CNN の逆伝搬を利用した食事画像の領域分割

下田 和[1]　柳井 啓司[1]

1. 導入

これまで，日々の食事のカロリー計算を容易に行うといった目的のために，食事画像の認識が行われてきた [1]．従来の食事画像認識は多くの場合，画像 1 枚に食事 1 品を想定していたため，定食などの複数品が含まれる場合にも主菜のみしか認識出来なかった．より精度の高いカロリー推定を実現するには，食事画像の自動セグメンテーションを行い一品ずつ種類を認識することが望ましい．

食事画像の認識には，一般的な画像認識と同様に SIFT や HOG などの特徴量が主に使われてきた．しかし，近年は Deep lerning による Covolutional newral network(CNN) 特徴量 [2] が注目されている．CNN の活性化の状況を視覚的に表現する手法として，CNN の逆伝搬を利用する方法がある．近年，この視覚化のための逆伝搬を応用したセグメンテーションの手法が発表されている [3]．Simonyan らはバックプロパゲーションとグラブカットを用いたセグメンテーションを提案した．この手法は，既存の高精度なセグメンテーション手法 [4] などと異なり，ピクセルレベルでのアノテーションを必要とせず，バウンディングボックスも不要である．一般的な画像単位にラベル付された学習データから学習した CNN を用いて，セグメンテーションを行うことができるという特徴がある．

本研究では, Simonyan らの手法 [3] と, Regions with CNN features (RCNN) [5] で使われている selevtive search [6] を用いて，食事画像のセグメンテーションを行う．

2. 概要

まず，selective search [6] を用いて，約 2000 のバウンディングボックス候補を抽出する (図 1-(1))．次に，バウンディングボックスのクラスタリングを行い，2000 のバウンディングボックスを約 100 の領域候補にグループ化する (図 1-(2))．

次にこの各領域候補について, 15 のバウンディングボックスについて逆伝搬とグラブカットによるセグメンテーションを行う (図 1-(3), (4), (5))．セグメンテーションした結果のアウトラインからバウンディングボックスを再設定し，これを CNN で認識をして, 最後に Non-Maximum Supression (NMS) を行い統合し，最終結果を得る (図 1-(6))．

3. 手法の詳細

3.1 セレクティブサーチ

本手法では逆伝搬により得られたサリエンシーマップを用いてグラブカットを行う．グラブカットは，サリエンシーマップ内の複数のオブジェクトにピークがある場合，オブジェクトが繋がってセグメンテーションされてしまう．そこで，このグラブカットの結果を改善するために，本研究では，最初に selective search [6] を用いた．Selective search は，約 2000 の物体猟奇候補のバウンディングボックスを提案する．領域を制限することにより，複数のオブジェクトについてサリエンシーマップのピークが出てくる可能性を抑えることが期待できる．

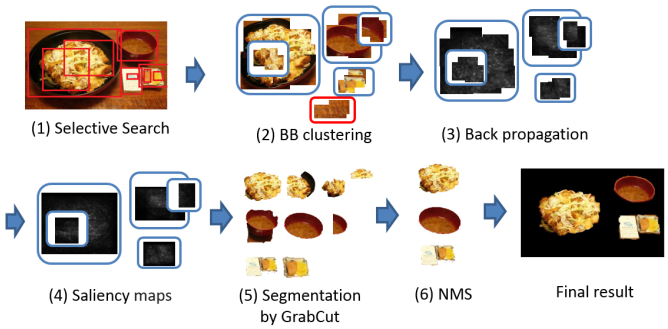

図 1　処理の流れ.

