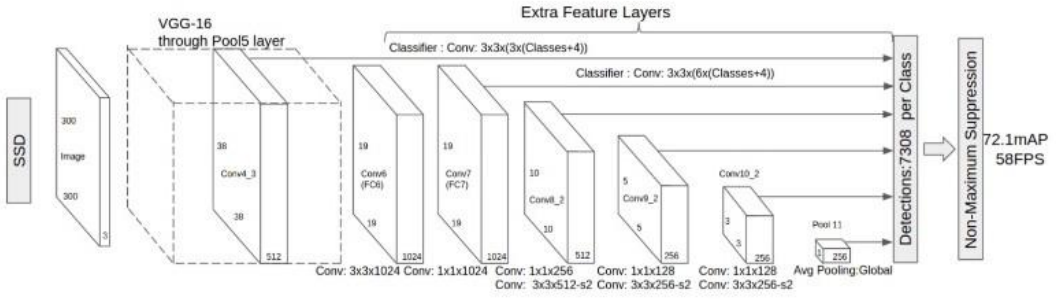

図 2　SSD モデル：ベースネットワークの最後に複数のフィーチャマップ層を追加する。[3]より引用。

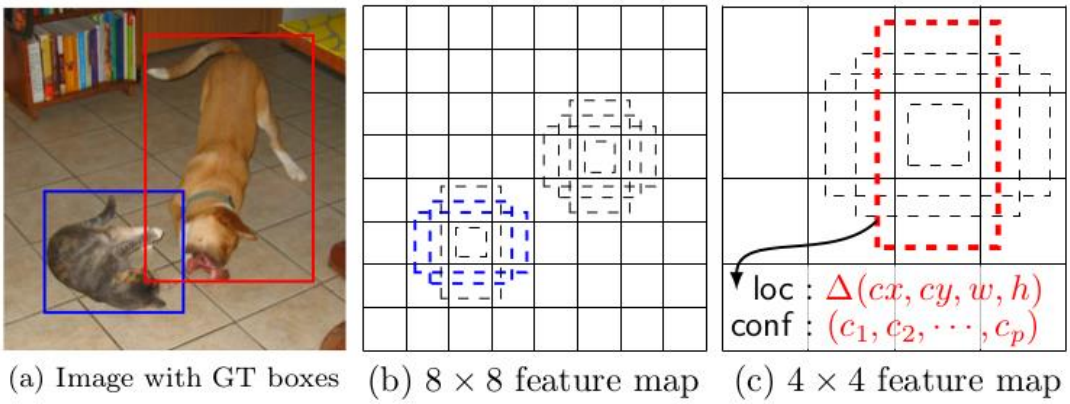

図 3　検出用畳み込みフィルタ：3x3 の小さなフィルタを掛け、物体クラス、検出ボックス、スコアを出力。[3]より引用。

SSD を利用するためには 3 種類のデータセットが必要となる。学習するための訓練用画像と、学習時正常に学習が行われているか確認(テスト)するためのテスト用画像、そして学習によって生成した学習モデルを用いて、実際に検出の成否を評価する評価用画像である。データセットの画像には、画像内にある物体について物体の種類(クラス)と位置(正解ボックス)の情報が与えられている。

本研究では PASCAL VOC データセット[5]を SSD に事前学習させたモデルを用いる。このデータセットは画像分類器、物体検出器などを構築および評価するために用いられるデータセットである。11,530 枚の画像に、人、鳥、車など 20 種類のクラスの 27,450 個の物体が含まれている。

3.2. 画像変換の種類

本研究では 7 種類の画像変換を利用する。各処理について以下に示す。

- スムージング（平均化フィルタ）

  中心画素と 3x3 範囲の周辺画素の輝度値の平均をとり、中心画素値をその平均値に変換する処理。

- 均等化（ヒストグラム平坦化）

  各画素をヒストグラム化したものに対して、画素が集中している部分を平坦化させることにより見やすい画像にする処理。

- メディアン処理

  中心画素と 3x3 範囲の周辺画素の輝度値を昇順に並べ、中心画素値をその中央値に変換する処理。

- モノクロ

  画像をモノクロ化させる処理。

- 階調変化

  256 階調で表現されている画像の階調数を 64 に減らす処理。

- エンボス処理

  画像の濃淡差を用いて輪郭部分を立体化させる処理。

- 補色変換

  画像内の色をその色と補色関係にある色に変換する処理。

画像変換手法としてスムージング、均等化、メディアン処理を選んだ理由は、[4]の研究において有効な画像変換とされていたためである。その他の手法は原画像と比較して見た目が大きく変わる画像変換であり、原画像で検出漏れした物体の検出に成功するなど、異なる検出結果となることを期待して選出した。実際に画像変換した画像を図 4 に示す。

