社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

弱教師学習手法を用いたWebからの食事検出器の自動学習

下田　　和†　　柳井　啓司†

† 電気通信大学 大学院情報理工学研究科 総合情報学専攻

あらまし　これまで、物体検出や領域分割などの画像認識タスクの精度向上のために、様々な物体の候補領域をObjectnessとしてプロポーザルする手法が研究されてきた。しかし、多くの既存手法は画像内における物体を網羅するために、大量の領域候補を提示する。近年の深層学習においては精度向上とともに計算コストが増加しており、これらのことを考慮すると、大量の領域候補を認識する手法は計算速度の面で今後問題を抱えることが予測される。一方で、近年領域候補を提案する手法を用いずに画像における物体の位置を直接推定することが可能であるFully Convolutional Network(FCN)が提案された。この手法は高速かつ高精度な物体の位置の推定が可能であり様々な分野のタスクで応用されている。しかし、FCNにおける物体の位置推定手法は、学習画像とテスト画像のドメインが異なる場合、著しく精度が低下してしまう場合がある。近年、深層学習を実現するために必要な画像枚数が問題になる場合があり、これを解決する手法として、Web上に大量に存在するWeb画像を用いて学習を行うWeb教師学習が注目されている。このWeb教師学習から実世界画像における認識を行う場合、学習画像とテスト画像のドメインが異なることが想定される。学習画像とテスト画像におけるドメインの違いにおいて頑強な位置推定手法が必要とされている。そこで、本研究においては、Web教師学習による食事領域候補をプロポーザルする手法を提案する。特に、単品のみしか映っていないWeb食事画像で学習を行い、複数品目の映るテスト画像において有効な手法を想定し、ドメインの変化に頑強な手法を実現するためにプロポーザルによる認識を行う。ただし、少ない領域候補で正確な食事検出を行うことに着眼し、なるべく精度を下げずに計算コストを低下させる。

キーワード　Convolutional Neural Network，領域分割

Webly-supervised Automatic Construction of Food Detectors

Wataru SHIMODA† and Keiji YANAI†

† Department of Informatics,　The University of Electro-Communications,　Tokyo

1. はじめに

近年、食事記録は健康の指標として重要視されるようになった。食事の記録により、ユーザーは様々なカロリーや滋養の統計を確認することができ、これらの統計は食生活の偏りなどの解決に役立つ。食事記録はユーザーにとって役立つものであるが、いくつかの手間が必要である。特に、手動によるテキストでの食事記録でこの問題は顕著になることが予測され、食事記録をより簡略化する方法が必要である。

食事認識は食事記録の簡略化において、重要な役割を担っている。食事の記録を写真をとるだけで行うことができれば、ユーザーの負担を大きく軽減できる。また、近年のDeep Convolutional Neural Network (DCNN)の進歩により、食事画像の認識精度は飛躍的に向上しており、高精度な食事の認識が可能である。

食事認識において物体検出や領域分割はともに重要なタスクである。物体検出は対象の物体にバウンディングボックスを付与するタスク、領域分割はピクセルレベルで物体の位置を推定するタスクである。これらのタスクは対象の物体の大きさを推定することにつながる。食事のサイズは食事の量を反映しているはずであり、これらのタスクを用いることで、より高い精度でカロリー推定ができる可能性がある。

近年、物体検出や領域分割においてもDCNNの進歩とともに大きく精度が向上した。しかしながら、多くのDCNNに基づいた物体検出や領域分割の手法は、バウンディングボックスやピクセルレベルのアノテーションなど高度な教師情報を必要とする。一方で、Web画像の存在などから、画像レベルのアノテーションは比較的に容易に手に入れることができる。一般に、高度な教師情報を必要とする学習を完全教師あり学習、必要としない学習を弱教師あり学習と呼ぶ。近年は、様々な弱教師ありにおける物体検出手法や領域分割手法が提案されている。

既存の弱教師あり領域分割手法の多くは、PASCAL VOCデータセットのみで実験されている。パスカルのデータセットの学習画像がマルチラベル画像であるのに対して、Web画像の多くはシングルラベル画像であり、既存の弱教師ありの位置

推定手法の多くは Web 画像を用いることを想定していない。Web 画像のみで学習を行うことができれば、学習データの収集はより容易となる。Web 画像のみを用いて学習する手法は、Web 教師学習と呼ばれ、近年注目されている [1]。本研究では、食事 Web 画像のみの教師情報を用いて、食事の検出を行う。

