画素単位アノテーション付きの食事画像データセットの構築と 認識・生成への応用

岡本 開夢[1,a)]　柳井 啓司[1,b)]

概要

現在、領域分割の画像データセットは多数公開されているが、そのなかで食事におけるカテゴリは少数に限られている。また、食事画像データセットとして画像毎に単独の食事名がアノテーションされたものや、UECFood [11] の様なバウンディングボックス付きのものもあるが、大規模な食事領域推定データセットは存在していない。一方、食事画像におけるタスクとしてカロリー量推定などが行われているが、これらのタスクは画像内における食事領域の面積に対して関連しており、食事領域の領域分割が不可欠である。そこで、本研究では食事領域用データセットである UECFoodPix [4] を発展させることで、高品質な領域分割様データセットとして、UECFoodPix Complete を作成した。また、このデータセットの活用例として Deeplab V3+ [2] による領域分割と SPADE [14] による画像生成を行い有用性を示す。

1. はじめに

現在、深層学習の発達により画像認識の精度が飛躍的に向上した他、画像生成や領域分割といったタスクにおいても優れた成果を残している。深層学習による教師あり領域分割においては、学習画像に画素ごとにアノテーションされたマスク画像データセットが必要となる。領域分割に必要な大規模データセットとして PASCAL VOC 2012 [5] や MS COCO [10] が挙げられるが、これらに含まれる画像は動物や乗り物のラベル付けされたものが大半を占めており、食事については少数のカテゴリに限られている。また単体の食事データセットにおいては単独の食事名がアノテーションされたものが大部分で、UECFood-100 [11] といった画像内の複数の食事に対してバウンディングボックスがアノテーションされているものは少数である。画素ごとにアノテーションされたデータセットとして、會下 [4] らによって、UECFood-100 をもとに作成された UECFoodPix が存在する。しかし、これはバウンディングボックスに基づいて Grab-Cut[15] によって半自動的にアノテーションされたため、誤りが含まれているという問題点がある。

一方で画像生成分野においては、Generative Adversarial Networks(GAN) [6] の登場により実際の画像に近しい画像が生成することが可能となった。GAN はノイズから画像を生成するが、最近ではマスク画像を用いた [14] ものも登場しており、よりいっそう領域分割データセットの価値が高くなっている。

こうした状況を踏まえて、本研究では會下 [4] らによって作成された一部不完全なアノテーションが含まれる UECFoodPix を人手で修正することによって、完全な領域アノテーション付き食事画像データセット UECFoodPix Complete を作成する。さらに、作成したデータセットを用いて領域推定と画像生成を行い、その活用例を示す。本研究で作成した画像データセットは近日中に公開予定である。

2. 関連研究

領域分割に用いられるデータセットとして PASCAL VOC 2012 [5] や MS COCO [10] がベンチマークとして頻繁に用いられている。PASCALVOC は 2005 年から 2012 年に行われたコンペティションに用いられたデータセットで 2012 年版では飛行機や自転車を含む 22 クラス 9,993 枚を含む。MS COCO は Microsoft 社が提供するデータセットで 80 クラス 33 万枚の画像を含む。しかし食事クラスに着目した際に MS COCO に含まれるカテゴリとして、ピザやホットドックといったわずか 10 クラスしか含まれておらず、食事領域分割モデルの学習画像としては限られたクラスにしか対応出来ていない。

一方、食事領域を抽出し応用した研究として、岡元ら [13] の研究が挙げられる。これは、大きさが既知の基準物体とともに食品を上から撮影することでカロリー量を推定するシステムを考案している。GrabCut [15] を用いて画像内の基準物体と食事の領域を抽出することで実面積を算出し、事前に作成したカロリーと面積の回帰式をもとにカロリー量の推定を行っている。GrabCut は前景と背景を分割する手法で領域を分割することができ CNN が登場する以前の主流な領域分割手法である。しかし、この研究は撮影角度

[1] 電気通信大学大学院情報理工学研究科
[a)] okamoto-ka@mm.inf.uec.ac.jp
[b)] yanai@mm.inf.uec.ac.jp

も上からのみと限定的なものであった。

食事領域分割用データセットと、食事領域を抽出した研究として、會下らの [4] 研究が挙げられる。これは、元の UECFood データセットのバウンディングボックス内の食事領域に対して GrabCut [15] を用いることで画素単位に半自動にアノテーションし、このデータセットを用いて米飯のカロリー量を推定した研究である。しかし、この研究によって作成された UECFoodPix は半自動で領域アノテーションが生成されているため、ノイズが含まれており、領域分割モデルの学習に利用した場合，十分な精度を出すことができていない。