a)原画像

b)スムージング

c)均等化

d)メディアン処理

e)モノクロ

f)階調変化

g)エンボス処理

h)補色

図４：画像変換の例

(原画像の出典：

https://upload.wikimedia.org/wikipedia/commons/1/1f/U.S._Army_Spc_130226-A-MX357-082c.jpg)

3.3.　組み合わせ手法

深層学習と画像変換を組み合わせる手法２つを提案する。いずれの手法でもPASCAL VOC データセットと複雑背景画像データセットを学習させた SSD を用いる。

手法１では、評価用画像に画像変換を施して検出を行う。図 5 に手法１のイメージ図を示す。この手法は[4]の実験と同様、画像変換によって多数の検出結果を求めることで検出漏れを減らすことを目的としている。

また、より性能を向上させるため、画像変換ごとの検出結果の和をとって適合率、再現率を求める実験も行う。

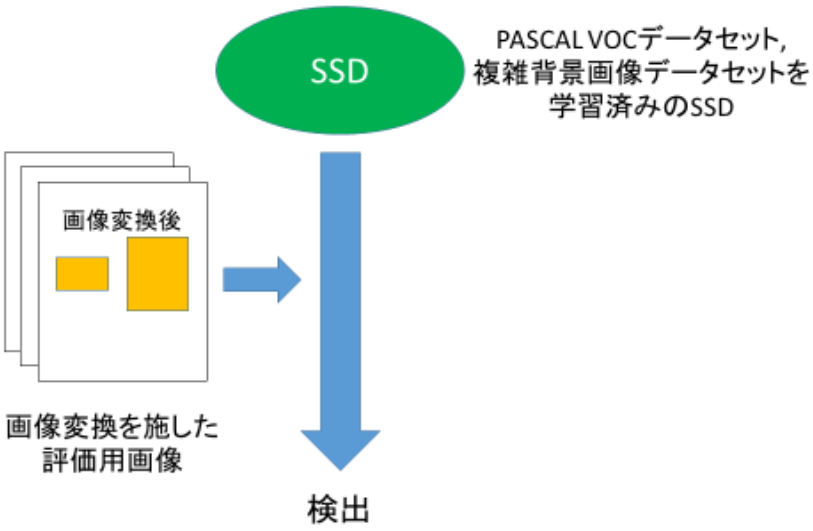

図 5：手法１：評価用画像のみを画像変換

手法 2 では、複雑背景画像データセットの訓練用画像に画像変換を施して追加学習する。図 6 に手法 2 のイメージ図を示す。こうして生成した学習モデルを用いて複雑背景画像中の対象物体を検出する。この時、手法 1 と同様に評価用画像にも画像変換を施して検出する実験も行う。この手法は変換画像のデータを水増し(データ拡張)して学習することで、SSD の性能を向上させ、誤検出を減らすことを目的としている。一般にデータ拡張には画像のコントラスト変換や拡大縮小、回転などを行うが、本研究では複雑背景を想定して先に挙げた 7 つの画像変換を行う。

