Vol.2021-MPS-132 No.4
2021/3/1

近代書籍における文字切り出し手法の検討

飯田紗也香[†1]　竹本有紀[†1]　石川由羽[†2]　髙田雅美[†1]　城和貴[†1]

概要：本稿では，近代書籍のテキスト化に用いるレイアウト解析処理のうち，文字切り出し手法について検討する．情景画像から文字切り出しを行う CRAFT 手法を用いて，近代書籍を対象に評価する．近代書籍とは，明治から昭和初期に刊行された活版印刷の文書である．評価に用いる近代書籍は，帝国議会会議録とする．

キーワード：近代書籍，レイアウト解析，CRAFT，文字切り出し

A Study of Text Detection Methods for Early-Modern Japanese Books

SAYAKA IIDA[†1]　YUKI TAKEMOTO[†1]　YU ISHIKAWA[†2]
MASAMI TAKATA[†1]　KAZUKI JOE[†1]

1. はじめに

現代に発行されている書籍は，文書作成ソフトウェアにより執筆されているため，対応するテキストデータが必ず存在する．そのため，テキストデータを用いた本文内容の検索や音声読み上げ，翻訳，データマイニングなど，さまざまな方法により活用することができる．一方，明治から昭和初期に刊行された近代書籍については，多くの場合，対応するテキストデータが存在しない．帝国議会会議録検索システム[1]では，帝国議会の本会議・委員会の速記録が活版印刷された画像をデジタルデータにより公開している．帝国議会は明治から昭和初期にかけて全 92 回行われている．議会内では，当時の世情を知ることができる議論が行われており，その速記録は近代日本における重要な歴史的資料である．帝国議会会議録をテキストデータ化することで活用の幅が広がり，多くの人が知見を得ることが可能となる．現代の印刷された文書は，光学文字認識(Optical Character Recognition, OCR)を用いて自動的にテキスト化を行うことができる．一般的な OCR ソフトウェアは印刷された文書画像を対象としている．その精度は書籍の保存状態や撮影状態により左右される．また，印刷様式が規格に従っていることを前提としている．近代書籍はノイズを多く含み，フォントの規格が作成されていない．そのため，近代書籍である帝国議会会議録に対して OCR ソフトウェアを用いた正確な文字認識は難しい．

そこで，近代書籍画像に特化した文字認識[2]に用いることを目的として，畳み込みニューラルネットワーク(Convolution Neural Network, CNN)による Semantic Segmentation[3]を用いたレイアウト解析手法[4,5]を提案している．Semantic Segmentation とは，画像を構成する画素ごとに意味を自動的に割り当てる手法である． Semantic Segmentation を用いたレイアウト解析では，近代書籍画像から文字領域，枠領域，文書領域を抽出する． この手法を帝国議会会議録画像に適用したところ，99%の文書領域抽出を実現している．しかし，この手法は抽出された文書領域からの文字切り出しには対応していない．そのため，抽出された文書領域からの文字切り出しを行うには，既存のボトムアップなアプローチによるレイアウト解析手法を用いている．ヒストグラムを用いる手法では，漢数字の「二」のような文字を構成する部品が上下に分かれている文字や，文字列間に余白が存在しない場合への対応が難しい．そこで，本稿では CNN による文字切り出し手法の検討を行う．

CNN を用いる情景画像からのテキスト領域検出手法として，CRAFT(Character Region Awareness for Text Detection)[6]という手法が提案されている．CRAFT とは，RGB の情景画像から任意の形状の文字列を抽出するフレームワークである．CRAFT を白黒 2 値画像である会議録画像に試験適用したところ，8 割程度の文字が切り出された．この試験適用では，公式から配布されている CRAFT 用の学習データを用いており，近代書籍画像を一切学習していない状態である．CRAFT の手法を用いて適切なデータセットの学習を行えば，近代書籍からの文字切り出しに利用できると期待できる．本稿では，CRAFT の手法を用いて近代書籍画像に対応したデータセットを学習した場合について，近代書籍画像からの文字切り出しの有用性を検証する．

