一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

HCGシンポジウム2019
HUMAN COMMUNICATION GROUP SYMPOSIUM 2019
HCG2019-C-4-3

深層学習を用いたコミックからの料理画像検出の一検討

小林　達哉†　　松下　光範†

† 関西大学総合情報学部　〒569-1095 大阪府高槻市霊仙寺町 2-1-1
E-mail: †k252475@kansai-u.ac.jp, ††mat@res.kutc.kansai-u.ac.jp

あらまし　本研究では深層学習を用いて実世界の料理画像を学習させることで、コミック画像中に含まれる料理画像の検出を行う。現状では、コミック画像を学習させることでコミックからの物体検出を行っている。しかし、従来の手法ではコミック中の物体検出のために使用されるデータはコミック画像に限られているため、用意できるデータセットの量に限りがある。そこで本研究では、自然画像を用いることでマンガ内の料理画像検出を目指す。その端緒として (1) 無加工 (カラー)、(2) グレイスケール化、(3) コミック風画像化、の 3 種類の料理画像を用いて各々 SSD300 による物体検出を行い、それらを用いてコミック内の料理画像抽出を行った際の精度を比較した。
キーワード　マンガ, 画像処理, 物体検知, 料理

Detecting Dish Images from Comics using Deep Learning

Tatsuya KOBAYASHI† and Mitsunori MATSUSHITA†

† Faculty of Informatics, Kansai University　2-1-1 Ryozenji-cho, Takatsuki-shi, Osaka, 569-1095 Japan
E-mail: †k252475@kansai-u.ac.jp, ††mat@res.kutc.kansai-u.ac.jp

Abstract　In recent years, it has been studied that various methods for detecting objects depicted in a comic by learning comic images. With conventional methods, however, the learning data used to detect objects in the comics is line drawings, and the amount of data sets for that purpose is limited. To solve this problem, this study examines how to detect dishes depicted in comics by learning actual cooking photos with a deep learning method. Our proposed method uses three types of cooking images such as (1) unprocessed (color), (2) gray scale, and (3) comic-style image. In our method, three types of cooking images were prepared: (1) unprocessed (color), (2) gray-scale, and (3) comic-like imaging. The objects in comics were detected by SSD300. This paper compared the accuracy of cooking image extraction in comics using this method.
Key words　comics, image processing, detecting of object, dish images

1. はじめに

コミックは日本の代表的なポップカルチャーコンテンツであり、年間 10,000 タイトル以上もの新刊タイトルが出版され流通している。それらのコミックで取り扱われるテーマは多岐にわたり、エンタテインメントとしての利用にとどまらず、未知の分野に対する興味喚起のトリガーとなったり、学びのための手段として用いられたりしている。こうした利用を促進するためには、コミックで描かれている内容と現実世界の情報とをリンクさせ、シームレスに辿ることができる環境が求められる。こうした観点の下、本研究ではコミックに描かれている「料理」に着目し、Web 上で閲覧可能な料理情報とコミックに描かれる料理画像とを紐付け、コミックの中の料理から実際の料理の作り方やそれを提供する店舗などの情報へのアクセスを可能にすることを目指す。

近年は、料理を作る過程や登場人物が料理を食べるシーンなどが描かれるコミック作品（以下、料理マンガと呼ぶ）が多く出版されている。こうした料理マンガに触発されて描かれている料理に興味を持ち、実際にその料理を作成する読者も増えている。料理を作成する工程が物語中に記載されていたり、レシピ本が別途出版されていたりする作品もあり、その料理を容易に作成できるようになっているが、こうした作品は一部にとどまっている。また、料理を作る工程やレシピが描かれているコミックであっても、一部の手順が省略されているものもある。こうしたケースでは、読者がそのような料理を再現することは困難である。

そこで本研究では、コミック中に現れる料理レシピを実際の料理と紐付けてユーザに提示することを目的とする。その端緒として、写真などの自然画像を用いてマンガ内に出現する料理の位置情報と料理名の判定を試みる。