また、より性能を向上させるため、手法 1 と同様、画像変換ごとの検出結果の和をとって適合率、再現率を求める実験も行う。

次節の実験で 2 つの手法をそれぞれ評価する。

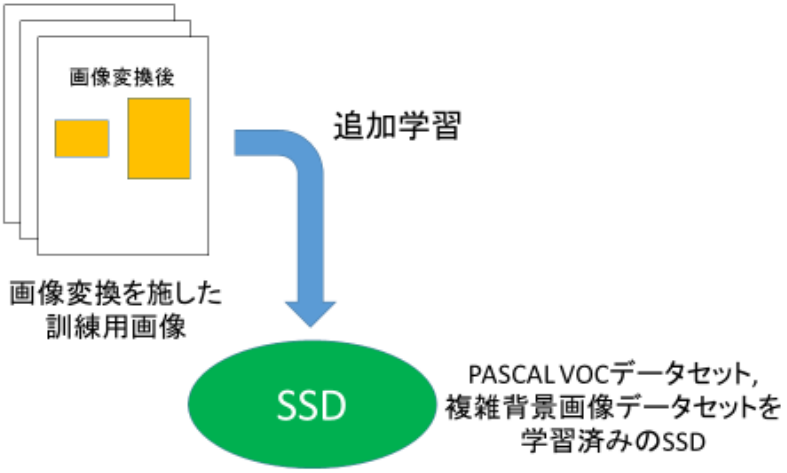

図 6：手法 2：訓練用画像を画像変換

4. 実験

4.1. 複雑背景画像

本研究の実験では、複雑背景画像中で迷彩服を着用した人物を検出対象物体とする。画像はインターネット上から集めた 120 枚を用いる。検出対象物体のクラス(人)と位置座標(正解ボックス)は人手により作成した。

学習では画像 1~10 はテスト用画像、11~20 は評価用画像、21~120 は訓練用画像としたものをテストケース 1 とし、テストケース 2 は画像 11~20 はテスト用画像、21~30 は評価用画像、31~120,1~10 は訓練用画像のように 10 ずつずらしながらテストケース 12 まで作成し、それぞれ学習、実験を行う(12-分割交差検証)。評価結果はこれら 12 通りの結果を合わせたものである。

4.2. 評価方法

物体検出の成功の判定は、正解ボックス A、検出ボックス B に対し、式(1)によって得られる Jaccard Overlap が 50%以上の場合検出成功とする。この時 B を検出成功ボックスと呼ぶ。

$$\frac{|\boldsymbol{A} \wedge \boldsymbol{B}|}{|\boldsymbol{A} \vee \boldsymbol{B}|} \qquad (1)$$

実験結果は、次の 3 つの指標を用いて評価する。

- 適合率(precision):正と予測したデータのうち、実際に正であるものの割合。誤検出が多いほど低下する。今回の実験では「検出ボックスの総数」に対する「検出成功ボックスの総数」の割合である。

- 再現率(recall):実際に正であるデータのうち、正であると予測されたものの割合。検出漏れが多いほど低下する。今回の実験では「正解ボックスの総数」に対する「検出成功ボックスの総数」の割合である。
- F 値(F-measure):式(2)で求められる、適合率と再現率の調和平均。

$$\text{F値} = \frac{2 \times \textit{適合率} \times \textit{再現率}}{\textit{適合率} + \textit{再現率}} \qquad (2)$$

検出物体クラスのスコアについて閾値 10~90 における適合率、再現率、F 値を算出し、最終的には閾値 70 の値で評価する。

4.3. 実験結果

4.3.1. 複雑背景画像の学習

表 1 上段に、PASCAL VOC データセットのみを学習したモデルの閾値 70 における実験結果を、下段に複雑背景画像を学習したモデルの実験結果を、また実際の検出結果画像の例を図 7 に示す。複雑背景画像を学習したことで再現率が向上し、F 値では 23 ポイント上回った。

表 1:複雑背景画像結果(閾値 70)

	適合率	再現率	F 値
PASCAL VOC のみ	88.00	31.43	46.32
複雑背景を追加	77.01	63.81	69.79

図 7：検出結果の例(class15 は「人」のクラス)
(原画像の出典：
https://upload.wikimedia.org/wikipedia/commons/0/09/JGSDF_22nd_Inf._official.jpg)

4.3.2. 手法 1

本節では手法 1 である、評価用画像に画像変換を施して検出した場合についての実験を行う。表 2 に画像変換ごとの閾値 70 における適合率、再現率、F 値を、図 8 に画像変換ごとの適合率、再現率を示す。閾値 70 の F 値においてメディアン処理が最もよい結果(73.42%)であり、次いでスムージング処理(71.31%)、原画像(69.79%)であった。階調変化、エンボス処理、モノクロ、補色変換の結果は適合率、再現率、F 値ともに大きく下回った。

表 2:閾値 70 における手法 1 結果

画像変換	適合率	再現率	F 値
メディアン	86.45	63.81	73.42
スムージング	85.91	60.95	71.31
原画像	77.01	63.81	69.79
均等化	72.34	64.76	68.34
階調変化	69.62	52.38	59.78
エンボス処理	65.13	47.14	54.69
モノクロ	44.20	29.05	35.06
補色	53.52	18.10	27.05
手法 1	86.96	66.67	75.48