3.2 バウンディングボックスクラスタリング

Selective search により得られた 2000 のバウンディングボックスについてバウンディングボックスクラスタリングを行い，各クラスタを領域候補とする．ここで各領域候補について，15 個以上のバウンディングボックスの割り当てられなかった場合は除去する．クラスタリングはオーバラップの割合を類似度として，類似度に基づいてグループ化を行った．ここでは，各領域候補に属するバウンディングボックスのうち，面積の大きい順にそれぞれ 15 個のバウンディングボックスを選び，次のステップの逆伝搬に用いることとする．

3.3 逆伝搬によるセグメンテーション

Simonyan ら [3] は，CNN の学習時に用いる Back Propagation (BP) を入力画像のレイヤーまで行うことで，入力画像に対する指定したカテゴリの存在可能性を示す heat map が作成できることを示した．この処理は，一枚の画像とそのカテゴリを与えて，順伝搬と逆伝搬を行うだけで可能であり，瞬時に計算することが可能である．

Simonyan らは 1 枚の画像毎にこの処理を行ったが，本研究では前のステップで選ばれたバウンディングボックス全てについて，この処理を行い，それぞれのバウンディングボックスについて図 1(3) のような heat map を得た．なお，逆伝搬時に与える教示信号は，順伝搬時に得られた結果の上位 5 候補を 1，それ以外を 0 として，与えた．

なお，逆伝搬による視覚化の方法としては，BP の他に，Deconvolution [7] や，Guided back propagation (Guided BP) [8] がある．これらの逆伝搬は，それぞれ同様に，CNN が認識の際に反応した入力画像の部分を可視化することができるが，ReLU レイヤーの逆伝搬時の処理の違いから得られる結果に違いがある．本研究ではこの 3 つの逆伝搬法を用いて heat map を得た．

次に，各領域候補クラスタ毎に，得られた 15 枚の heat map のをピクセルごとに平均して，領域候補毎に食事の saliency map を作成する．このマップは，各バウンディングの認識結果上位 5 位の出力値で重み付けして，食事種類ごとに行う．

次に，各領域候補毎に，15 個のバウンディングの出力値の平均の上位 5 種類について，saliency map を利用して

[1] 電気通信大学大学院 情報理工学研究科 総合情報学専攻

GrabCut を実行し，食事領域抽出を行う．図 1(5) のような結果が得られる．

3.4 CNN 特徴量の抽出と NMS

各領域候補は互いにオーバラップすることもあるので，セグメンテーション結果の領域を囲む部分画像を CNN で再度認識し，この時の出力値を使って領域候補について Non-Maximum Supression (NMS) を行い，ピクセルレベルで領域を統合した．

4. 実験

UECFOOD100 [9] と Pascal VOC 2007 の二つのデータセットについて，物体検出の評価を行った．なおここでは，UECFOOD100 の評価データにピクセルレベルでのアノテーションがなく，バウンディングボックス (BB) レベルでのアノテーションしかないため，提案手法に付いては領域抽出まで行った上で，BB レベルでの検出率評価を行った．評価方法は Pascal VOC detection task と同様で，推定 BB と正解 BB が 50 正解とカウントし，平均適合率 (Average Precision, AP) の全カテゴリの平均 mAP (mean AP) で評価した．

まず，食事画像に対する実験として，UECFOOD100 データセットにおける複数品の食事画像 1174 枚について評価を行った．RCNN [5] と提案手法の比較を行った結果，特に食事画像において RCNN より大幅によい精度となった (表 1)．また，BP, Deconv, Guided BP の比較を行ったが，BP が他よりも若干，良い精度となった．図 2 に検出結果の例を示す．図 3 には領域抽出結果を示す．

UECFOOD mAP	100class	53class	11class
RCNN	26.0	21.8	25.7
Ours(GBP)	**50.7**	52.5	51.4
Ours(Deconv)	48.0	54.1	**55.4**
Ours(BP)	49.9	**55.3**	**55.4**

表 1 UECfood の複数食事画像データセットにおける RCNN と提案手法の比較．UECfood は 100 カテゴリあるが，複数品食事画像に含まれる各カテゴリの物体数に偏りがあったので 3 段階に分けて評価を行った．100class は objcect num >= 0，53 クラスは objcect num >= 10 11class は objcect num >= 50

