Vol.2019-MPS-126 No.14
2019/12/12

ホワイトボードからの文字抽出手法の検討

山本祐美[†1†2]　本郷仁志[†2]　森吉弘[†2]　城和貴[†1]

概要：ホワイトボードに書かれた文字やイラストから文字領域を抽出することを目的として，風景画像から文字抽出する CRAFT 手法に対して，日本語特有の文字変動を学習したときの評価実験を行った．日本語手書き文字を転移学習させることで，適切に文字領域を抽出できる結果が得られた．

キーワード：ディープラーニング，文字領域抽出，手書き文字，セマンティックセグメンテーション

Extraction of Character from Whiteboard

YUMI YAMAMOTO[†1†2]　HITOSHI HONGO[†2]
YOSHIHIRO MORI[†2]　KAZUKI JOE[†1]

Abstract: In order to extract character regions from characters and illustrations written on a whiteboard, we conduct evaluation experiments with learning Japanese-specific character variations for the CRAFT method that extracts characters from landscape images. By transferring learning of Japanese handwritten characters, we obtain the results that can extract the character area appropriately.

Keywords: deep learning, character detection, handwritten character, semantic segmentation

1. はじめに

近年，スマートフォンの普及によりメモ代わりにカメラ画像で記録することが一般的になってきている．例えば，ホワイトボードを議事録として画像で保存することが多い．これらの画像から文字情報を抽出することで，情報検索が可能となり活用の幅が広がることが期待される．しかしながら，ホワイトボードでは自由な位置，大きさで文字が書かれ，またイラストやマークなど文字以外も一緒に描かれるため，情景画像中からのロバストな文字領域抽出手法が求められている．

情景画像からの文字領域抽出の手法として，古くは文字色や文字の線幅などベーシックな画像特徴を用いた MSER[1]，SWT[2]が提案されている．これらは想定した条件に沿わない文字の場合は文字抽出が困難となる課題がある．近年，ディープラーニングを用いて探索枠を変動させて認識する SSD[3], R-CNN[4]が提案されている．しかし，手書き文字の変動吸収と誤検知抑制するモデルの学習は容易ではない．ディープラーニングを用いた文字領域分割手法として CRAFT[5]が提案されており，高性能な文字領域抽出を実現している．

CRAFT は, 情景画像から文字領域を抽出するセマンティックセグメンテーション[6]の 1 手法である. 情景画像から，英語だけでなく，中国語，日本語，ハングルなど多様な文字を学習し，多言語に対応した文字領域抽出を実現している. 高精度な文字領域抽出を実現しているが, 看板や建物，物などに印刷した活字を対象としている．日本語の手書き文字は変動が大きく，字形だけでなく，文字サイズ，文字幅が変化し，さらには文字が連結する場合もある．これらの変動に対する効果については報告されていない．そこで我々は, CRAFT に日本語の手書き文字を学習させることで, ホワイトボードに書いた文字に対する効果を検証した．

本稿の構成は，以下の通りである．第 2 章では CRAFT の手法を説明し, 手書き文字に対する課題について触れる．第 3 章では手書き文字学習方法について述べる．第 4 章では文字抽出の実験結果を示す．第 5 章で考察し，第 6 章でまとめを述べる．

2. CRAFT とその課題

CRAFT（Character Region Awareness for Text Detection）[5]は，情景画像から文字領域を抽出するフレームワークである．情景画像中から文字領域と，隣接する文字の連結を学習し領域分割を行う．この手法には，バッチ正規化を備えた VGG16[7]に基づいた全層畳み込みニューラルネットワークが採用されている．ネットワーク構造を図 1 に示す．カラー画像を入力し，リージョンスコアとアフィニティスコアを出力する．リージョンスコアは文字の中心である確率を表し，アフィニティスコアは隣接文字間の中心である確率を表す. これらの出力値から文字領域を判定している．