2. 先行研究

画像内の位置情報と物体を推測する研究としては物体検出という方法が多く用いられている。

柳澤らはマンガの物体検出について、PASCAL VOC detection task の評価手法(注1) に基づいた平均適合率を用いて物体の検出率を評価している。PASCAL VOC detection task は与えられた画像 20 個のクラスから、画像内の物体がどのクラスに属し、画像のどの位置に存在するかを予測するタスクである。このタスクで用いられた評価の元、柳澤らは物体検出のモデルである Faster R-CNN のモデルを用いてキャラクターの顔、フキダシ、活字のテキスト、コマ枠のそれぞれ 4 種類の物体を検出させ、モデルの評価をした [7]。しかし本研究における評価手法はこの平均適合率を利用せず、IoU (Intersection over Union) 評価(注2)を用いる。IoU 評価とは定めた位置情報とモデルにより求めた位置情報とで重なった部分の割合を計算する評価指標である。本研究では、人間が目で見る料理の位置とモデルが認識する料理の位置の差分を重視し、コンピュータが料理であることを判断できるか確認する必要があるため、この IoU 評価を採用した。

小川らは、マンガ内の物体間には重なりが大きくモデルの性能が下がることを問題とし、この問題を解決するための新しい物体検出モデルである SSD300-fork を提案した [5]。このモデルを用いると、従来のモデルより平均適合率が向上することが確認された。しかし、このモデルはマンガ画像を学習させるためのモデルであり、自然画像を学習することを目的としていない。そのため、本研究では、小川らの実験に適用された物体検出モデルの SSD300-fork を除くモデルのうち、最も平均適合率が高かった SSD300 モデルを用いることとした。

3. 手　　法

ここでは本研究で提案する手法と実験概要について説明する。

3.1 SSD300

SSD300 は物体検出をするためのモデルである [4]。VGG16 ネットワークをボトルネック層とし、追加ネットワーク層を加えたネットワークで構成され、後段になるほど畳み込み層が小さくなる。(図 1)。このネットワークの特性から、前段の畳み込み層は大きくなることで小さな物体を検出することができる。同様に、後段の畳み込み層は小さくなることで大きな物体を検出することができる。この検出は、入力画像から大きさの異なる畳み込み層を抽出し、縦横比が異なるボックスを畳み込み層に適用することで可能となる (図 2)。なお、抽出される畳み込み層は図 1 の通りである。

このボックスは物体からの位置のズレとボックス内の物体がどのカテゴリに属するかを表す確信度を求める。SSD300 はこのボックスを計 8,732 個作成し、画像の大きさを 300×300 に設定する。SSD300 のネットワークはボトルネック層である VGG16 から追加ネットワークの順に順伝播が計算され、ボックスの位置のズレとクラスが推論されていく。損失関数は物体の位置のズレとクラス分類の 2 点で計算され、勾配がネットワーク全体に伝播される。以上より、End-to-End の計算を実現可能にし、複数の物体を検出しクラスを推論させることができる (図 2 参照)。

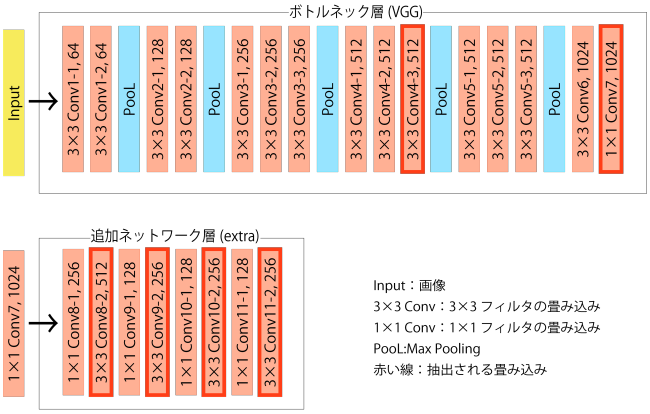
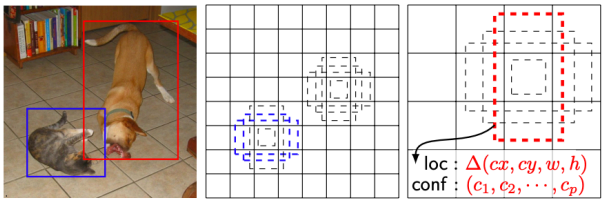