本稿の構成は，以下の通りである．第 2 章では既存のレイアウト解析手法について説明する．第 3 章では，CRAFT の手法を説明し，同手法を近代書籍画像に対して用いる場合の課題について述べる．第 4 章では，近代書籍画像に対する文字切り出しの実験を行う．文字切り出しの実験方法について説明し，その実験結果と考察を示す．

1 奈良女子大学
Nara Women's University
2 滋賀大学
Shiga University

2. 既存のレイアウト解析手法

文書画像に対するレイアウト解析では，画像内の領域抽出と分類を行う．一般的な OCR ソフトウェアに用いられるレイアウト解析手法には，ボトムアップなアプローチと，トップダウンなアプローチの 2 つの方法がある[7].

ボトムアップなアプローチでは，まず，文書画像に含まれる文字や記号を構成する細かいパーツの検出を行う．次に，検出されたパーツを同じ文字や単語，文章，段落で繰り返し統合を行うという流れで，反復的な解析を行う．この方法の利点は，任意の形状をした領域を処理することが可能な点である．しかし，テキストの並び方などのページの構造を考慮するためには，文書中の文字や記号全てについて，繰り返し領域の分類と検出を行わなくてはならない．そのため，計算コストがかかるという短所がある．

トップダウンなアプローチでは，書籍画像に含まれる空白領域や幾何学的情報に基づいて，文書を段落や行の集合ごとに分割する．その領域が単一の領域になるまで分割を繰り返すという流れでレイアウト解析を行う．トップダウンな方法の利点として，段落などのページの構造を直接解析できること，反復的に分類を行う必要がなく高速であることが上げられる．欠点として，安定した解析結果を得るためには，文書のレイアウトについて前提条件が必要である．前提条件とは，文書を構成する要素に図や枠線が含まれるか，非矩形領域が存在するか等の情報である．

文書画像のレイアウト解析における共通した問題として，ノイズと画像の傾きが挙げられる．ノイズとは，インクの染みや書籍の汚れによるごましおノイズなどを指す．画像の傾きとは，文字列の並びが完全な水平もしくは垂直ではない状態を指す．書籍を撮影する際，ページのたわみや撮影位置などにより文書画像が回転して撮影されることで傾きが生じる．近代書籍は，明治から昭和初期に刊行されているため，書籍自体の状態が劣化している．また，印刷された当時にはフォントや印刷様式など，統一された規格が存在しない．そのため，文字列の並びが完全に垂直および水平ではない場合がある．書籍画像の状態によっては，ノイズや傾きが多くなり，レイアウト解析が困難となる．

Semantic Segmentation を用いたレイアウト解析手法では，非矩形領域の構造をしている文書領域の検出が可能である．しかし，1 文字ごとの文字切り出しなど，レイアウト構造の細部までの検出を行うことができない．CRAFT を用いた情景画像からのテキスト抽出手法では，文字ごとの切り出しや，任意の角度および形状の文字列の切り出しが可能である．CRAFT を用いれば，書籍画像からの文字切り出し等．細部の検出が期待できる．

3. CRAFT

CRAFT(Character Region Awareness for Text Detection)とは，RGB の情景画像から，任意の形状に並ぶテキスト領域を抽出する手法である．CRAFT では，一般的な CNN のようにラベル付けされた画像中の領域を学習するのではなく，検出対象における中心位置の確率を，ヒートマップにより学習する．出力されたヒートマップを用いて，文字ごとの矩形切り出しを行う．それらの連結点を検出し，テキスト領域の推定を行う．CRAFT では，このボトムアップなアプローチにより，任意の方向を向いたテキスト，湾曲したテキスト，変形したテキストなど，情景画像に含まれる複雑な形状のテキスト検出が可能となる．