CRAFT では，日本語も学習しているが看板や建物，物などに印刷された活字を対象としている．そのため，文字の

†1 奈良女子大学
Nara Women's University
†2 株式会社コネクテッド
Connected Co

Vol.2019-MPS-126 No.14
2019/12/12

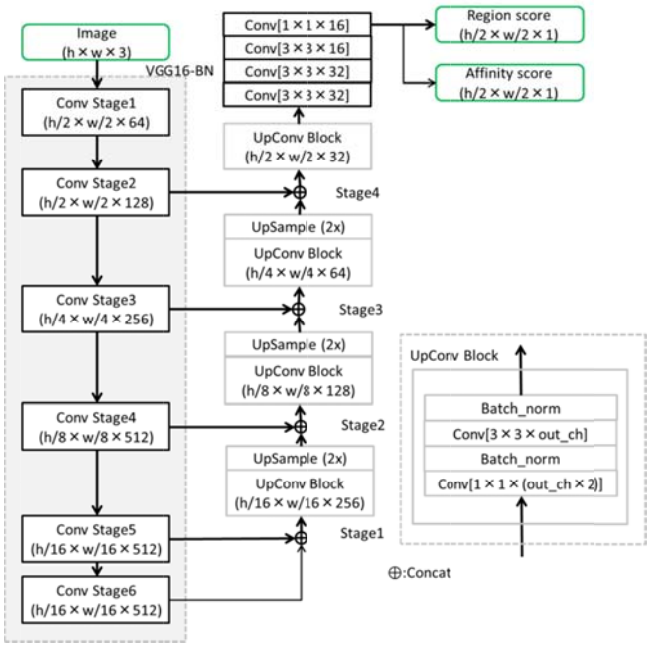


図 1　CRAFT のネットワーク構造

変形としては，フォントデザインやカメラアングル，歪曲した表面に印刷されたことによる変形が主となる．ホワイトボード画像では, 手書き特有の文字変動がある. 句読点，拗音，促音などによる文字の連続性や，文字を囲む線，イラストなどの文字以外が描かれることによる文字領域抽出への影響を評価する必要がある．さらに，CRAFT では VGG16 をベースとして転移学習が可能である. 手書き文字の転移学習による効果も併せて評価する．

3. 手書き文字学習方法

本稿では, CRAFT で用いられたネットワーク構造と同じモデルを使用する．学習データとして，ホワイトボードに書かれた日本語の手書き文字と図やイラストを含むものを 21 枚収集する．まず，各画像に対して，文字ボックスとアフィニティボックスを作成する．今回，文字ボックスは，図 2 の赤枠ように各文字の外接矩形とする．ただし，句読点や長音などのボックスが小さすぎるものや細長すぎるものは，一回り大きめに外接矩形を設定する．次に，アフィニティボックスを作成する．アフィニティボックスは隣接文字間を表すので，図 3 のように，2 つの文字ボックスから作成する．まず，青線で示すように，文字ボックスの対角線を結び，外接矩形を 4 つの三角形に分割する．生成した各三角形の中心をアフィニティボックスの頂点として設定する．この時，隣接文字が横に並んでいる場合，上下の三角形の中心を頂点する．また，縦に並んでいる場合は，左右の三角形の中心を頂点とする．隣接する文字ボックスに対しても，同様の処理を行う．4 つの頂点を結んでできた赤線の四角形をアフィニティボックスとする．最後に，

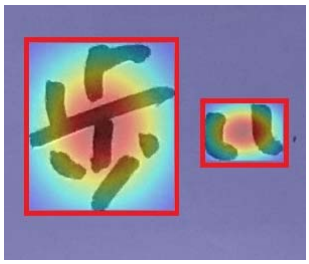


図 2　文字ボックス

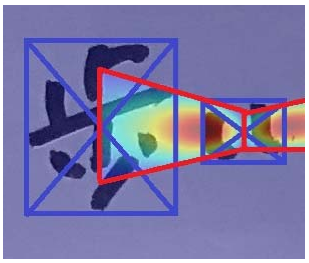


図 3 アフィニティボックス

表 1　文字抽出結果

	手書きモデル	オリジナルモデル
総文字数	2176	2176
検知数	1802	1674
未検知	124	57
文字誤検知数	228	416
イラスト誤検知数	22	29

生成した文字ボックスとアフィニティボックスからリージョンスコアとアフィニティスコアを生成する．2 次元ガウス分布を用意し，各ボックス領域に合わせてマッピングした画像が学習データとなる．作成したスコアの例を図 2, 図 3 に示す．