特に、弱教師あり領域分割手法において高精度を達成した手法である DCSM [2] をベースにし、これを Web 教師学習に応用した。DCSM は学習とテストにおけるドメインの変化に対して頑強でなく、Web 教師学習において有効ではなかった。そこで、本研究では、この DCSM を食事のためのプロポーザルに応用し、RCNN や SDS といった従来手法に倣い物体検出を行った。以下に本手法の新規性をまとめる。

- Web 教師学習による食事の検出を行った。
- Foodness を認識し食事の候補領域を提案する手法を考案した。

2. 関 連 研 究

本研究では、食事の領域のプロポーザルによる食事の検出を行う。本研究は、物体検出、領域分割などの研究と強く関連している。

2.1 食 事 認 識

食事画像の認識はカロリー推定や食生活の改善に役立つものとして、画像認識におけるアプリケーションの一つとして研究され、これまで多くの [3]〜[9] 研究が発表されてきた。しかしながら、多くの研究は画像に一枚しか食事が含まれていない場合を想定している。複数の食事を認識し、位置の推定を行うことで、より詳細なカロリー推定への応用を行うことが可能であり、これまでいくつかの研究が発表されている [6], [10]〜[12]。

松田ら [6] は複数の食事を検出するために、Deformable Part Model (DPM) [13]、円検出、JSEG による領域分割 [14] を用いた。He ら [12] は Local Variation [15] を食事の領域分割に活用しカロリー推定を行った。モバイル食事認識アプリとしては河野ら [10], [11] の研究があり、ユーザーのインタラクティブな操作から GrabCut により食事の領域分割を行った。近年の研究としては、Myers ら [16] による、”im2calorie”と呼ばれるモバイルフレームワークがある。Myers らはピクセルの深度情報を Deep learning により推定し、カロリーを推定した。

2.2 完全教師あり領域分割

CNN を用いた領域分割の初期における研究としては、[17]、[18], [19] がある。[17]、[18], [19] は CNN の高いクラス分類精度を利用し、プロポーザル [20]、[21] によって得られた多数の領域候補から、認識結果のよいものを選ぶことで、高い領域分割精度を達成した。しかし、プロポーザルを用いる領域分割手法は、CNN を使う以前の領域分割の精度を大幅に上回る精度を達成したものの、全ての領域候補に対して forward の計算を要するために、かなりの時間がかかるという欠点があった。

[22] は物体検出において、CNN の hidden layer が位置情報を反映した特徴 [17] であり、位置の推定に有効であることを示した。hidden layer を用いた位置推定は、一度の Forward の計算のみで、[17] に劣らない高い精度を達成した。[22] は領域分割の研究にも影響を与えている。特に、[23], [24] は hidden layer の位置情報は Upsampling することで、任意のサイズの位置推定結果として扱うことが可能であることを示し、[22] の手法を拡張し、領域分割の研究に取り入れている。

[25] によって提案された、Fully convolutional network は、領域分割の精度向上において大きな役割を果たしている。この手法は、CNN の高層における Inner product 層を Convolution 層の一種として扱うことで、任意のサイズの入力を可能にした。入力画像のサイズを大きくすると、hidden layer の位置情報はきめ細かくなり、CNN の最終的な出力も粗い位置情報を得る。

また、FCN により生成された特徴は、領域の表現に十分な位置情報を持っており、この出力と領域ラベルの誤差を Stchastic gradient discent によって最適化することで、領域の推定に最適化したネットワークが学習であることが示されている [23], [24]、[26]。これにより、一度の Forward プロセスのみで領域分割が可能となり、高速で領域分割が可能となるだけでなく、既存手法を大きく上回る精度を達成した。

2.3 弱教師あり領域分割

弱教師あり領域分割における伝統的な手法の多く [27]、[28]、[29] は、グラフィカルモデルをベースとしており、各ノードの結びつきに low level feature を用いた。