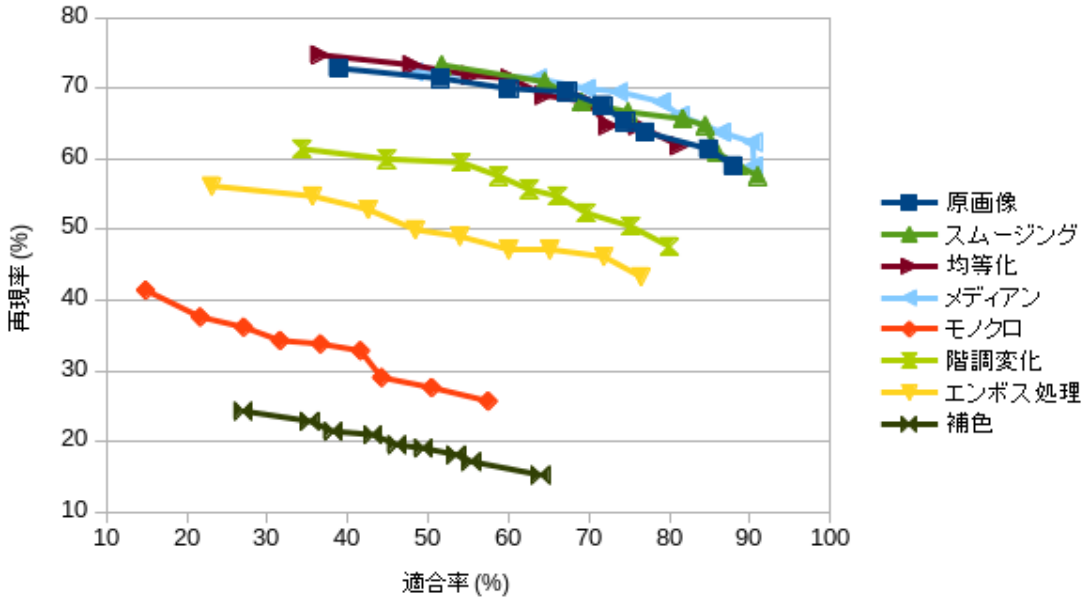

図 8：画像変換ごとの適合率・再現率(手法 1)

ここで、より高い性能を目指すために、メディアン処理での検出結果をベースとして、スムージング、原画像の検出結果の和をとって適合率、再現率を求める実験を行った。この時、誤検出の増加を抑えるため、スムージング、原画像の結果については閾値 90 以上のみの検出結果を採用した。閾値 70 とした時の適合率、再現率、F 値を表 2 の最下段(手法 1)に、手法 1、メディアン処理画像、原画像の適合率、再現率のグラフを図 9 に示す。

メディアン処理単体の結果と比較して再現率が向上し、検出成功ボックス数が増えたために適合率も維持している。調和平均である F 値も 2 ポイント向上している。また、原画像のみの検出結果と比較すると適合率、再現率ともにより向上しており、閾値 70 における F 値は 6 ポイント上回っている。この結果より手法 1 は有効であったといえる。

