The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

半教師あり学習による商品画像中の個数と位置の同時推定

Estimation of Number and Locations of Products in Pictures by Semi-Supervised Learning

藤橋 一輝[*1] Kazuki Fujihashi　木村 雅之[*1] Masayuki Kimura　金崎 朝子[*2] Asako Kanezaki　小澤 順[*2] Jun Ozawa

[*1] パナソニック株式会社 Panasonic Corporation　[*2] 国立研究開発法人 産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

Abstruct: We propose a semi-supervised method for estimating the number and locations of products in pictures. Many existing approaches can estimate objects locations in images by supervised learning which needs images annotated with objects locations. On the other hand, our method needs only numbers of objects in images. The experiment shows effectiveness of our method.

1. はじめに

Deep Neural Network（DNN）を用いて，画像中の物体認識を行う手法が多く提案されている．中でも，特定の物体の種類の認識だけでなく，画像中のどこに位置しているかを提示する物体検出手法は，ロボティクス，医療，スポーツ，セキュリティ等，様々な分野での応用が期待され，注目されている．

DNN を用いた物体検出手法として，Faster R-CNN[Ren et al., 2016]が提案されている．Faster R-CNN は end-to-end 学習により物体検出を行うことで，精度よく高速な検出を可能とした．また，Faster R-CNN より高精度，高速な物体検出手法として，SSD [Liu et al., 2016]，YOLO [Redmon et al., 2017]が提案されている．さらに，Faster R-CNN，SSD，YOLO が物体位置を矩形で提示するのに対し，Mask R-CNN [He et al., 2017]は画素レベルでの物体位置の提示を可能とした．

前述の物体検出手法は非常に優れた性能を誇るものの，実用にあたり障害となる点がある．それは，使用する DNN の学習に用いるデータとして，認識したい物体の画像データと，画像中の物体の位置，大きさなどの詳細なアノテーションが必要である点である．一般に，画像中の物体の位置，大きさなどの情報を自動で抽出することは困難であり，これらの情報は人手で付与する必要がある．しかし，画像データの量が多いほど，あるいはシーンが複雑になるほど，このような人手でのアノテーション付与にかかる負担は大きくなる．特に店舗内の商品の認識などの，認識対象の改廃が頻繁に発生するようなケースにおいてその負担はピークに達するであろうことは想像に難くない．そこでこのような負担を軽減するために，詳細なアノテーションを用いない学習手法が求められる．

本研究では，多くの従来手法で必要とされる画像中の物体の位置，大きさを必要としない，物体検出手法を提案する．提案手法では，画像中の物体の個数のみを用いる半教師あり学習を行い，画像中の任意の領域に含まれる物体の個数を推定する．さらにこの結果を利用して物体の位置を推定する．

2. 提案手法

提案手法では，あらかじめ入力として，画像と画像中の物体の個数のみを与え，畳み込みニューラルネットワーク（CNN）の学習を行う．学習した CNN を用いた処理を行うことで，物体の個数，位置が未知である画像を与えたとき，画像中の物体の個数，位置を出力する．以下 2.1 節で CNN の学習手法について，2.2 節でテスト時における画像中の物体の個数，位置の推定手法について述べる．

2.1 CNN の学習

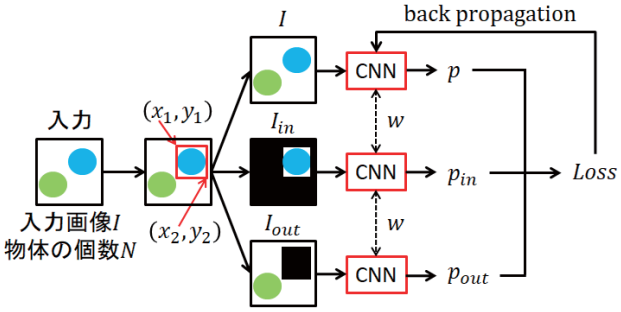

図 1：CNN の学習の流れ．各 CNN は重みwを共有する．

CNN の学習の流れを図 1 に示す．入力画像をI，画像中の物体の個数をN，(x, y)におけるIの画素値を$I(x, y)$とする．Iに対し画像中の(x_1, y_1)，(x_2, y_2)の 2 点をランダムに与え$(x_1 < x_2,\ y_1 < y_2)$，画像I_{in}，I_{out}をそれぞれ式(1)，式(2)により生成する．

$$I_{\text{in}}(x_i, y_i) = \begin{cases} I(x_i, y_i) & if\ \ (x_i, y_i) \in U \\ 0 & otherwise \end{cases} \quad (1)$$

$$I_{\text{out}}(x_i, y_i) = \begin{cases} 0 & if\ \ (x_i, y_i) \in U \\ I(x_i, y_i) & otherwise \end{cases} \quad (2)$$

$$(U = \{(x, y) | x_1 \le x \le x_2,\ y_1 \le y \le y_2\})$$

次に，I，I_{in}，I_{out}を入力として CNN により畳み込み処理を行い，それぞれの出力p，p_{in}，p_{out}を得る．CNN のネットワークとしては，VGG16 [Simonyan et al., 2015]の最終出力ノードを 1 つとしたものを用いる．そして，損失関数Eを式(3)で定義し，CNN の学習を行う．

$$E = (p - N)^2 + \left((p_{\text{in}} + p_{\text{out}}) - N\right)^2 \quad (3)$$