一方で、Zhan et al. [30] は、従来におけるグラフィカルモデルベースの手法に対して、各ノードの結びつきの表現に CNN 特徴量を用いることで、高い精度の弱教師あり領域分割を実現した。また、ソーシャルイメージのノイジーなタグからでも学習が可能であり、CNN がノイズに頑強であることを示した。[31]、[32] は、完全教師あり学習で有効であった手法を弱教師あり学習に適応させた。[31] は完全教師ありの領域分割で高い精度を達成した [23] のネットワークを弱教師あり学習に適応させた。[31] はグローバルプーリングにより、行列の出力をベクトルに変換し、通常のラベルとの誤差の最適化に落とし込んでも大まかな物体の位置の学習が可能であることを示した。[32] は、Overfeat [25] のネットワークに複数の Convolution 層を追加し、粗い位置情報を保持した Features map を出力する Segmentation net を構築した。[32] は [31] と同様に位置情報を持つ出力に処理を加えることで、ラベルとの誤差の最適化を行ったが、後処理としてプロポーザル [21] による Smoothing prior を用いることで、領域分割結果の精度を向上させた。EM-adopt [33]、CCNN [34] は、プーリング、アグリゲーションは使わずに、位置情報をもつネットワークの出力と Pseudo positive 領域ラベルの誤差の最適化を行い、弱教師ありによる領域分割のためのネットワークを構築している。EM-adopt [33]、CCNN [34] はともにこの Pseudo positive 領域ラベルを、ネットワークの出力と弱教師ありラベルから生成した。特に、ラベルの一貫性と、各領域ラベルの領域の大きさの誤差が極端にならないように制限をかけることで、弱教師から擬似の領域ラベルを生成している。[34] はさらに物体の領域のサイズにタグをつけ、この Constrain を用いることでさらに高い領域分割結果が得られることを示した。[35] は Covolutional neural network における学習アルゴリズムに着目し、信号を Back propagation により各階層を通して画像レベルにまで伝搬させると、その伝搬値は大まかな物体の位置を反映していることを示した。また、この位置推定結果を grab cut [36] の種として扱い、弱教師あり領域分割を行った。[35] の手法は弱教師あり領域分割に応用が可能であると主張しているが、物体の位置推定結果がカテゴリスペシフィックでないために精度が。下田ら [2] はこれを各カテゴリについての勾配の差分をとることで解決し、Backward による弱教師あり領域分割で高い精度を達成した。

これらの弱教師あり領域分割は精度を向上させているが、実験のほとんどは PASCAL VOC 2012 データセットのみで行われており、学習画像のドメインはテスト画像と同様であり、マルチラベル画像を含んでいる。本研究では、Web 教師学習に着

目し、シングルラベルのみで学習を行い、テスト時には複数品目の食事を検出する。特に、Fully Convolutional Network をプロポーザルに活用し、ロバストで高速な食事の検出を行う。

3. 手　　法

本研究では、Web 教師学習により、Foodness を検出し食事領域候補を生成する。本手法は Simonyan ら [35] の手法の拡張である DCSM [2] を基にプロポーザルを行い、それらを認識することで物体検出を行う。そのために、プロポーザルのためのネットワークと認識のためのネットワーク 2 つのネットワークを用いて物体検出を行う。図 1 に本手法のネットワークの概要を示す。

3.1 DCSM

本セクションでは、まず本手法の基になっている手法である DCSM について言及をする。Simonayn et al. [35] により提唱された画像レベル I_0 における逆伝搬値は疎な位置推定結果となる傾向がある。そこで、下田ら [2] は Upsampling を用いて中間層の特徴量を補間しこれらを統合した。カテゴリ c における信号 S について、階層 L_i の逆伝搬値 v_i は以下の式で表すことができる。

$$v_i = \left.\frac{\partial S_c}{\partial L}\right|_{L_i} \tag{1}$$

v_i を Upsampling することで、画像と同じサイズ w_i とする。$M^c_{i,x,y}$ を以下の式から計算した。

$$M^c_{i,x,y} = \max |w_{i,h_i(x,y,k)}| \tag{2}$$

$h_i(x,y,k)$ は w_i の要素のインデックスであり、 k はカーネルを表している。ただし、$M^c_{i,x,y}$ はカテゴリごとの応答が小さいために $M^c_{i,x,y} \simeq M_{i,x,y}$ とした。この $M_{i,x,y}$ を統合することで、$g_{x,y}$ を得る。L は統合するためのレイヤー、α はスカラー値である。