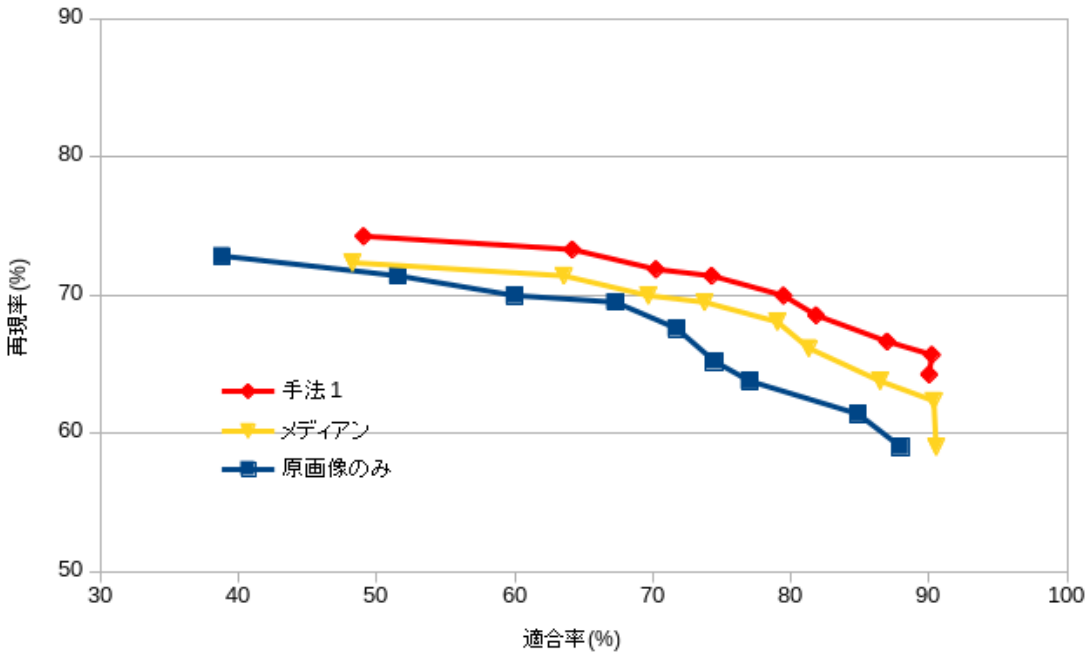

図 9：閾値を変えた時の適合率・再現率(手法 1)

4.3.3. 手法 2

本節では手法 2 である、訓練用画像に画像変換を施し追加学習させた場合についての実験を行う。図 10 に原画像のみを学習した学習モデル(4.3.1.節の実験結果)と変換画像をすべて追加学習した学習モデルで、原画像を評価用画像として与えた場合の適合率と再現率を示す。訓練用画像に画像変換を施して検出した結果では各閾値の結果において適合率は向上していたが、再現率が 10 ポイントほど減少した。この結果は、誤検出は減少していたが検出漏れが増えていることを表す。これは、原画像のみの学習では検出できていたボックスが、変換画像を追加学習したことで検出に失敗したためである。

続いて、変換画像を追加学習した学習モデルを用いて、手法 1 と同様に評価用画像にも画像変換を施した実験を行った。閾値 70 における結果を表 3 に、閾値を変えた時の適合率、再現率を図 11 に示す。結果は均等化が最も高く(70.30%)、続いてスムージング(65.88%)、原画像のみ(63.03%)の結果となった。手法 1 とは異なり補色変換の結果も比較的高い数値となっている。しかし、手法 2 において最も高かった均等化画像の検出結果は、手法 1 において最も精度の高かったメディアン処理の結果(表 2)と比較して適合率、再現率ともに低い値となっている。

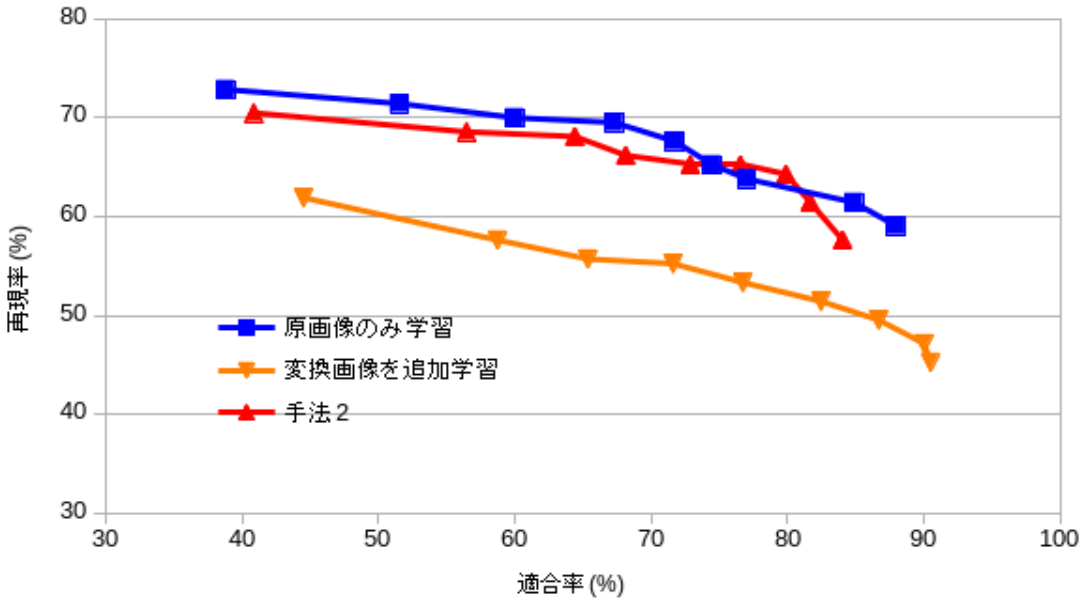

図 10：閾値を変えた時の適合率・再現率(手法 2)