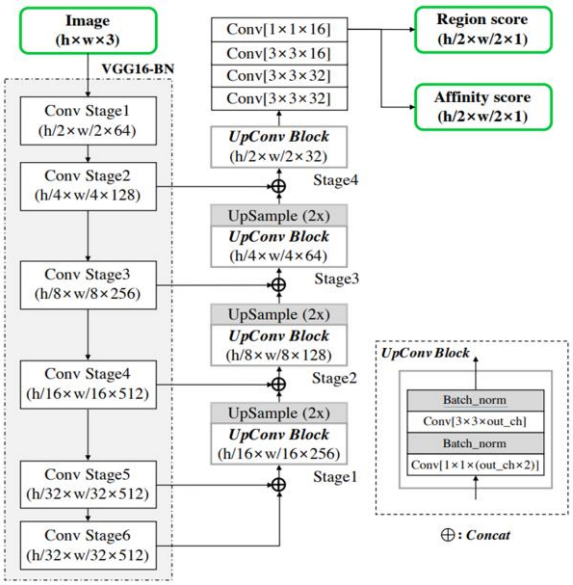

図 1　CRAFT に用いられる CNN アーキテクチャ

CRAFT に用いられる CNN の構造は，図 1 に示されるとおりの，VGG-16[8]に基づいた完全畳み込みネットワーク(Fully convolution network, FCN)[9]である．このニューラルネットワークは FCN の一種である U-net[10]と類似した構造をしており，エンコーダからデコーダにかけてスキップ接続を持つ．スキップ接続では，畳み込み層による特徴抽出を行った後の特徴マップを保存し，逆畳み込み層に足し合わせる処理を行う．この処理により，プーリング層において失われる認識対象の位置情報を復元することができる．このニューラルネットワークに対して，RGB の情景画像を入力すると，文字の中心である確率と，文字間における連結点の中心である確率を示すヒートマップを，それぞれ出力する．出力されるヒートマップは，入力画像の半分の解像度である．文字の中心である確率を示すヒートマップをリージョンスコアと呼ぶ．文字間の文字の隣接領域の中心である確率を表すヒートマップをアフィニティスコアと呼ぶ．これらのヒートマップをもとに，文字ごとの矩形切り出しや単語領域の判定を行う．

情景画像を対象とした学習済みモデルを用いて，CRAFT による帝国議会会議録画像の文字切り出しを行う．その結果，図 2 の右側に示すようなリージョンスコアが出力され

Vol.2021-MPS-132 No.4
2021/3/1

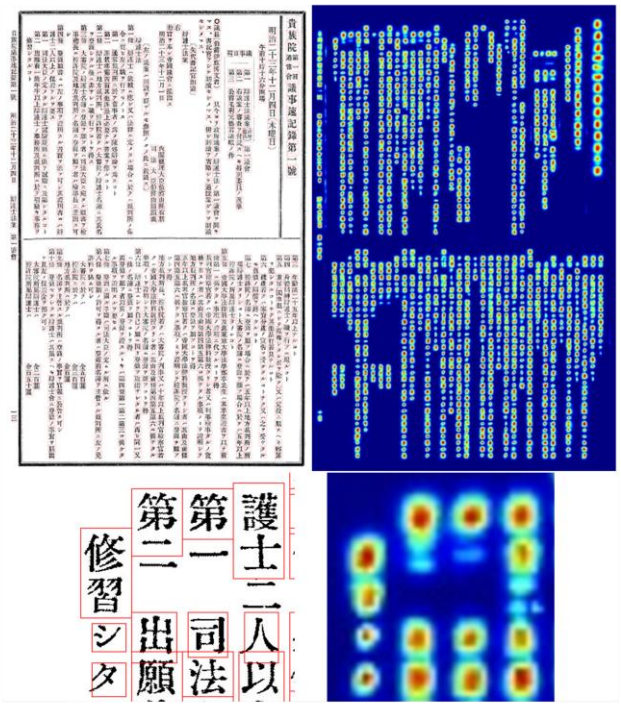

図 2　情景画像用の学習モデルを用いた会議録の
切り出し結果(左)とリージョンスコア(右)