$$g_{x,y} = \frac{1}{L}\sum \tanh(\alpha \cdot M_{i,x,y}) \tag{3}$$

また、下田らは誤差を伝搬させる際に、単純な Backpropagation ではなく、Guided backpropagation [37] を採用した。

Simonyan et al. [35] の手法により得られる逆伝搬値は異なるカテゴリによる応答の差が小さい。これは、教師信号のクラスと CNN が学習している他のクラスで、認識に寄与している領域に共通部分があるためであると考えられる。そこで、下田らはこの問題を解決するために、逆伝搬値の差をとることで、クラスごとに共通している領域を排除した。式 2 から、それぞれのクラス c における差分によるマップ $\tilde{M}^c_i$ を得る。

$$\tilde{M}^c_{i,x,y} = \sum_{c' \in D} \max\left(M^c_{i,x,y} - M^{c'}_{i,x,y}, 0\right)[c \neq c'], \tag{4}$$

ただし、D は選択したクラス集合である。また、本セクションにおいては、複数のサイズの入力画像により得られる逆伝搬値を統合し、これにより位置推定結果の改善が可能であることを示した。最終的なクラスマップ $\hat{M}^c$ を以下の式により得る。

$$\hat{M}^c_{x,y} = \frac{1}{|T||L|}\sum_{j\in T}\sum_{i\in L} \tanh(\alpha \tilde{M}^c_{j,i,x,y}), \tag{5}$$

L は統合するためのレイヤー、T はスケールについての集合、α は固定値である。ただし、$\tilde{M}_{j,i}$ は 式 5 の計算の前に画像サイズにリサイズし、正規化を行っている。また、セクション**??**と同じく Guided Backpropagation による伝搬を行った。

3.2 Foodness のプロポーザル

本研究では、単品の食事画像で学習を行い、複数品の食事画像の位置を推定する。一般にこのように学習時とテスト時にドメインが変わると精度が低下し、クロスドメイン問題やドメインアダプテーション問題として知られている。この問題は、DCSM [2] においても同様に現れ、著しい精度低下がみられた。図 2 に本研究におけるクロスドメイン問題のイメージ図を示した。

本研究では、シンプルにプロポーザルを採用することでこのドメインの問題を避けた。プロポーザルは認識対象の領域を含む、領域候補を生成する。対象の食事領域候補の認識は単純なテスト画像の認識と比較して学習画像の状態に近いため、容易になるはずである。

RCNN [38] や SDS [18] は DCNN と領域候補を活用した物体検出、領域分割手法である。RCNN や SDS は Selective Search [20] あるいは MCG [21] といったプロポーザルを用いている。これらの伝統的なプロポーザルの手法は約 2000 の膨大な領域候補を Low lebel feature から生成する。このような膨大な量の領域候補は再現率を上昇させるものの、計算コストが著しく上昇する。そこで、本研究では、約 2000 の領域候補は食事の検出において不要な領域が多く含まれていると考え、新しい食事のためのプロポーザル手法を考案する。

本研究においては、弱教師あり領域分割手法である DCSM を Foodness の検出に応用した。DCSM はドメインの変化に対して脆く、Web 教師学習による物体検出には有効ではなかった。しかしながら、これらの領域の推定結果は正しく位置の推定ができていないものの、食事の領域に反応する。また、推定された領域のいくつかは、画像に含まれていない食事の認識結果であっても、画像のいずれかの食事の領域に反応していた。これらの結果から、異なるドメインにおける画像を学習する際に、正しい食事のクラスの学習は行えなかったものの、食事の概念について学習ができていることが推測できる。食事の認識は各食事のクラスの持つ共通している部分が含まれることから、しばしば詳細画像分類の一つとして扱われる。本研究では、この現象の要因は食事のクラスの皿などの共通している領域があることによると考え、これらの DCSM により粗く推定された領域を Foodness として食事領域候補として扱った。

実際に DCSM を Foodness のプロポーザルとして扱うために、本研究では式**??**における *candidate* を増加させた。ただし、計算コストを考慮し、複数の入力画像の結果を統合することはしなかった。認識候補のクラスの数 P のぶんだけ、領域候補を得ることができる。

$$\mathrm{P^c} = \frac{1}{|\mathrm{L}|}\sum_{\mathrm{i}\in \mathrm{L}} \tanh(\alpha \tilde{\mathrm{M}}^{\mathrm{c}}_{\mathrm{i}}), \tag{6}$$