検出精度を向上させるために、均等化画像の検出結果をベースとして、スムージング、原画像、補色の検出結果の和をとった結果を求めた。手法 1 と同様に誤検出の増加を抑えるため、スムージング、原画像、補色の結果については閾値 90%以上の結果のみを採用した。実験結果を表 3 の最下段(手法 2)と図 10 に示す。均等化画像のみの結果と比較して F 値が 1 ポイント向上しているが、手法 1(表 2)と比較して結果は下回っていた。この結果から手法 2 は有効ではなかったといえる。

表 3:閾値 70 における手法 2 結果

画像変換	適合率	再現率	F 値
均等化	82.17	61.43	70.30
スムージング	86.15	53.33	65.88
原画像	86.67	49.52	63.03
補色	78.79	49.52	60.82
メディアン	80.17	46.19	58.61
階調変化	80.67	45.71	58.35
エンボス処理	71.09	43.33	53.84
モノクロ	62.20	37.62	46.88
手法 2	79.88	64.29	71.24

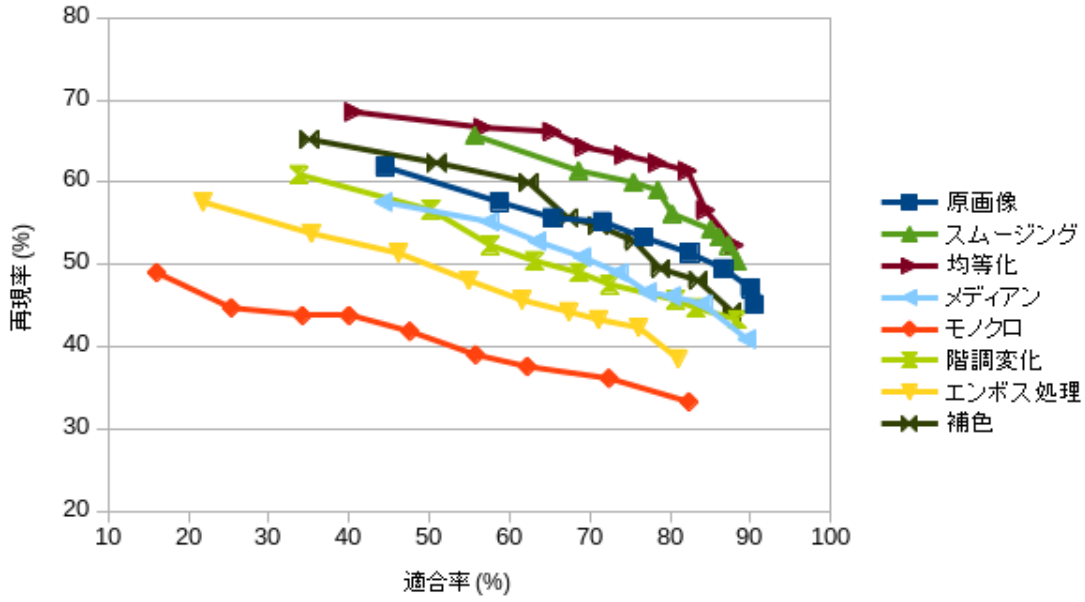

図 11：画像変換ごとの適合率・再現率(手法 2)

4.4.　一般背景画像での実験

本研究では複雑背景中の物体検出性能の向上を想定して手法を考案した。画像変換によって検出性能が向上した結果が、複雑背景特有であるかを評価するため、一般背景に対する実験を行った。

4.4.1.　実験の手順

PASCAL VOC データセットの訓練用画像(5,011 枚)を学習した学習モデルで、同データセットの評価用画像(訓練用画像と異なる画像 120 枚)について、原画像のままの場合と、メディアン処理を施した場合の検出結果を比較する実験を行う。評価対象は「人」クラスである。このデータセット内の画像は背景と検出対象物体の区別がつきやすいものであり、一般背景画像である。画像の例を図 12 に示す。ここで、メディアン処理は複雑背景で実験した手法 1 において、原画像よりも性能が高かった画像変換である。