る．リージョンスコアを元に文字ごとの矩形切り出しを行うと，図の左側のように 8 割程度の文字が切り出された．この学習済みモデルには，英語に対応した学習データに加えて，中国語の看板画像等をアノテーションしたデータセットである CTW-1500[12]が用いられている．CTW-1500 は漢字に対応しているが，図 2 の左側下部に示すように漢数字の「二」等の分離した構造を持つ文字の切り出しに失敗する．また，文字の位置が上下に近い場合や，画数の多い文字が連続する場合，文字ごとの中心位置を示すヒートマップの距離が近く境界があいまいになり，切り出し矩形が連結する傾向にある．CRAFT の対象は，情景画像中に含まれる看板や建物等に印刷された活字文字である．そのため，書籍画像のように，1 枚に含まれる文字数が多い画像や，白黒 2 値画像には対応していない．

本稿では，CRAFT の手法を用いて近代書籍画像からの文字切り出しが可能であるか検証を行う．近代書籍文字に対応したデータセットの作成し，学習を行う．

4. 実験

4.1 実験方法

本稿では，CRAFT に用いられるニューラルネットワークと同じ構造のモデルを学習に使用する．学習データとして，近代書籍文字画像を用いて書籍画像を生成する．生成された書籍画像を学習したモデルを用いて，文字切り出し矩形の検出について実験を行う．

近代書籍文字を用いた書籍画像の生成方法について述べる．学習データの生成は，文字画像の配置，リージョンボックスとアフィニティボックスの作成，ヒートマップの作成という流れで行う．作成する書籍画像は，一辺 1024px の正方形に正規化する．画像内に文字を納めるため，最大文字数は約 700 文字とする．

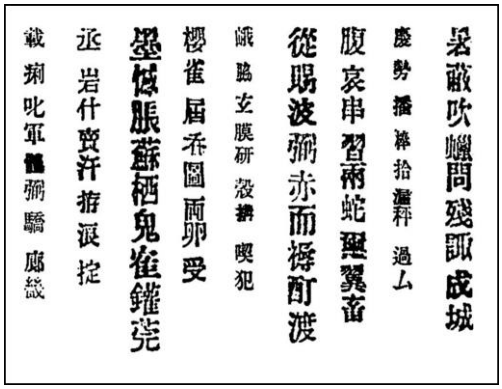

図 3　生成された近代書籍文字画像

まず，文字の配置について説明する．配置する文字画像には，国立国会図書館デジタルコレクション[12]で公開されている近代書籍画像から収集されたひらがな，カタカナ，漢字の文字画像 3044 種を用いる．記号や句読点，括弧などは含まれない．文字の配置は，図 3 に示すように縦書き方向上詰めのレイアウトとする．1 行の文字数と使用する文字種は，乱数によりランダムに決定する．文字画像は行ごとにランダムな大きさに拡大縮小する．文字画像を配置する際，上下方向に 0 から 16px のずれをランダムに入れる．配置にずれを入れることで，近代書籍の規格化されていない印刷様式を再現する．

次に，リージョンボックスとアフィニティボックスの作成について説明する．リージョンボックスとは図 4(a)に示すように，文字画像の角 4 点の座標により構成された矩形領域である．文字画像の貼り付け位置と，文字画像の幅および高さを用いて作成する．アフィニティボックスとは，隣接文字間の領域であり，2 つの文字ボックスから作成する．図 4(b)に示すように，隣接するリージョンボックスを用いてアフィニティボックスを定義する．それぞれのリージョンボックスの対角線を結ぶことで，2 つの三角形を作成する．リージョンボックスが縦方向に並ぶ場合はリージョンボックス内で左右に並ぶ三角形，横方向の場合は上下に並ぶ三角形を用いる．上下に隣接する 2 つの文字ボックスについて，三角形の重心 4 点を，それぞれアフィニティボックスの頂点とする．最後に，ヒートマップの作成について述べる．2 次元ガウス分布を作成し，図 4 のように各矩形領域に合わせてマッピングを行うことでヒートマップを作成する．学習に用いるニューラルネットワークの出力サイズにあわせて，作成するヒートマップは対応する書籍画像の半分の解像度に縮小する．