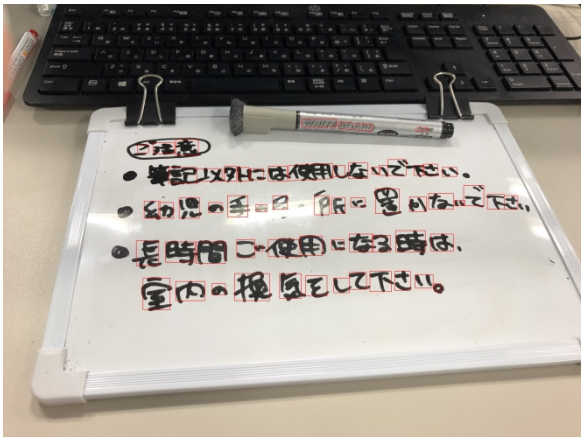
作成したデータセットを用いて学習を行う．しかし，今回収集した学習データ数は少ないため，過学習が起きやすく，未知データに対する性能がでない．そこで，ランダムクロップにより，データ拡張を行った．また，小さい文字に対応するため，元画像を縮小したパターンも追加した．学習データは合計 231 種類用意した．

学習の最適化手法として，Adam[8]を用いた．また，学習率は 0.0001 で固定とした．損失関数は平均二乗誤差を用いた．学習を 100 回行い，2 時間かかった．実験環境は，CPU はインテル® Core™ Intel core i7-7700K プロセッサ 4.2 GHz，GPU は NVIDIA GeForce GTX1080，メモリは 64GB(16GB×4)の計算機を使用した．

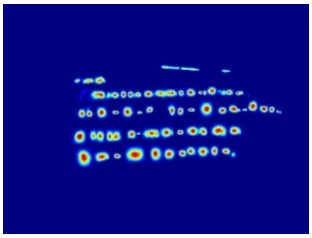
4. 文字抽出実験

テストデータとして, ホワイトボード画像を 20 枚収集した．学習データと同様に，日本語の手書き文字と図やイラストを含む画像である．これらのデータと CRAFT を用いて，ホワイトボードの画像から文字抽出実験を行う．

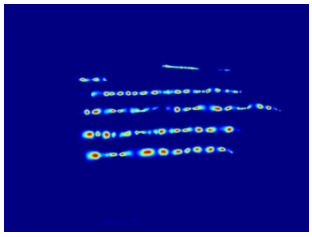
本実験では，手書き文字を学習したモデル（以降，手書きモデルと呼ぶ）と CRAFT のオリジナルモデル（オリジナルモデルと呼ぶ）との比較を行った．表 1 に文字抽出結果を示す．総文字数は，テストデータ 20 枚に含まれる文字の総数を表す．文字全体を適切に検知できた場合を正解とした．未検知数は検知されなかった数を示す．複数文字を



（a）文字抽出結果

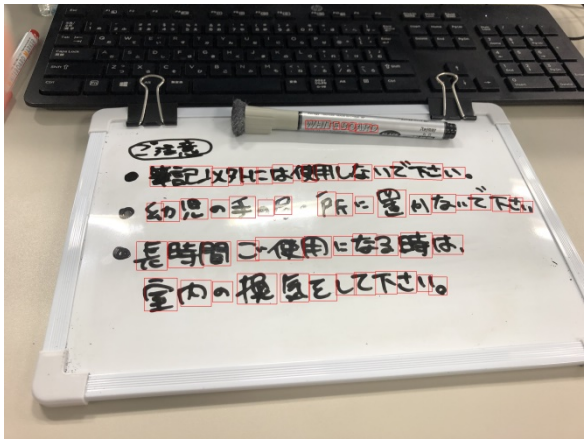


（b）リージョンスコア

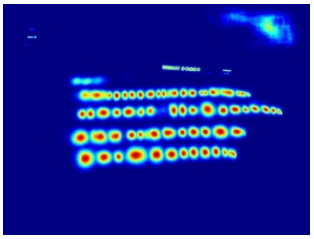


（c）アフィニティスコア

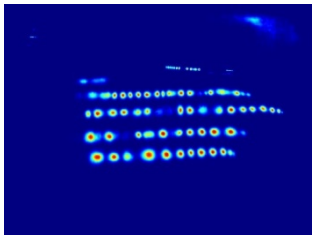
図 4　手書きモデル結果例 1



（a）文字抽出結果



（b）リージョンスコア



（c）アフィニティスコア

図 5　オリジナルモデル結果例 1

1 文字として検知した場合や，文字の一部を検知した場合は誤検知とした．イラストの誤検知は，イラストを文字として検知した数である．

手書きモデルでは検知率が向上し，文字の誤検知数も減りオリジナルモデルより向上した結果が得られた．しかし，未検知数は増加した．オリジナルモデルと比較して，手書きモデルでは句読点が検知されないことが多かった．