2.2 画像中の物体の個数，位置の推定

未知の画像中の物体の個数，位置の推定の流れを図 2 に示す．まず，入力画像Iに対し，2.1 節で学習した CNN を用いて，画像中の物体の推定個数N_{pred}を得る．次に，画像Iに対して Selective Search [Sande et al., 2011]を用いて，複数の物体位置候補矩形を得る．そして，矩形ごとに 2.1 と同様に式(1)によりI_{in}，

連絡先：藤橋 一輝，所属：パナソニック株式会社，〒108-0075 東京都港区港南 4-1-8 リバージュ品川 14 階，TEL：03-5796-2565，E-mail：fujihashi.kazuki@jp.panasonic.com

I_{out}を生成し，2.1 で学習した CNN を用いて，矩形内外の物体の推定個数$N_{\text{pred_in}}$，$N_{\text{pred_out}}$を得る．最後に，$N_{\text{pred\_in}}$，$N_{\text{pred\_out}}$，式(4)により定義されるIoU(Intersection over Union)をもとに，N_{pred}個の矩形を物体の位置として提示する．ここで，$S(A)$は領域Aの面積を表す．矩形提示の処理をアルゴリズム 1 に示す．

$$\text{IoU} = \frac{S(A_g \cap A_e)}{S(A_g \cup A_e)} \tag{4}$$

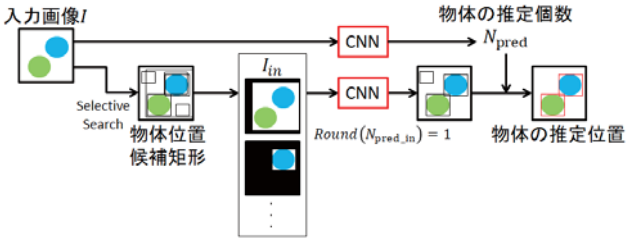

図 2:物体の個数，位置推定の流れ

アルゴリズム 1:矩形提示の処理

```
Input : Input Image I;
        Rect candidates Rc;
        Count estimation in the Image I : Npred
Output : Predicted Rect Rp;
Initialization : a number of Rect candidates Nc;
                 res ← 0.
for iter1 = 1 to Npred do
  if iter1 > 1 then
    I ← Make_Iout(I, Rc(Rp_index_list(iter1 - 1)));
  end
  Initialize : output_in_list, output_out_list, dist_list;
  for iter2 = 1 to Nc do
    Iin ← Make_Iin(I, Rc (iter2));
    Iout ← Make_Iout(I, Rc (iter2));
    output_in ← CNN(Iin);
    output_out ← CNN(Iout);
    dist_list(iter2) ← Abs(output_in - 1) + Abs((output_out
      + output_in + res) - output_whole) + S(Rc(iter2));
    output_in_list(iter2) ← output_in;
    output_out_list(iter2) ← output_out;
  end
  iter3 ← 1;
  IoU ← 1.0;
  while IoU > 0.5 then
    Rp_index ← ArgSort(dist_list)(iter3);  # ascending order
    IoU ← CalculateIoU(Rc(Rp_index), Rp);
    iter3 ←iter3 + 1;
  end
  Rp_index_list(iter1) ← Rp_index;
  Rp (iter1) ← Rc (Rp_index);
  res ← res + output_in_list(Rp_index);
end
return Rp.
```

3. 実験

3.1 実験条件

図 3:商品画像データセットの画像の例

実験には，3D スキャナ(Ortery 社 3D MFP)で取り込んだ一般的な商品の CG データを用いて作成した CG 画像と背景画像を合成した商品画像データセットを用いた．CG 画像は 3DCG ソフトを用いて種類・数量・配置をランダムに設定したものを事前にレンダリングして用いている．なお、画像に含まれる商品の種類は 9 種類で、各画像には 2 から 8 個の商品が種類の重複を許して出現する．一方，背景画像は実物のショッピングバスケットに白または黒の発泡スチロール板を敷いたものをランダムに配置して撮影されたカゴ画像 93 枚である．合成時には，CG 画像に対し，背景画像としてランダムに選出したカゴ画像に色彩変化を加えた上で処理を実施した．商品画像データセットの画像の例を図 3 に示す．商品画像データのうち 5948 枚を学習用，661 枚を評価用に分割し，学習用画像を用いて 2.1 節の手法で CNN を学習し，評価用画像に 2.2 節の手法を適用した．