次に食事以外の一般物体に付いて，Pascal VOC 2007 について RCNN と提案手法について検出精度の比較を行った．Pascal VOC 2007 においても，本手法により約 4 %精度が改善した (表 2)．

Pascal VOC 2007	mAP
RCNN	54.2
Ours(BP)	58.7

表 2 Pascal VOC 2007 データセットにおける 20 カテゴリの mAP

5. 考察

食事画像の物体検出において，RCNN は低い精度になった．これは食事画像が一般的な物体と比較して，一部分の認識で高い認識率になるためであると考えられる．テクスチャのように，一部からでもカテゴリの判断ができる．そのため，図 2 に示すように，小さいバウンディングボックスが検出結果として出力される傾向があった．

しかし，提案手法ではよい精度のまま検出を行うことができている．また，食事画像においては Pascal と比べてセグメンテーションが容易であったため，提案手法がより適していた可能性も考えられる．

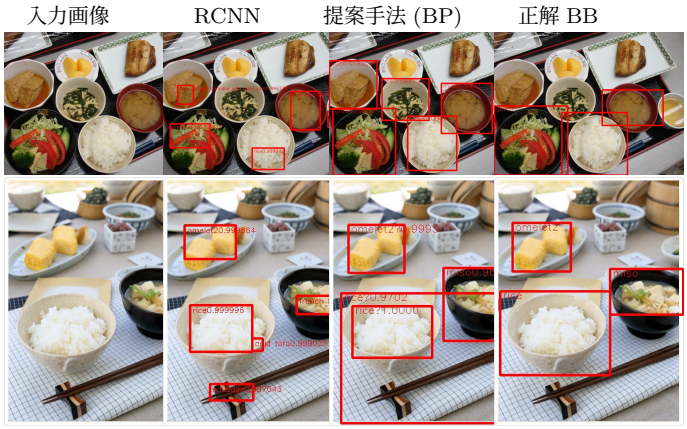

図 2 RCNN と提案手法の検出結果の比較．

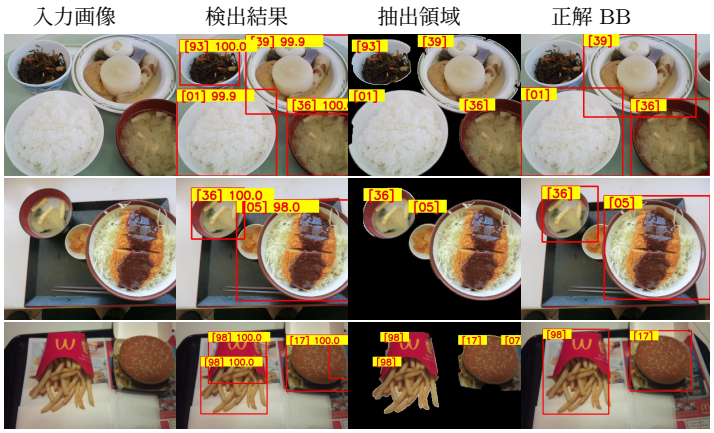

図 3 提案手法による出力結果例．図中の (ID:食事) は，(01:ご飯), (05:カツどん), (17:ハンバーガー), (36:味噌汁), (39:おでん), (93:きんぴら), (98:フライドポテト) を表す．

参考文献

[1] Y. Kawano and K. Yanai. Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications*, pp. 1–25, 2014.

[2] A. Krizhevsky, I. Sutskever, and G E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[3] K. Simonyan, A. Vedaldi, and A Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

[4] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[6] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.

[7] M. Zeiler, G. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature. In *ICCV*, 2011.

[8] J. T. Springenberg, A. Dosovitskiy, T Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR 2015 Workshop Track*, 2015.

[9] Y. Matsuda, H. Hoashi, and K. Yanai. Recognition of multiple-food images by detecting candidate regions. In *ICME*, 2012.