また、手法 1 と同様に原画像での検出結果と、メディアン処理を施した画像で閾値を 90 に設定した検出結果との和をとって適合率、再現率を求める実験も行う。

図 12：一般背景画像例(PASCAL VOC データセットより)

4.4.2. 実験結果

閾値 70 における実験結果を表 4 に示す。適合率はともに 100%であるが、複雑背景で実験した場合とは異なり、メディアン処理を施したことにより性能が向上することはなく、むしろ悪化していることがわかる。また、閾値を 90 に設定したメディアン処理での検出結果の詳細を確認すると、原画像で検出できなかった検出対象物が新たに検出できた例はなかったため、和をとっても再現率は向上せず、性能はよくならなかった(表 4　原画像、メディアン処理の和(閾値 90))。閾値を 70%に設定したところ、F 値は 0.85 ポイント向上した(同　閾値 70)。複雑背景を対象とした 4.3.2 節の手法 1 では、原画像のみの結果より 5.69 ポイント向上していることから、一般背景では画像変換による効果は乏しいといえる。

これらの結果から一般背景画像において画像変換は有効ではなく、複雑背景での検出特有であることが確認できた。また一般背景画像では、表 1 の結果よりも F 値が高いことから複雑背景における物体検出が一般背景よりも難しいことも確認できた。

表 4:閾値 70 における一般背景での実験結果

	適合率	再現率	F 値
原画像	100.00	66.27	79.71
メディアン処理	100.00	61.83	76.41
原画像、メディアン処理の和 (閾値 90)	100.00	61.83	76.41
原画像、メディアン処理の和 (閾値 70)	100.00	67.45	80.56

4.5. 回転画像での実験

メディアン処理やスムージングのような画像変換が複雑画像に対して有効であるか確認するために、データ拡張で一般的である回転処理を施した画像での検出結果と比較した。

4.5.1. 実験手順

複雑背景画像を 90 度、180 度、270 度回転した画像を作成し、それらを評価用画像として検出する。さらに、回転していない原画像の検出結果をベースとして、回転画像の検出結果の和をとって適合率、再現率を求める実験を行った。手法 1,2 と同様、誤検出の増加を抑えるため、回転画像の検出結果については閾値 90 以上のみの検出結果を採用した。

4.5.2. 実験結果

表 5 に回転画像の閾値 70 における適合率、再現率、F 値を、図 13 に原画像のみを検出した結果(4.3.1.節の実験結果)と、手法 1、手法 2 の結果と回転画像の結果を比較したグラフを示す。

原画像のみの検出結果と比較して回転画像では再現率が 2 ポイント増加しているが、適合率が大きく低下している。これは単に回転処理でデータを水増ししただけでは検出漏れを減少できず、さらに誤検出を増やしてしまうということを示している。この結果から手法 1 で行ったメディアン処理、スムージングが