学習条件を説明する．使用する学習データの数は 10000

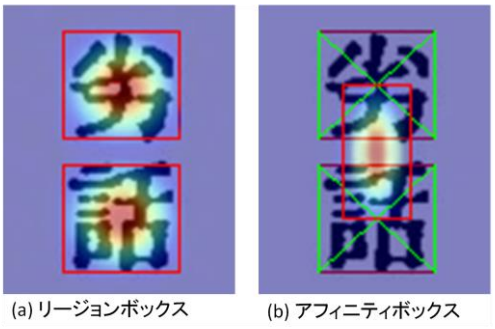

図 4　生成された矩形とヒートマップ

表 1　生成画像からの切り出し結果

	抽出文字数	総文字数	割合
画像 1	706	706	100.0%
画像 2	706	706	100.0%
画像 3	703	703	100.0%
画像 4	710	710	100.0%
画像 5	709	709	100.0%
画像 6	707	707	100.0%
画像 7	704	704	100.0%
画像 8	707	707	100.0%
画像 9	704	704	100.0%
画像 10	702	702	100.0%

種類で，ミニバッチ学習により学習を行う．バッチサイズは 8 である．最適化手法には Adam を用いて，学習率は 0.0001 とする．損失関数には平均二乗誤差を用いる．学習回数は 1000Epoch で, 計算時間は 9 日 1 時間 11 分である．

学習終了後，学習モデルの文字切り出し精度の検証として，書籍画像からの文字切り出し実験を行う．今回の実験では，アフィニティスコアによる文字の連結は行わない．入力画像 1 枚に含まれる全ての文字のうち，切り出された文字の割合を用いて評価を行う．テストデータとして，学習データと同じ方法により生成された近代書籍文字画像と，帝国議会会議録画像を用いる．まず，生成画像に対する評価について説明する．文字切り出し矩形では，文字の端が 2,3px 切り出し矩形からはみ出している文字も正常に切り出されたと判定する．これは，文字のストロークが十分に保たれていればオフライン文字認識において十分認識が可能なためである．文字の端が途切れる量の測定のため，文字として切り出された矩形の範囲外に存在する黒画素の量を用いる．生成画像には枠線や図，インクの染みなど，文字以外の黒画素が含まれていない．切り出された文字領域を白画素で埋め, その画像 1 枚に含まれる黒画素の量が 1% 未満であれば，切り出された文字の欠損は文字認識に支障が出ない程度であると判定することができる．次に，会議録画像に対する評価について述べる．帝国議会会議録検索

表 2　生成画像の切り出し矩形外の黒画素の割合

画像 1	0.16%	画像 6	0.15%
画像 2	0.17%	画像 7	0.15%
画像 3	0.15%	画像 8	0.20%
画像 4	0.17%	画像 9	0.15%
画像 5	0.20%	画像 10	0.13%

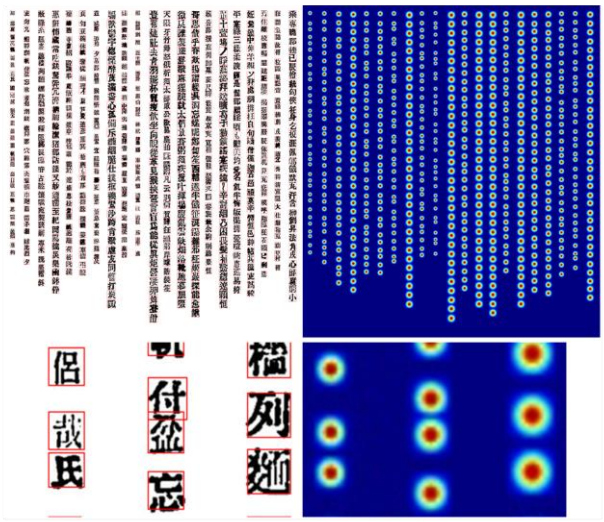

図 5　生成画像からの切り出し結果(左)と
リージョンスコア(右)