図 1　SSD300 の構成基礎。これらの畳み込み層の Conv4-3, Conv7, Conv8-2, Conv9-2, Conv10-2, Conv11-2 が抽出され、縦横比が異なるボックスが作成される。



(a) Image with GT boxes　(b) 8 × 8 feature map　(c) 4 × 4 feature map

図 2　(b) の青い点線が (a) の青い実線の位置に一致するように重みが更新され、カテゴリが猫であると予測されるようにする。抽出される畳み込み層は (b) と (c) のようにスケールが違うため、ボックスは相対的に (b) では小さくなり、(c) では大きくなるため様々な大きさの物体を検出できる。(文献 [4] より図引用)

3.2 学習用データセットの前処理

柳澤らはマンガ画像を用いてキャラクターなどの位置情報を推定している [7] が、実験に使用しているデータセットが少ないという問題がある。柳澤らの手法では Manga109 [2] をデータセットとして利用している。このデータセットは 1970 年代からのマンガ作品をはじめ、日本のプロの漫画家によって描かれた 109 冊のマンガで構成されている画像データセットである。公開されているデータセットにはアノテーションがあり、キャラクターの顔、キャラクターの全身、活字のテキスト、コマ枠のそれぞれの位置情報が付与されている。しかし、この方法においてコミック中の物体検出のために使用されるデータはコミック画像に限られているため、用意できるデータセットの量に限りがある。例えば、料理を検出するために必要な教師データは Manga109 のみでは不十分である。そこで本研究では、コミック画像のデータを用いずに、自然画像を用いることで料理マン

(注1)：http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html
(注2)：Intersection over Union (IoU) for object detection - pyimagesearch (2016)

図 3　無加工 (カラー)

図 4　グレイスケール

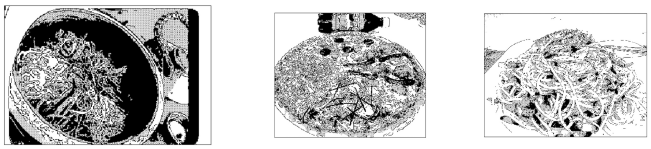


図 5　マンガ風加工

ガの料理画像検出を目指す。自然画像とは実世界における画像のことを表すとし、本研究において以降そのように扱う。

学習用に用いる自然画像の料理画像には、UECFOOD-100 データセットを用いる [1]。このデータセットは 100 種類の料理画像が含まれ、単品料理の画像に加え複数料理の画像が存在する。単品料理の画像とは一つの画像に対して一種類の料理しか含まれておらず、複数料理の画像は一つの画像に対して複数の料理が存在する画像のことである。UEC FOOD-100 の料理画像の合計は単品料理と複数の料理の画像を合わせ 12740 枚の画像データセットで構築されており、それぞれの料理にはバウンディングボックスが付与されている。この料理画像は SSD300 のモデルの学習用画像として扱う。しかし、マンガ画像は本来白黒でありカラーではないため、自然画像をカラーで学習させても効果がない可能性がある。そのため自然画像を学習させるその端緒として (1) 無加工、(2) グレイスケール化、(3) コミック風画像化、の 3 種類の異なる前処理を施し各々 SSD300 による物体検出を行い、コミック内の料理画像の検出を行った際の精度を比較する。(1) の無加工は自然画像に対して何も加工せず学習させる。(2) のグレイスケール化は、画像を 0〜255 の値に正規化し学習させる。(3) のコミック風画像化はグレースケール変換、輪郭検出、多値化、マスク処理を順に行い学習させる。なお、輪郭抽出は Canny アルゴリズムを用いる。それぞれの前処理をした料理画像の一例を図 3〜図 5 に示す。