また、画像生成の分野においては GAN の発展により劇的な進歩を遂げている。GAN は生成器と判別器を敵対的に学習することで実際の画像に近しい画像を生成できるが、Conditional GAN [12] のようにノイズに条件を与えることでより高品質な画像生成を行う手法も提案されている。その他にもマスク画像から画像を生成する研究も行われており、Park [14] らは、conditional normalization 手法、Spatially-Adaptive Denormalization(SPADE) を提案している。SPADE はセグメンテーションマスク画像からの情報をネットワークの noramlization に与えることで、よりリアルな画像を生成することが可能となっている。食事画像の生成を行なった研究として、Cho [3] の提案した “RamenAsYouLike” が存在する。これは、Deeplab V3+ [2] と pix2pix [8] を組み合わせた研究である。食事画像から推定した領域をもとにスケッチ画像を作成しこのスケッチ画像を手動で変更することで自在な “ラーメン” 画像を作成でき食事画像生成をアプリケーションに応用した例である。食事画像生成は、様々なタスクにおいて応用されつつあり今後も発展していくと予想される。本研究では SPADE を用いて画像生成を行う。

3. UECFoodPix Complete

現在、多くの食事画像データセットが公開されており、食事分類タスクにおいては Food-101 [1]、UECFood-100 [11] や UECFood-256 [9] などが標準的なベンチマークとして用いられている。画像内の複数の食事に対してバウンディングボックスを持つものに関しては、UECFood-100/256 など少数に限られている。また、セグメンテーションマスク付きの大規模な食事画像データセットに関しては、會下らの UECFoodPix [4] が存在する。しかし、このデータセットは食事領域の境界においてマスクにばらつきを含む。そこで本研究では、UECFoodPix を手作業で拡張することにより、より高品質な食事領域分割用データセットとして **UECFoodPix Complete** を作成した。

UECFoodPix はバウンディングボックスから GrabCut[15] を用いてマスク画像を生成しているため、図 1 の中の “とんかつ” の画像のように “とんかつ” と “キャベツ” が分離していない。また、“コーンスープ” の画像のように食事領域の境界面に凹凸が生じており、適切なマスク画像となっていない。そこで、手作業によりマスク画像の精査を行なった。境界面に注意しながら、数人によって UECFoodPix のマスク画像をある程度精査したのち、人によるばらつきを防ぐために、筆頭著者自身 1 人のみで 1 万枚のマスク画像の最終確認を行った。作成したデータセット UECFoodPix Complete は，学習用 9000 枚、評価用 1000 枚から成っている．

図 1 マスク画像 (左:実画像, 中:自動マスク画像, 右:手動マスク画像)

4. 領域分割モデルの学習と評価

まず、作成した食事領域分割のデータセット (UECFoodPix Complete) を用いた領域分割を行う。領域分割用のモデルは Deeplab V3+ [2] とし Accuracy と mean Intersection over Union (mIoU) による定量評価を行い、評価用データを用いて分類された食事領域を示す。作成した食事画像データセットのベンチマークテストとして、Chen らによって提案された Deeplab V3+ [2] による食事領域抽出を行った。Deeplab V3+は複数のスケールで格子状に分割しそれぞれの画像に対して畳み込みを行うピラミッド構造とエンコーダーデコーダーモデルを組み合わせたセマンティックセグメンテーションモデルであり、領域分割手法として頻繁に使われる手法である。今回のこの Deeplab V3+の画像特徴量を抽出するモデルとして ResNet-101 [7] を用いる。学習画像を変えた 3 つのモデルを用意しそれぞれ比較を行う。モデル A をもとの UECFoodPix を用いて学習した場合、モデル B を UECFoodPix の学習画像の 9,000 枚のうち 2,000 枚のマスク画像を人手で精査した場

表 1　領域分割の精度

学習モデル	Acc	mIoU
モデル A：すべて自動	0.560	0.416
モデル B：2000 枚手動	0.597	0.436
モデル C：9000 枚手動	0.668	0.555

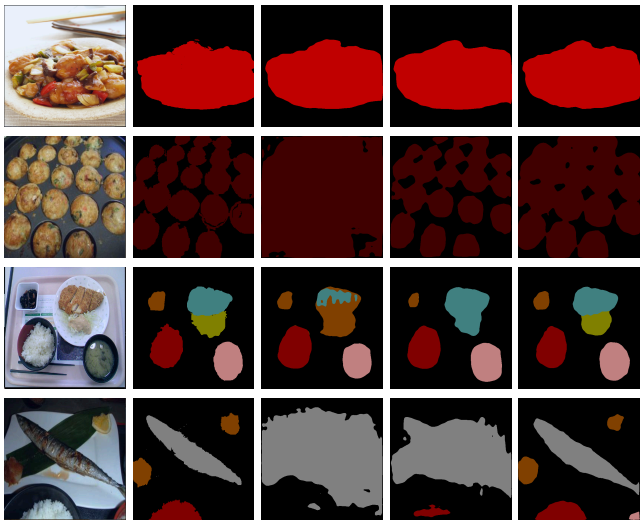

図 2　Deeplab V3 による領域分割の結果 (1 列目：入力画像、2 列目：評価マスク画像、3 列目：モデル A の結果、4 列目：モデル B の結果、5 列目：モデル C の結果)