$\mathrm{P^c}$ は サリエンシーマップのような強弱のある確率分布になっている。この分布は複数の食事領域に反応する場合があるために、$\mathrm{P^c}$ を閾値により分割する。$k \in \{1, 2, \cdots, K\}$ は領域の要素、K は領域の数である。小領域はこの際に考慮せずに切り捨てた。

$$\hat{\mathrm{P}}^c_k \in \mathrm{P^c} \tag{7}$$

最終的に、これらの領域 $\hat{\mathrm{P}}^c_k$ を Foodness のプロポーザルとして扱った。本研究では、DCSM における認識対象のクラスを増加させることで、そのクラスの数だけの領域の推定結果を得

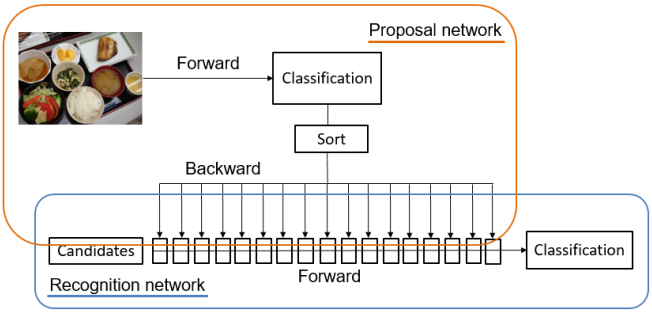

図 1　Processing flow of the our method.

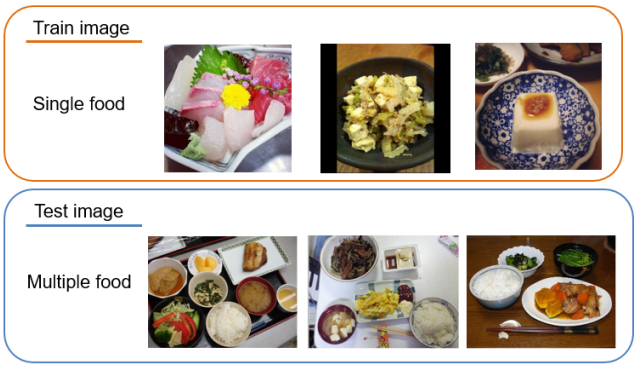

図 2　Example of our cross domain situation.

る。また、認識対象のクラス数は、認識対象のデータセットのクラスの数まで増加させることである。この認識対象のクラス数については、Section 4. 2 で議論する。

4.　CNN training

本研究においては、プロポーザルを行うネットワークとプロポーザルを認識するネットワークの二つを用いる。両方のネットワークにおいて、VGG16 モデルを用い、食事画像によりファインチューニングを行った。ただし、プロポーザルのネットワークにおいては FCN とし、認識においては通常のネットワークを用いた。本セクションでは、この二つのネットワークの詳細について述べる。

4. 1　Proposal network

本研究では、画像レベルのアノテーションのみで学習を行う。近年、Fully Convolutional Networks (FCN) が物体検出、領域分割において広く用いられている。FCN は全結合層を畳み込みの特別な場合と考えることにより、任意の入力画像を入力として扱うことが可能になる。ただし、画像ラベルのみで学習するために、グローバルプーリングを取り入れた。グローバルプーリングを用いることで出力をベクトルとして扱うことが可能となり、通常の画像レベルのロスで学習ができる。特に、VGG16 を FCN として、食事画像 100 種類の UECFOOD100 データセットで学習を行った。

4. 2　Recognition Network

認識のためのネットワークにおいては、通常のロス関数で学習を行い、最後の出力の食事カテゴリ数のみを変える。ただし、プロポーザルのネットワークとは異なり、カテゴリを追加する。このネットワークの目的は、領域候補をより正確に認識することである。領域候補の認識は食事以外の画像と小領域の食事画像を多分に含んでいるという点において、学習時とは条件が異なる。RCNN と SDS においては対象が一般物体であるので、対象のクラス以外の領域候補は背景としてのみ考えた。しかしながら、食事認識は一般の認識とは異なる部分がある。食事画像の認識はテクスチャーの認識としての側面があり、DCNN を用いて食事のパッチ画像を認識する際に、食事画像として識別されることがある。例えば、一般物体において、犬の足や肌の一部は信頼度の低い認識結果となる。一方で、食事の認識においては、ご飯のパッチ画像は高い信頼度でご飯として認識されることが多くある。そのために、追加のカテゴリとして、食事画像のパッチカテゴリを用意した。