3.3　評価用データセットの構築

評価用のデータセットはマンガ図書館 Z の料理マンガを使用する。マンガ図書館 Z とは漫画家・権利者の方々の許諾や好意により絶版になってしまったコミックや出版社の許諾を得たコミックなどが無料で読むことができる電子書籍サイトである(注3)。本研究では、5 巻以上出版されている料理マンガという条件の下、『ザ・シェフ』、『海猫亭へようこそ』、『天才料理少年味の助』の 3 作品を選択した。これらの料理マンガより料理が描かれているコマの一枠を切り抜き、マンガ料理画像を作成する。切り抜かれる画像は UEC FOOD-100 のカテゴリの料理のみとし、マンガ内のテキストやイラストにより UEC FOOD-100 に存在するカテゴリの料理であるかを判断しコマの切り抜きを行った (図 6)。それぞれのコマはバウンディングボックスが付与されていないため、全ての画像に対して手動でバウンディングボックスを付与する。以上の手続きに従って料理画像を作成し、合計画像枚数 154 枚のバウンディングボックス付き画像データセットを構築した。

(注3)：https://www.mangaz.com

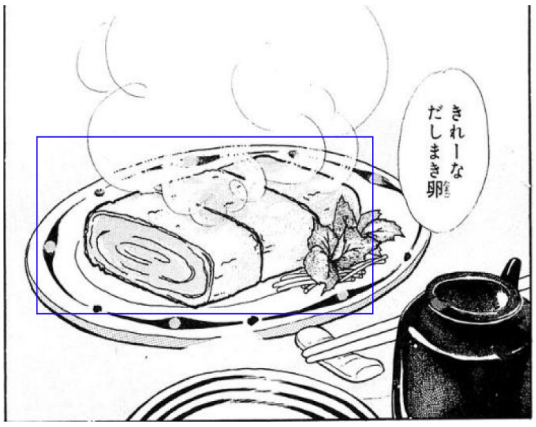


図 6　切り取られた作品に対してバウンディングボックスを付与させた例(注4)

評価指標として、物体検出には PASCAL VOC detection task の評価手法に基づく平均適合率が標準的に用いられるが、本研究ではラベルの正解を重視せずマンガ料理画像が検出できるかに焦点を当てる。そのため、人間が料理であると認識した位置と SSD300 モデルにて推論された位置との誤差を測る必要があるため、本研究では IoU 評価を用いて評価することとした。加えて、切り取られたマンガのコマに対して、料理が一つ以上検出されるかどうかを確認する。切り取られたコマ y_i に対して以下のように定義する。

$$y_i = \begin{cases} 1 & \text{(料理画像が検出された時)} \\ 0 & \text{(料理画像が検出されなかった時)} \end{cases} \tag{1}$$

y_i は一つのコマを表す。また，割合 $f(x)$ と各作品の画像の合計数 L は次式のように定義する。

$$f(x) = \sum_{i=1}^{L} \frac{y_i}{L} \tag{2}$$

これにより各作品の検出率を計算した。

4.　実　　験

まず、SSD300 のボトルネック層となる VGG16 を、imagenet の学習済みモデルを用いて再学習させる [6]。追加ネット

(注4)："ゆめ色クッキング" @くりた陸 (今回集めたデータセットの中にはこの作品を含めていません)

表 1　各作品における IoU 評価 (%)

作品	無加工 (カラー)	グレイスケール	マンガ風加工
A	46.3	36.8	25.4
B	37.4	12.2	17.3
C	43.5	47.5	31.5

ワーク層に対しての初期値は He らによる初期値方法を使用し初期化させている [3]。He らによる初期化方法は活性化関数が ReLU である場合、情報の流れを維持させることができるため使用する。最適化手法は SGD を使用し、momentum 値を 0.9、weightdecay 値を 0.0005 とし、バッチサイズは 4 と定め、学習率は 10^{-3} において epoch40 回、10^{-4} において epoch20 回反復学習させる。data augmentation は random expand,random crop,random flip,mean subtraction を使用し、mean subsection における平均輝度値は UEC FOOD-100 の画像の平均輝度値を求め利用する。学習に用いる料理画像は単品料理と複数料理も兼ねて学習させる。