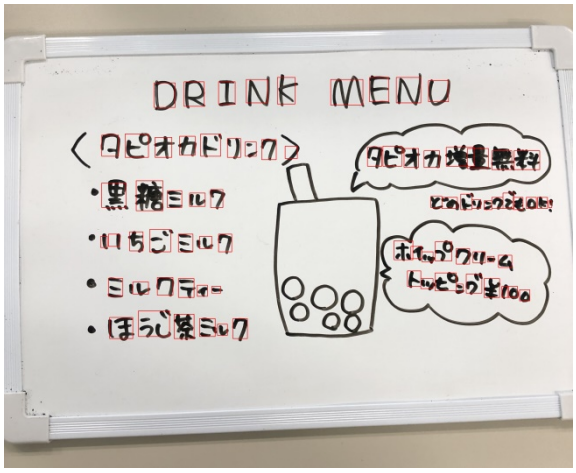
文字抽出結果例を図 4 に示す．赤枠が文字領域として抽出された箇所である．図 4（b）と図 4（c）は，それぞれ図 4（a）に対して出力されたリージョンスコアとアフィニティスコアをヒートマップで表したものである．同画像に対するオリジナルモデルによる結果を図 5 に示す．どちらのモデルも大半の手書き文字に対して適切に文字領域を検知していることが分かる．両モデルでの違いとしては，オリジナルモデルでは近接する文字を分離することができず複数文字を 1 つの文字として検知する傾向が見られた．例えば，図 5 に示した 1 行目の箇条書きでは，「筆記」，「で下さ」など文字と文字が接している箇所では 1 つの文字として検知している．一方，手書きモデルでは文字の外接矩形を文字領域として検知しており，1 文字 1 文字が正確に抽出される傾向が得られた．箇条書き 2 行目の「で」の濁点を手書きモデルでは検知しているが，オリジナルモデルでは検知できなかった．左上に書かれた文字周辺を線で囲んだ「ご注意」は，オリジナルモデルでは検知できなかったが，手書きモデルでは検知できた．

図 6，図 7 はイラストを含む画像である．手書きモデルでは，イラスト部分に対して誤検知しなかった．オリジナルモデルでは，リージョンスコアとアフィニティスコアが反応しているため〇を誤検知している．この理由としては，〇だけでは文字かイラストか判断が難しいと考えられる．

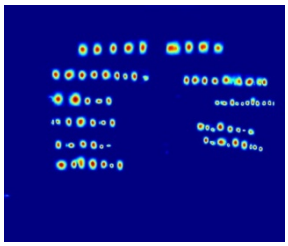
5. 考察

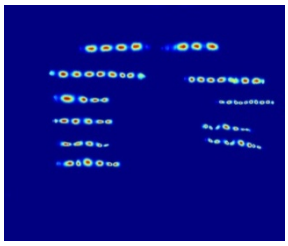
本実験結果から，手書き文字を学習することで文字抽出の精度が高まり，文字と文字がつながった場合や，濁点など手書きの変動を吸収できる結果が得られた．これは，手書きによる崩れた文字を学習したことによる効果と考える．手書きの文字変動は大きく，活字をベースとした学習ではカバーしきれなかったと思われる．また，文字と文字がつながった場合や，文字を線で囲んだ場合も文字領域を外接矩形で学習したことにより外乱による影響を受けにくいモデルが構築できたことが推察される．

両モデルで共通した誤検知としては，1 つの文字を分離して検知する傾向がある．例えば，「い」は 2 文字に分離して抽出される場合が多く見られた．リージョンスコアが 2 箇所に分離して反応し，アフィニティスコアも「い」の中心にあたる部分が少し反応している．「幼」や「所」も同様に分離して検知されている．偏旁と個々で文字が成立する場合，文字領域を分離するか統合するかの選択は本手法では困難であり，文字認識が必要になると思われる．句読点