さらに、本研究では、低解像度画像を追加した。低解像度の画像が、食事のパッチ画像のカテゴリとして認識される場合があった。これは、食事のパッチ画像が低解像度になりがちであるためであることが原因として考えられる。そこで、低解像度画像を学習画像に含めることで、この現象を避ける。

本研究では、学習画像カテゴリを食事 100 種類＋非食事画像クラスを合わせた 101 種類のカテゴリについて、それぞれ食事パッチ画像として、元の学習画像からクロップした画像を用いて、202 種類のカテゴリを作った。具体的には、各学習画像から三枚の画像をランダムなサイズで切り出し、これを食事パッチ画像とした。短辺 50 ピクセルを最小サイズ、長辺 150 ピクセルを最大サイズとした。ただし、元画像のサイズは 256 × 256 のサイズである。つまり、各パッチ画像は元画像に対して、0.2〜0.6 のスケールで切り出す。また、本研究では元画像に対して縮小、拡大を行うことで、低解像度画像を生成した。縮小する際の比率はランダムに決定した。最小の縮小サイズは短辺 10 ピクセル、最大の縮小サイズは元画像と同じ長辺 256 ピクセルとした。最終的に、以上のデータ拡張により、学習画像は約 7 倍の量となる。

5.　実　　験

実験においては UEC-FOOD100 dataset [6] と Web 食事画像を用いた。UEC-FOOD100 dataset [6] は 100 クラスの食事カテゴリからなり各クラスについて 100 枚の画像がある。また、各食事画像はバウンディングボックスによるアノテーションが付与されている。一方で、Web 食事画像は UEC-FOOD100 と同じ食事カテゴリからなり、各クラスについて約 1000 枚の画像がある。ただし、バウンディングボックスによるアノテーションは付与されていない。Web 食事画像の多くは、Twitter ストリームから収集し、不足分を Bing API から収集している。

実験においては、UEC-FOOD100 における複数品食事画像データセットを用いて、食事の検出を行う。なお、全ての食事検出結果の評価は、Pascal VOC の物体検出において採用されている mean average precision(mAP) に基づいている。

5. 1　食事検出の精度評価

本研究においては、二つの学習データセットを用いる。一つは UECFOOD100 と Web 画像両方を含む学習データセット、もう一方は Web 画像のみを含む学習データセットである。まず、三種類の認識のネットワークについての実験の評価を行う。この際に、プロポーザルのネットワークは固定し、同様のものを用いている。

- 食事画像 100 クラス+非食事画像 1 クラス（101 クラス）
- 食事画像 100 クラス+非食事画像 1 クラス+各クラスのパッチ画像クラス（202 クラス）
- 食事画像 100 クラス+非食事画像 1 クラス+各クラスの

パッチ画像クラス+低解像画像を各クラスについて追加（202クラス）
また、これらの三つの認識のためのネットワークについて、二種の学習データセットを試し、計 6 パターンの評価を行った。Table 1 はその実験についての評価結果である。

表 1　Difference in some recognition models and training image condition

method	non-food class	small-patch class	low-resolution images	training with only web images	100class (all)	53class (#item ≧ 10)	11class (#item ≧ 50)
Foodness 1	✓	-	-	-	30.0	29.3	31.9
Foodness 2	✓	✓	-	-	33.7	39.0	33.6
Foodness 3	✓	✓	✓	-	39.5	46.0	38.9
Foodness 4	✓	-	-	✓	33.5	35.1	33.3
Foodness 5	✓	✓	-	✓	32.2	34.8	31.8
Foodness 6	✓	✓	✓	✓	36.4	39.9	36.3

次に、二つの一般的なグローバルプーリング手法、グローバルアベレージプーリング、グローバルマックスプーリングについての評価を行う。Table 2 は二つのプーリング手法の比較である。Zhou ら [39] の研究などの先行研究においては、グローバルアベレージプーリングにおける優位性が主張されているが、Foodness のプロポーザルにおいては、グローバルマックスプーリングにおける精度が上回っている。これは、グローバルマックスプーリングがより小さい物体をとらえることに優位性があり、それが食事の検出に有効であったことが原因として考えられる。