選択したマンガ 3 作品 A〜C について、3 種類の異なる前処理を施し各作品において検出の精度を評価した。各作品における IoU 評価は表 1 に示し、それぞれのコマに対して 1 つ以上料理が検出された割合を表 2 に示す。なお、検出させる領域は SSD300 モデルにおける確信度が 40% 以上越えるとき，その場所を領域とする。

実験結果よりそれぞれの処理において、各作品における IoU 評価及び検出率が異なった。これはマンガの料理がテキストやキャラクターと重なることが原因となったり、各作者における絵の線画が異なるという影響があるため、作品毎に評価の数値が異なる。検出率の精度において、無加工の前処理における検出率は作品 B を除き 60% を超えた。一方で他の処理においては 40% 程度となり、検出に対しては無加工で行えることを確認した。しかし、IoU 評価はどの手法においても 50% を超えることはなかった。これは物体の検出をすることは可能であるが領域の特定が難しいことを意味する。IoU 評価の値が低い原因として、UEC FOOD-100 が定めるバウンディングボックスと手動で定めたバウンディングボックスの作り方の基準が異なるためである。加えて、物体を検出させることができなかった場合も含めて計算させているため、数値が下がる。

マンガの料理画像を可視化させクラス分類を確認したとき、無加工におけるクラス分類はおにぎり、サンドウィッチ、ざるそば、寿司、のいずれか 4 つの種類に 70% が分類される結果となった。グレイスケール及びマンガ風加工においては検出率は下がるものの、4 つの種類の料理だけでなく他の種類の料理が確認され多様性があった。以上の結果を纏めると、料理マンガの料理を入力させた時、料理を検出させることは可能であるが、領域の特定とその料理の種類を判断することは困難であることが確認された。

5. ま と め

本研究では、自然画像である料理画像を用いてマンガ内の料理画像を検出した。精度は料理マンガの作品に対して影響するが、最も検出率が高かったのは無加工の前処理であることを確認した。しかし、クラス分類の観点において、無加工はおにぎり、サンドウィッチ、ざるそば、寿司、のいずれか 4 つの種類に 70% が分類された。一方で、グレイスケール及びマンガ風加工は 4 つの種類だけでなく様々なクラスが割り当てられることを確認した。

表 2　各作品におけるそれぞれのコマに対して 1 つ以上料理が検出された割合 (%)

作品	無加工 (カラー)	グレイスケール	マンガ風加工
A	62.1	48.5	43.9
B	44.8	28.1	34.4
C	60.7	53.6	46.4

今後の展望として、料理画像を用いて料理を検出させることは可能であるが領域の特定が困難であるため、マンガのテキストを用いることにより精度が向上するかを考えている。また、検出させた料理画像を用いて、料理レシピの自動生成を試みる。

文　献

[1] Ege, T. and Yanai, K.: Simultaneous Estimation of Food Categories and Calories with Multi-task CNN, *2017 Fifteenth IAPR International Conference on Machine Vision Applications*, pp. 198–201 (2017).

[2] Fujimoto, A., Ogawa, T., Yamamoto, K., Matsui, Y., Yamasaki, T. and Aizawa, K.: Manga109 dataset and creation of metadata, *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding* (2016).

[3] He, K., Zhang, X., Ren, S. and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-level Performance on Imagenet Classification, *Proc. IEEE international conference on computer vision*, pp. 1026–1034 (2015).

[4] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. and Berg, A. C.: SSD: Single shot multibox detector, *European conference on computer vision*, pp. 21–37 (2016).

[5] Ogawa, T., Otsubo, A., Narita, R., Matsui, Y., Yamasaki, T. and Aizawa, K.: Object Detection for Comics Using Manga109 Annotations, *arXiv preprint arXiv:1803.08670* (2018).

[6] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).

[7] 柳澤秀彰, 渡辺裕: Faster R-CNN を用いたマンガ画像からのメタデータ抽出, 2016 年映像情報メディア学会年次大会, 14B-1 (2016).