システムにより公開されている会議録画像は，縦幅約 4600px，横幅約 3200px と解像度が高く，1 ページにつきおよそ 1000 から 2000 の文字が含まれている．学習された近代書籍画像は一辺が 1024px の正方形であり, 文字数は最大で約 700 文字である．そのため，今回作成する学習モデルへ会議録画像をそのまま入力する場合，精度の高い認識結果を得ることは期待できない．そこで，学習モデルへ入力する会議録画像を，学習データの近代書籍文字画像の文字数と解像度に合わせて分割する．このとき，重複部分を含めて分割を行うことで，切り出された文字の取りこぼしを防ぐ．画像の統合は，出力画像のヒートマップをもとに，重複部分の削除を行う．統合された会議録画像から正常に切り出された文字矩形を数え，評価を行う．

4.2 実験結果と考察

切り出し精度の評価方法として，画像 1 枚に含まれる文字のうち，切り出された文字の割合を示す．学習データと同じ方法で生成された近代書籍文字画像と，帝国議会会議録画像を学習モデルに入力する．それぞれの画像に含まれる文字のうち，切り出された文字の割合を算出する．

まず，学習データと同じ方法で生成された近代書籍文字画像 10 枚に対して, 学習モデルを用いた文字切り出しを行う．結果は表 1 に示す通り，画像 1 枚に含まれる文字のうち 100%の文字が切り出される．画像 1 枚に対する切り出し矩形外にある黒画素の割合は，表 2 に示す通りで，平均

表 3　会議録画像からの切り出し結果

	抽出文字数	総文字数	割合
画像 A	848	938	90.4%
画像 B	1358	1388	97.8%
画像 C	1761	1895	92.9%

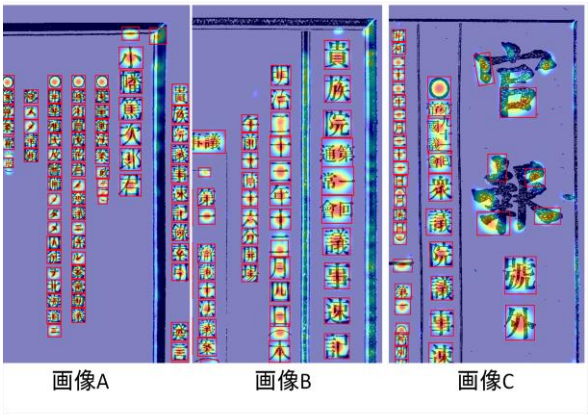

図 6　会議録画像からの切り出し結果と
リージョンスコア

すると 0.16%である．これは 1%を下回るため，切り出された文字の欠損は文字認識に支障が出ない程度であるといえる．上下の距離が近い文字では，図 5 左下のように文字切り出し矩形が上下に重複する場合がある．矩形が重複する場合も図 5 の右下に示すように，リージョンスコアは文字ごとに明確に分かれている．よって，学習データと同じ方法により生成される近代書籍文字画像に対しては，十分な精度で文字切り出しを行うことが可能であるといえる．

帝国議会会議録画像 3 枚に対して，学習モデルを用いた文字切り出しを行う．その結果，画像 1 枚から切り出される文字の割合は，表 3 に示す通り 3 枚ともが 90%を超える精度である．表 3 の画像 1 枚に含まれる文字を平均すると 93.7%の文字が切り出される．文字以外の誤検知には，枠線等の文字ではない領域が切り出されることがある．

会議録画像に含まれる，文字以外の領域における誤検知について述べる．会議録画像 3 枚からの文字切り出し矩形とリージョンスコアの例を図 6 へ示す．リージョンスコアに着目すると，枠線部分にヒートマップが反応する傾向が確認できる．これにより，枠線の一部が文字領域と誤認識され切り出される．誤検知の原因として，学習に用いる近代書籍文字画像には枠線など，文字以外の情報が含まれていないことが上げられる．図 7 に正しく切り出されなかった文字の例を挙げる．本文の文字とサイズが大きく異なる文字は，正しく切り出されない傾向にある．図 7 の(a)に示すように，小さい文字は，リージョンスコアが結合して検出される．そのため，複数の文字が 1 つの文字として結合して切り出される．図 6 の画像 C に含まれる表題の「官報」のような大きい文字について，リージョンスコアが分離し