合、モデル C を今回作成した UECFoodPix Complete を用いた場合とする。評価用の画像は、1,000 枚人手で精査したものをそれぞれのモデルで用いる。

評価には各クラスにおける Accuracy と mIoU を用いており結果は表 1 の通りになった。また、図 2 は評価画像における食事領域抽出の結果の一例であり、1 列目が入力画像、2 列目が正解マスク画像、3 列目がモデル A の結果、4 列目がモデル B の結果、5 列目がモデル C の結果となる。結果として、元のデータセットで学習したモデル A と今回作成した UECFoodPix Complete で学習したモデル C を比べると Accuracy においては約 0.1、mIoU においては約 0.14 の精度向上が見られた。また、領域推定の結果から、単品料理のみの場合 (酢豚) は学習画像に違いにより領域分割結果にさほど変化はないが、複数品目の場合 (定食の画像) には手動で精査した画像を用いる枚数が増えることで、より複雑な形状や境界に対応できる傾向が見られた。これらは、元の UECFood [11] のバウンディングボックスが、境界が曖昧な複数品目に対応しておらず “唐揚げ” と “キャベツ” が同じボックス内に含まれている場合や、“秋刀魚” のように細長い食事がボックス内に含みれきれていない場合があり、これらのボックスに対して、GrabCut [15] のみによるマスク生成だけでは、複数食材や複雑な形をもつ食事の境界を十分に分割できないためと考えられる。

5.　画像生成モデルの学習

次に、作成した食事領域分割用のデータセットの活用として画像生成を行なった。画像生成には SPADE [14] を用いて画像を作成した。SPADE はマスク画像から抽出した、スケーリング項とバイアス項をもとに正規化することで意味情報をよりよく反映できるようにしたもので、マスク画像を用いた画像生成において頻繁に用いられるモデルである。学習に用いた画像は手作業で精査したマスク画像 9000 枚を用いて学習し、残りの 1000 枚を用いて画像生成を行なった。

図 3　SPADE に生成画像 (上：入力したマスク画像、下：SPADE による生成画像)

結果画像は図 3 のようになり、上段が入力したマスク画像、下段が SPADE による生成した結果である。“天ぷら丼” や “ビビンバ” のような単品品目は生成できている他、定食のような複数品目においても現実味のある画像が生成された結果となった。また、学習画像のアノテーションは食事領域のみ設けられており、皿や茶碗などの領域は一律背景としているが、これらに関しても食事にそったものが生成される結果となった。そのため、UECFoodPix Complete は単品品目や定食のような食事同士がはっきりとした境界をもつ食事画像に対して実用的なデータセットであると言える。しかし、UECFoodPix Complete が食事領域を用いたアプリケーションを見据えて作成されているため、対応できない画像も存在する。4 行目の真ん中の “天ぷら” の画像の様に、容器から食事が飛び出しているものに関しては、丼の部分までもが変形してしまっている。これはマスク画像が食事領域にのみ付けられており、容器の領域に対してマスクが設けられていないため、食事領域全体を覆う様に背景が生成されてしまったからである。また、4 行目の “ロールパン” の画像の場合には、マスク画像がインスタンスを考慮していないため、複数の食事が一つの食事として生成されてしまっている。

次に、マスク画像のラベルを変化させながら画像を生成した。変化させるために用いる、クラスラベルとして “牛丼”，“親子丼”，“冷し中華”，“ラーメン” の 4 つを用いた。また、“定食” の画像に関して複数品目の変換を行なった。“ごはん” の領域に関しては、上記のクラスラベルを適用し

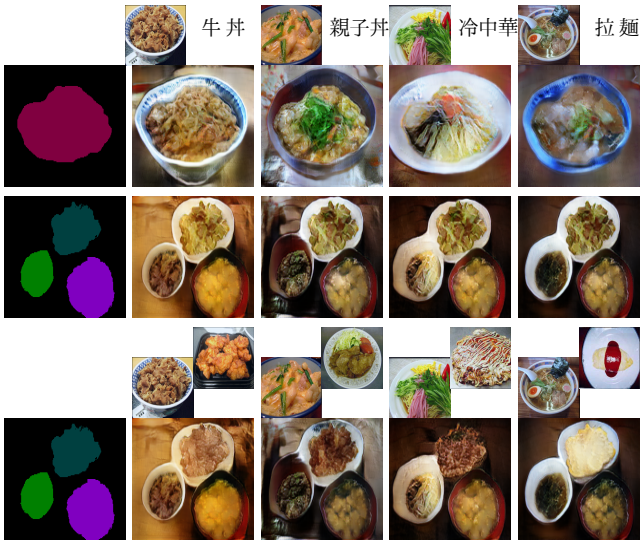

図 4 マスクラベルを変更した画像生成の結果 (1 行目、4 行目:クラスラベル画像、2 行目:単品品目による結果、3 行目：複数品目による結果、4 行目:複数品目における同時変換の結果)