検出性能の向上に有効であったことが確認できた。

表 5:閾値 70 における回転画像での実験結果

	適合率	再現率	F 値
原画像	77.01	63.81	46.32
手法 1	86.96	66.67	75.48
回転画像	50.89	65.24	57.18

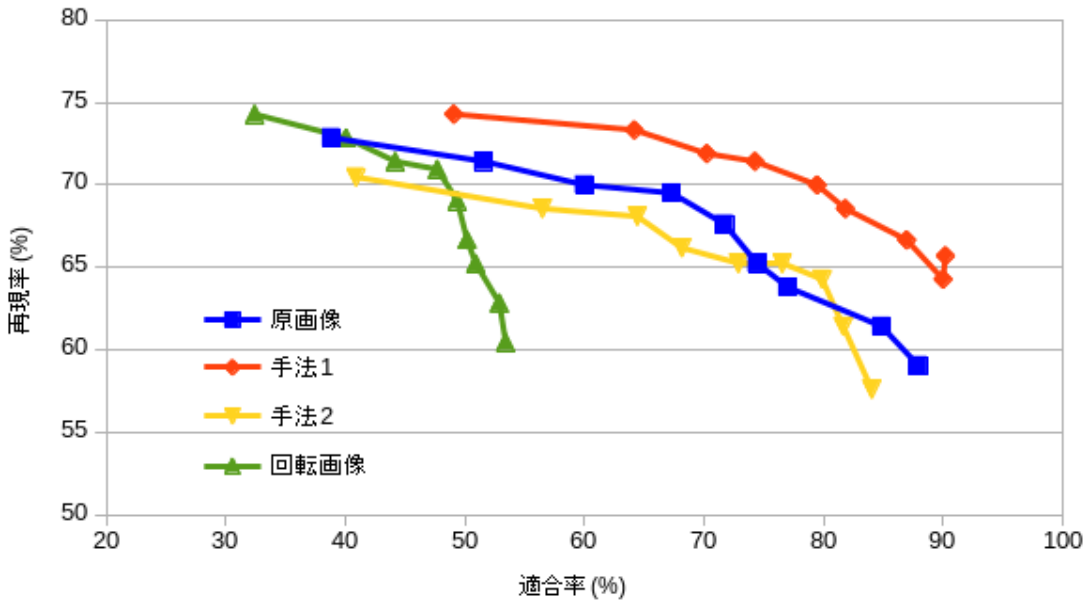

図 13：閾値を変えた時の適合率・再現率(回転画像)

4.6. 考察

手法 1 の実験結果から、評価用画像に画像変換を行うことで、原画像のみでは検出できなかった物体の検出に成功した例が存在し、複雑背景での物体検出に効果があることが分かった。ここでモノクロ、補色のように色情報が減少、または大きく変化する画像変換では性能が下がる結果となり、メディアン処理やスムージングといった画像のノイズを軽減させる画像変換では性能が向上した。また、メディアン処理、スムージング、原画像での検出結果の和をとることでさらに再現率が向上した。しかし、画像変換後の検出結果の組み合わせ次第では誤検出も大幅に増加するため適合率が下がり、結果的に F 値が低い結果となることもある。検出結果の最適な組み合わせを追求する必要がある。

手法 2 の実験結果から、訓練用画像に画像変換を施して追加学習した学習モデルでは、原画像のみを学習した学習モデルでの検出と比較して検出漏れが多い結果となっていた。さらにこの学習モデルに、均等化、スムージング、原画像、補色処理を施した画像を評価用画像として与えた結果の和をとると性能は向上したが、手法 1 の結果には及ばなかった。これは補色、モノクロなど、原画像と比較して大きく見た目が変わる画像を同時に学習させたことで、かえって検出対象物らしさが曖昧になってしまった可能性がある。

5. おわりに

本研究では、深層学習による物体検出器と、画像変換を用いることで、複雑背景における物体検出の性能を改善する手法を提案した。検出時に評価用画像にメディアン処理、スムージングを施し、原画像での検出結果との和をとることで性能が向上した。しかし一般背景画像において人物を検出した結果には及ばない結果となっている。画像変換ごとの検出結果のより良い組み合わせによるさらなる検出性能の向上や、本研究で検出対象とした人物クラス以外の対象での提案手法の評価が今後の課題である。

謝辞

本研究を進めるにあたり、ご指導をいただいた卒業論文指導教員の椋木雅之教授に感謝いたします。指導教員である椋木雅之教授には、本研究で使用した SSD を導入するにあたっての環境設定や、論文に関する助言やご指導を沢山いただきました。また、SSD を提案された Liu Wei さんを始めとする研究者の皆様に感謝いたします。本研究では、画像データセットとして Pascal VOC を、SSD の実装には github にて提供されているものを使用させていただきました。両サービスの主催者の皆様に感謝いたします。最後に、椋木研究室の皆様、お忙しい中数々の助言やアドバイスをしていただきありがとうございました。

参考文献

[1] G.Ross, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. CVPR 2014, pp.580-587, 2014.
[2] R.Joseph, et al. You only look once: Unified, real-time object detection. Proc. CVPR 2016, pp.779-788, 2016.
[3] L.Wei, et al. SSD: Single Shot MultiBox Detector. Proc. ECCV 2016, LNCS 9905, pp.21-37, 2016.
[4]赤松龍太, 他. 画像変換を用いた多数の候補領域抽出と検証による物体検出, 情報処理学会, 火の国情報シンポジウム 2016, 1B-1, 2016.
[5]Pascal VOC, http://host.robots.ox.ac.uk/pascal/VOC/ (2018/02/09 アクセス).