なお，学習および評価プログラムは PyTorch を用いて実装し，learning rate=0.0001, batch size=100 で 50 epochs の学習を行った．

3.2 実験結果

提案手法を用いた物体位置推定結果の例を図 4 に示す．提案手法の評価として，画像中の正解の位置矩形領域をA_g，推定の位置矩形領域A_eとしたとき，IoUの平均 Mean IoU，Mean AP(Average Precision) [Everingham et al., 2009]，個数の正答率，平均誤差 を算出した．

また，教師あり学習である Faster R-CNN，弱教師あり学習である SPN [Zhu et al., 2017]を商品画像データセットに適用した．提案手法と既存手法の評価結果を表 1 に示す．

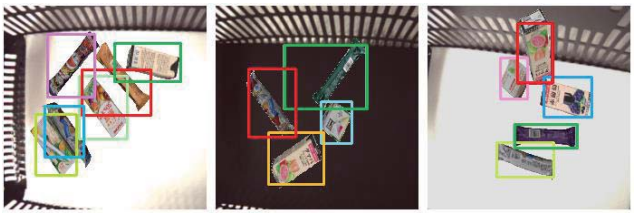

図 4:物体位置推定結果の例

表 1:実験結果の各評価値．教師あり 1 は物体クラスを与えた場合，教師あり 2 は物体クラスをすべて同一とした場合．

	Mean IoU	Mean AP	個数の正答率	個数の平均誤差
Faster R-CNN（教師あり 1）	0.918	0.995	0.939	0.061
Faster R-CNN（教師あり 2）	0.922	0.979	0.770	0.230
SPN（弱教師あり）	0.313	0.020	0.209	1.442
提案手法	0.687	0.504	0.989	0.011

3.3 考察

図 5:物体位置推定が不良な例

表 1 より, 提案手法が画像中の物体の個数, 位置の推定に有効であることが確認できた. ただし, 図 5 に示すように, 物体位置推定が不良な場合も見られた. 不良な場合は①矩形内の物体は 1 個であるが, 矩形提示が物体に対して非常に大きい, ②1 つの物体に対して矩形が 2 つ提示されており, 物体検出漏れがある, ③物体が密な箇所で, 矩形提示に誤りが見られ, それに応じる形で誤りが発生する, の 3 つに大別された. ①～③の不良の原因としては, Selective Search の結果で物体領域候補矩形がうまく得られていなかった, そもそも物体が密な場合には物体 1 つだけが囲われるような矩形をとれない, といったことが考えられる.

4. おわりに

画像中の物体の個数のみを画像のラベルとして与えた CNN の学習を行い, CNN を用いた処理を行って画像中の物体の個数, 位置の推定を行う手法を提案した. 独自の商品画像データセットを用いた実験により, 本手法の有効性が示された. 本手法でアノテーションとして用いる個数情報は比較的容易に取得可能であるケースが多く, 本手法を用いることで学習データを用意するコストを抑えた上で物体計数や位置推定を行うことが可能となり, より幅広いシーンでの利用が期待される. 今後の課題として, 別のデータセットにおける検証や, 物体位置推定手法において物体候補矩形選定の最適化, 単独の CNN で完結するいわゆる end to end での推定手法の検討, 物体が密に位置する場合や重なりがある場合での良好な矩形提示などが挙げられる.

参考文献

[Ren et al., 2016] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

[Liu et al., 2016] Wei Liu, Dragomir Anguelov, Dumitru Erthan, Christian Szegedy, Scott Reed,Cheng-Yang Fu, Alexander C.Berg: SSD: Single Shot MultiBox Detector, European Conference on Computer Vision, 2016.

[Redmon et al., 2017] Joseph Redmon, Ali Farhadi: YOLO9000: Better, Faster, Stronger, Computer Vision and Pattern Recognition, 2017.

[He et al., 2017] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick, Mask R-CNN, IEEE Internatinal Conference on Computer Vision , 2017.

[Simonyan et al., 2015] Karen Simonyan, Andrew Zisserman: Very Deep Convolutional Networks for Large-Scale Image Recognition, International Conference on Learning Representations, 2015.

[Sande et al., 2011] Koen E. A. van de Sande, Jasper R. R Uijlings, Theo Gevers, Arnold W. M. Smeulders: Segmentation as Selective Search for Object Recognition, IEEE Internatinal Conference on Computer Vision, 2011.

[Everingham et al., 2009] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman: The Pascal Visual Object Classes(VOC) Challenge, Int J Comput Vis(2010) 88. 303-338, 2009.

[Zhu et al., 2017] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, Jianbin Jiao: Soft Proposal Networks for WeakLy Supervised Object Localization, IEEE Internatinal Conference on Computer Vision, 2017.