“野菜炒め” の領域に関しては、“からあげ”，‘生姜焼き”，“お好み焼き”，“オムライス” の変換を行なった。結果は図 4 の通りとなった。左側がマスク画像、1 行目がクラスラベル、2 行目以降が生成結果となっている。

図 4 の 2 行目のように単一品目の単調で十分な大きさの食事領域をもつ画像に対しては、ラベルを入れ替えることで別の画像に変換することが可能であることが示された。一方で、3 行目の画像の “ごはん” のような小さい領域に関して “ラーメン” ラベルに変更した際に、圧縮されたような画像が生成された。これは、小さい領域に対して、複雑な特徴をもつ “ラーメン” のようなラベルを割り当てたことによるものであるため、より高解像度なマスク画像を用いることで、大小様々な複数品目を含む食事画像における画像変換アプリケーションにも応用可能となることが期待される。

6. おわりに

本研究では、既存の UECFood-100 [11] に複数品目を考慮したバウンディングボックスを付加し、そのバウンディングボックスに対して GrabCut [15] と手作業による精査を行うことで領域分割用の食事データセットを作成した。また、Deeplab V3+ [2] による領域分割を用いることで UECFoodPix の有用性を示した。画像生成においても、SPADE [14] を用いることで現実性のある画像を生成できることを示した。

このデータセットにより学習させた領域推定モデルを用いることで、食事領域と密接に関係するカロリー量推定タスクにおいて高精度なシステムの開発が見込まれる。今後の発展として、このデータセットに対してタスクの目的に対してさらなるアノテーションを行うことを検討中である。特に食事量やカロリー量といった “量” を考慮したベンチマークとなるデータセットが存在していないため、これらの情報を付加させることでより実用性の高いデータセットに発展させる予定である．

謝辞 アノテーション作業を手伝ってくれた多くの皆さんに感謝致します．本研究は JSPS 科研費 15H05915, 17H01745, 19H04929, 17H06100 の助成を受けたものです．

参考文献

[1] Bossard, L., Guillaumin, M. and Van Gool, L.: Food-101 – Mining Discriminative Components with Random Forests, *in Proc. of ECCV* (2014).

[2] Chen, L., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, *in Proc. of ECCV* (2018).

[3] Cho, J. and Yanai, K.: RamenAsYouLike: Sketch-based food image generation and editing, *in Proc. of ACM Multimedia, Demo Track* (2019).

[4] Ege, T. and Yanai, K.: A New Large-scale Food Image Segmentation Dataset and Its Application to Food Calorie Estimation Based on Grains of Rice, *in Proc. of MADiMa* (2019).

[5] Everingham, M., Eslami, S., Van Gool, L., Williams, C., Winn, J. and Zisserman, A.: The pascal visual object classes challenge: A retrospective, *International Journal of Computer Vision*, Vol. 88, No. 2 (2010).

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *in Proc. of NeurIPS* (2014).

[7] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *in Proc. of CVPR*, pp. 770–778 (2016).

[8] Isola, P., Zhu, T. and Efros, A.: Image-to-image translation with conditional adversarial networks, *in Proc. of CVPR* (2017).

[9] Kawano, Y. and Yanai, K.: Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation, *in Proc. ECCV WS on TASK-CV* (2014).

[10] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.: Microsoft COCO: Common objects in context, *in Proc. of ECCV* (2014).

[11] Matsuda, Y., Hoashi, H. and Yanai, K.: Recognition of Multiple-Food Images by Detecting Candidate Regions, *in Proc. of ICME*, pp. 25–30 (2012).

[12] Mirza, M. and Osindero, S.: Conditional generative adversarial nets, *arXiv preprint arXiv:1411.1784* (2014).

[13] Okamoto, K. and Yanai, K.: An Automatic Calorie Estimation System of Food Images on a Smartphone, *in Proc. of MADiMa* (2016).

[14] Park, T., Liu, M. and Zhu, J.: Semantic image synthesis with spatiallyadaptive normalization, *in Proc. of CVPR* (2019).

[15] Rother, C., Kolmogorov, V. and Blake, A.: GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts, *ACM Trans. Graph.*, Vol. 23, No. 3, pp. 309–314 (2004).