表 2　Comparison of global pooling operation for food-ness.

method	training with only web images	100class (all)	53class (#item ≧ 10)	11class (#item ≧ 50)
Foodness (average pooling)	-	39.5	46.0	38.9
Foodness (average pooling)	✓	36.4	39.9	36.3
Foodness (max pooling)	-	39.9	48.3	37.6
Foodness (max pooling)	✓	38.9	42.5	38.1

プロポーザルネットワークの精度と速度を既存のプロポーザル手法と比較し評価する。比較においてはベースラインとして二つのプロポーザル手法、Selective Search と Multiscale Combinatorial Grouping を用いる。これらは共に約 2000 の領域候補を提案するために、認識に計算コストを要する。本研究の速度麺の優位性を主張するために、認識対象のクラス数を変化させることで、領域候補の数を調整する。認識対象のクラス数を小さくすれば、Backward の計算コストを抑制することができるとともに、領域候補の数も減少するために、認識に必要な計算コストも減少する。

Table 3 は比較結果である。ただし、認識速度はプロポーザルから得られる領域候補数から求めた理論値である。本研究の手法は、40 倍以上少ない領域候補数であるものの、高い mean AP を達成している。さらに、認識対象のクラス数を小さくしても、30 %以上の mean AP を保っている。これは、本手法の Foodness の検出の精度が高いことを示している。

表 3　Comparison with other traditional proposal method.

method	100class (all)	53class (#item ≧ 10)	11class (#item ≧ 50)	proposal speed[s]	Recognition speed for candidates[s]
Selective Search [20]	38.3	39.1	35.7	7.6	35.0
Multiscale Combinatorial Grouping [21]	33.9	43.7	33.4	2.5	35.0
Foodness with 10 candidate classes	33.1	33.0	33.2	0.5	1.1
Foodness with 20 candidate classes	36.5	40.1	37.7	1.0	2.6
Foodness with 30 candidate classes	38.9	42.5	38.1	1.4	3.8

6. おわりに

本手法においては、弱教師あり学習による、CNN に基づいた Foodness についてのプロポーザル手法を提案した。特に、既存のプロポーザルと Fully Convolutional Network のアプローチを組み合わせることで、高速な食事領域の検出を行った。今後の課題として、食事の検出のみでなく、食事のインスタンスの領域分割結果を考慮し、更なる精度向上とより高度な認識を行いたい。

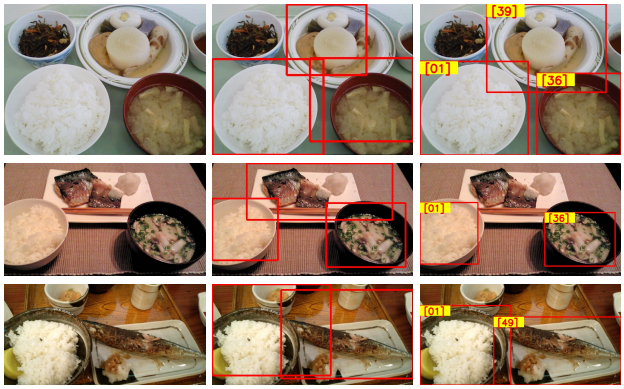

図 4　Examples of results.

文　献

[1] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proc. of IEEE International Conference on Computer Vision*, 2015.

[2] S. Wataru and Y. Keiji. Distinct class saliency maps for weakly supervised semantic segmentation. In *Proc. of European Conference on Computer Vision*, 2016.

[3] L. Bossard, M. Guillaumin, and L. V. Gool. Food-101 - mining discriminative components with random forests. In *Proc. of European Conference on Computer Vision*, 2014.

[4] H. Kagaya, K. Aizawa, and M. Ogawa. Food detection and recognition using convolutional neural network. In *Proc. of ACM International Conference Multimedia*, pp. 1085–1088, 2014.

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

[6] Y. Matsuda, H. Hoashi, and K. Yanai. Recognition of multiple-food images by detecting candidate regions. In *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 1554–1564, 2012.

[7] M. Chen, Y. Yang, C. Ho, S. Wang, S. Liu, E. Chang, C. Yeh, and M. Ouhyoung. Automatic chinese food identification and quantity estimation. In *SIGGRAPH Asia*, 2012.