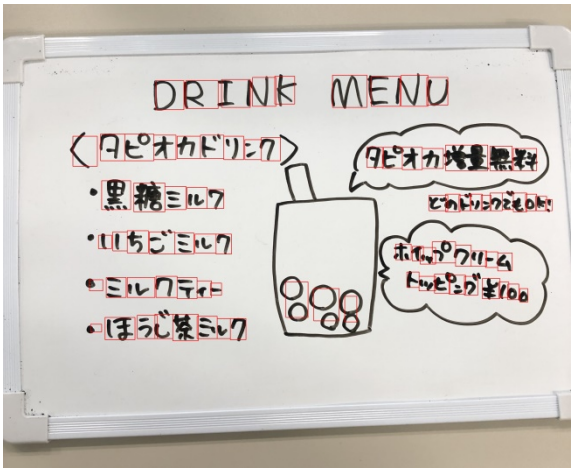
（a）文字抽出結果



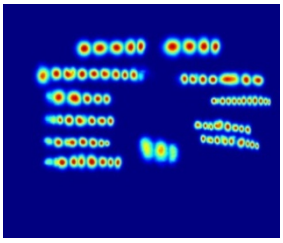


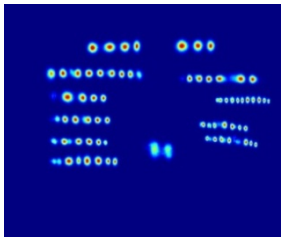
（b）リージョンスコア　（c）アフィニティスコア

図 6　手書きモデル結果例 2



（a）文字抽出結果





（b）リージョンスコア　（c）アフィニティスコア

図 7　オリジナルモデル結果例 2

に関しては, 手書きモデルでも検知率は改善されなかった．句読点の大きさが小さいことと，文字以外の特徴と差が出にくいため検知率が低くなったと考えられる．これには，アフィニティの学習を改善するなど工夫が必要と思われる．また，今回は文字の外接矩形領域を学習したことにより文字領域の抽出精度が向上したと考えるが，定量的な評価は今後の課題としたい．

今回，少ない画像データから転移学習により日本語の手書き文字抽出に対して検知性能を向上させる可能性が示唆された．

6. まとめ

本稿では，CRAFT を用いて，ホワイトボードに書かれた日本語の手書き文字に対して，文字領域の抽出を行った．ホワイトボードには，図やイラストなどの文字以外の線が描かれる場合や，文字を線で囲む場合がある．また．句読点や拗音，促音などの日本語特有の文字変動も含まれる．我々は，文字の外接矩形に 2 次元ガウス分布を設定し，文字領域と文字連結を学習させた．

手書きモデルとオリジナルモデルに対して文字抽出実験と比較を行った．手書きモデルの検知率は 82.8%で，オリジナルモデルの検知率は 76.9%だった．検知率は向上し，文字の誤検知も改善された．手書きモデルは，文字領域を外接矩形で検知し，1 文字を正確に抽出する傾向が得られた．文字領域を外接矩形で学習したことによる効果だと考えられる．定性的ではあるが，本実験から手書きモデルの方が文字領域を適切に抽出する結果が得られた．しかし，句読点などの小さい文字の抽出は改善できなかった．今後は，検知されなかった句読点などの小さい文字や分離されやすい文字に対して改良を行う．

参考文献

[1] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust widebaseline stereo from maximally stable extremal regions. Image and Vision Computing, 22(10):761–767, 2004.
[2] B. Epshtein, E. Ofek, and Y.Wexler. Detecting text in natural scenes with stroke width transform. In CVPR, pages 2963–2970. IEEE, 2010.
[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In ECCV, pages 21–37. Springer, 2016.
[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. PAMI, (6):1137–1149, 2017.
[5] Y. Baek, B. Lee, D. Han, S. Yun, H. Lee: Character Region Awareness for Text Detection. CVPR 2019
[6] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, pages 3431–3440, 2015. 2
[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
[8] Diederik P. Kingma, Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. ICLR, 2015.