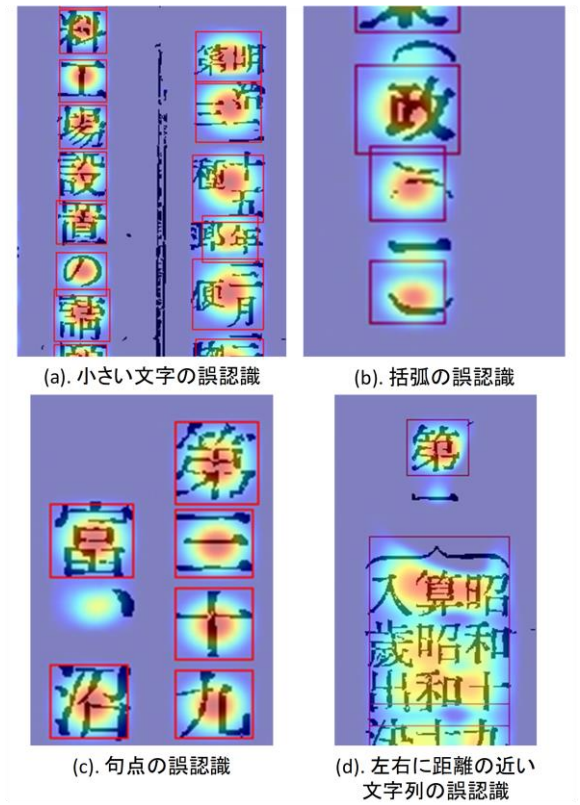

図 7　会議録画像の誤検出例

て検出される．誤検出の原因としてリージョンスコアに着目すると，本文の文字サイズに合わせヒートマップが出力される傾向が確認できる．そのため，本文の文字とサイズが大きく異なる文字は誤検出されると推測される．誤検出されやすい文字種は，読点や記号，括弧の誤検出が多く見られる．特に，図 7 の(b)に示すように，「)」と「(」の境界はうまく検出されず，2 文字が統合され切り出されやすい．表 3 および図 6 に示す画像 A は，文中に括弧を多く含む．画像 A では，正しく検出されない文字のほとんどが括弧である．図 7 の(c)に読点を誤認識した例を示す．リージョンスコアに着目すると，ヒートマップは反応しているが，強く反応しておらず，範囲も小さい．読点は他の文字と比較してサイズが小さい．そのため，文字矩形として切り出されにくいと考えられる．また，読点や括弧の誤認識の原因として，学習に用いるデータセットに，これらの文字種が含まれていないことが上げられる．上下の距離が近い文字は正常に切り出される．しかし，図 7 の(d)に示すように，左右の距離が近い文字はリージョンスコアが結合しており，文字領域の切り出しに失敗している．これは，学習データセットのレイアウトが縦書きであり，横方向の文字間距離が一定であるためと推定できる．

本実験により，近代書籍文字に対応するデータを学習することで, CRAFT の手法を用いた近代書籍からの文字切り出しの有用性が示唆される．学習データに句読点や括弧，