[8] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp. Combining global and local features for food identification in dietary assessment. In *Proc. of IEEE International Conference on Image Processing*, 2011.

[9] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar. Food recognition using statistics of pairwise local features. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2010.

[10] Chamin Morikawa, Haruki Sugiyama, and Kiyoharu Aizawa. Food region segmentation in meal images using touch points. In *Proc. of ACM MM WS on Multimedia for Cooking and Eating Activities (CEA)*, pp. 7–12, 2012.

[11] Y. Kawano and K. Yanai. Real-time mobile food recognition system. In *Proc. of IEEE CVPR International Workshop on Mobile Vision (IWMV)*, 2013.

[12] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp. Food image analysis: Segmentation, identification and weight estimation. In *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2013.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively

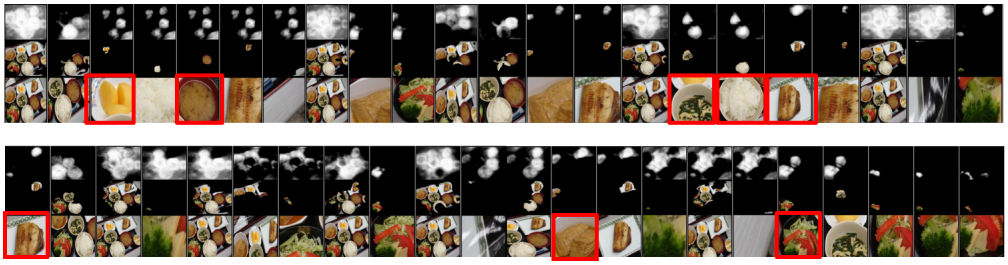

図 3 Proposal results.. First row show salieny obtained from DCRM. Second row shows regions obtained from saliecny maps. Third row shows bounding boxes we recognize. Red rectangle means it seems to be a good candidate.

trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 9, pp. 1627–1645, 2010.

[14] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 8, pp. 800–810, 2001.

[15] P. F. Felzenszwalb and D. P. Huttenlocher. Image segmentation using local variation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 98–104, 1998.

[16] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy. Im2calories: Towards an automated mobile vision food diary. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[17] R. Girshick. Fast r-cnn. In *Proc. of IEEE International Conference on Computer Vision*, 2015.

[18] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *Proc. of European Conference on Computer Vision*, 2014.

[19] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic iamge segmentation with deep convolutional nets and fully connected crfs. In *Proc. of International Conference on Learning Representations*, 2015.

[20] 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*, Vol. 104, No. 2, pp. 154–171, 2013.

[21] A. Pablo, Jonathan T. Jordi, P., M. Ferran, and M. Jitendra. Multiscale combinatorial grouping. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proc. of European Conference on Computer Vision*, pp. 346–361, 2014.

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[24] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[25] P. Sermanet, D. Eigen, X. Zhang, M.l Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. of International Conference on Learning Representations*, 2014.

[26] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and Yuille A. L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *Proc. of International Conference on Learning Representations*, 2015.

[27] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2010.

[28] A. Vezhnevets and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2012.

[29] L. Zhang, Y. Gao, Y. Xia, K. Lu, J. Shen, and R. Ji. Representative discovery of structure cues for weakly-supervised image segmentation. Vol. 16, No. 2, pp. 470–479, 2014.

[30] W. Zhang, S. Zeng, D. Wang, and X. Xue. Weakly supervised semantic segmentation for social images. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[31] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *Proc. of International Conference on Learning Representations*, 2015.

[32] P. Pedro and C. Ronan. From image-level to pixel-level labeling with convolutional networks. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[33] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *Proc. of IEEE International Conference on Computer Vision*, 2015.

[34] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proc. of IEEE International Conference on Computer Vision*, 2015.

[35] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proc. of International Conference on Learning Represenation Workshop Track*, 2014.

[36] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimality boundary & region segmentation of objects in n-d images. In *Proc. of IEEE International Conference on Computer Vision*, 2001.

[37] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *Proc. of International Conference on Learning Representations*, 2015.

[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[39] B. Zhou, A. Khosla, A. Lapedriza, and A. Oliva, A. Torralba. Learning deep features for discriminative localization. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2016.