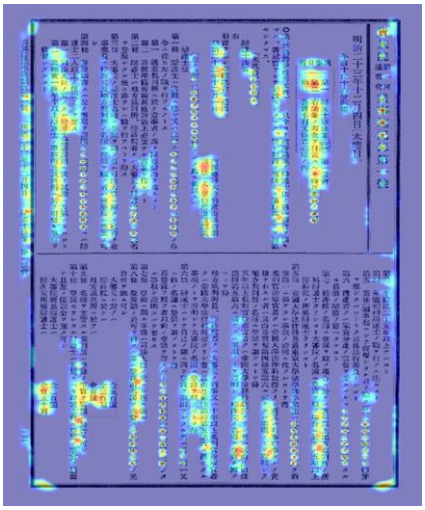

図 8　未分割で入力された会議録の
リージョンスコア出力結果

記号などの文字種を追加することで，誤検出された文字種に対応できると考えられる．また，文字以外の誤検知について，枠線やインクの染みなどのノイズを学習することにより削減できると推察する．今回，学習に用いる近代書籍文字画像では，文字を 1024px の正方形内に収めるため，最大文字数が約 700 文字となるよう設定している．これにより，会議録画像のような文字数の多い書籍画像に対応していない．学習モデルへ会議録をそのまま入力した場合，図 8 のように文字が存在する領域でリージョンスコアのヒートマップが反応しない．バッチサイズを減らし，学習データの解像度を上げることで，文字数の多い書籍画像への対応できると推測する. 今後は, 学習データの解像度を上げ，含める文字種を増やしていく．近代書籍画像からの文字切り出しを行うニューラルネットワークの精度向上を目指す.

5. まとめ

近代書籍画像に特化したレイアウト解析のために，近代書籍に対する文字切り出し手法の検討を行っている．情景画像を対象とする CRAFT 手法について，近代書籍である帝国議会会議録を対象に評価する.

CRAFT は RGB の情景画像を対象とする，テキスト領域の検出手法である．ヒートマップにより文字および文字の連結位置を検出することで，任意の形状の文字列領域を検出するボトムアップなアプローチである．本稿では，CRAFT と同じ方法を用いて, 近代書籍に対応する学習モデルの作成を行う．作成したデータセットを用いて，CRAFT と同じ構造のニューラルネットワークによって学習を行う. その学習モデルを用いて, 文字切り出しの精度を検証する.

近代書籍文字画像が学習されたモデルを用いて実験と比較を行う．学習データは近代書籍画像から抽出された文字画像を用いて，10000 種作成する．本実験では，学習データと同じ方法により生成された画像 10 枚と, 帝国議会会議録画像 3 枚に対して文字切り出しを行う．画像 1 枚に含まれる文字のうち，切り出された文字の割合を算出し，評価を行う．会議録画像については，学習データより解像度が高く，含まれる文字数が多いため，分割して入力を行う．出力された会議録は，文字の中心領域である確率を示すリージョンスコアのヒートマップを用いて統合する.

実験の結果，学習データと同じ方法により生成された画像は 10 枚全てについて，文字が 100%の精度で切り出される. 検出された文字領域矩形からはみ出る黒画素の割合も，平均で 0.16%と文字認識の妨げにならない程度である．会議録画像については，平均で 93.7%の精度で文字が切り出される．文字領域以外の誤検出や，正確に切り出されない文字種については，学習データの充足により対応できると推測される．本実験結果より，CRAFT の手法を用いる近代書籍画像からの文字切り出しの有用性が確認できる．今後は，学習データの解像度を上げ，含める文字種やレイアウト構造を増やすことで, 文字切り出しの精度向上を目指す.

謝辞　本稿は MEXT 科研費 JP20H04483 の助成を受けたものです.

参考文献

[1] 帝国議会会議録検索システム https://t eikokugikai-i.ndl.go.jp/(参照 2021-1-28)
[2] Yasunami, S., et al. "Applying CNNs to Early-Modern Japanese Printed Character Recognition."
[3] Thoma, Martin. "A survey of semantic segmentation." arXiv preprint arXiv:1602.06541 (2016).
[4] 飯田紗也香, 竹本有紀, 石川由羽, 髙田雅美, & 城和貴. (2019). 帝国会議会議録における semantic segmentation を用いたレイアウト解析. 研究報告数理モデル化と問題解決 (MPS), 2019(6), 1-4.
[5] Iida, Sayaka, et al. "Layout analysis using semantic segmentation for Imperial Meeting Minutes." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2019.
[6] Baek, Youngmin, et al. "Character region awareness for text detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[7] Lee, S. W., & Ryu, D. S. (2001). Parameter-free geometric document layout analysis. IEEE Transactions on pattern analysis and machine intelligence, 23(11), 1240-1256.
[8] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
[9] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[10] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.
[11] Yuliang, Liu, et al. "Detecting curve text in the wild: New dataset and new solution." arXiv preprint arXiv:1712.02170 (2017).
[12] 国立国会図書館デジタルコレクション https://dl.ndl.go.jp/(参照 2021-1-28)