奈良女子大学大学院修士論文

邦字新聞に対応したOCRシステムの開発

奈良女子大学大学院人間文化総合科学研究科

情報衣環境学専攻

（学籍番号：22720045）

熊谷　もも

令和5年1月

概要

本修士論文では，邦字新聞に対応した OCR システムを提案する．システム開発のため，既存システムである NDLOCR に用いられる一部手法の置き換え，追加を行う．

明治から昭和初期までの時代に刊行された近代書籍のうち，邦字新聞というメディアがある．邦字新聞とは，明治維新前後よりアメリカ大陸，アジアにおいて日本人移民により刊行された海外日系新聞の集合体である．邦字新聞は様々な出版社により刊行され，その性質は，当時の対象購読者や政治状況により，コミュニティ，政治プロパガンダなど多岐にわたる．現在，スタンフォード大学フーバー研究所にて収集された邦字新聞画像データが，邦字新聞デジタル・コレクションとして公開されている．邦字新聞は当時の暮らしや状況を知るための重要な資料であり，邦字新聞デジタル・コレクションの全文自動テキスト化が求められている．現在，光学文字認識(Optical Character Recognition，OCR)を用いた自動テキスト化が進められている．しかし，邦字新聞に対する既存 OCR システムの精度の低さにより，その成果は未だ十分ではない．

日本語書籍に対応した OCR システムの 1 つに，NDLOCR がある．現在，NDLOCR ver.1.0，ver.2.0，ver.2.1 のソースコードがウェブ上で公開されており，誰でも実行が可能である．NDLOCR は，国立国会図書館が所蔵する約 262 万タイトルの書籍資料画像によって学習データが構築されている．また，日本語文書に用いられるほとんどの文字種に対応している．よって，幅広い年代やジャンルの日本語文書に対して高精度なテキスト化が可能である．しかし，邦字新聞は多段組みレイアウトかつ，見出しや広告，図などの様々な要素を含む複雑なレイアウト構成を持つ．また，活版印刷により文字のフォントやサイズが規格化されていない．更に，出版社の技術レベルに問題があるだけではなく，保存状態も劣悪なものが多い．以上のような一般的な文書と異なる特徴を持つことにより，邦字新聞に対する NDLOCR の精度は不十分である．NDLOCR が邦字新聞に対応するには，3 つの課題点がある．

1 つ目の課題は，レイアウト解析精度の低さである．NDLOCR の学習に使われる文書画像は，その仕様上，1 段組みから 2 段組み程度の単純な構成をとるものが想定されている．邦字新聞のような多段組みかつ，段組み内に図や広告が挟まる複雑なレイアウト構成には当然対応できず，レイアウト解析精度が不十分である．一般に，OCR システムにおけるレイアウト解析精度が低いと，全体の認識精度が低下する．よって，邦字新聞に対応するにはレイアウト解析精度の改善が必要となる．

2 つ目の課題は，ルビが含まれる行に対する文字認識精度の低さである．邦字新聞のほとんどの漢字にはルビが振られているが，NDLOCR では邦字新聞のルビを検出する

ことができない．その原因として，邦字新聞の活版印刷により文字のサイズが不統一であることが挙げられる．ルビが検出できない場合，ルビがノイズとなり文字認識の精度が低下する．邦字新聞に対応するには，ルビ除去手法の検討が必要である．

3 つ目の課題は，読み順検出精度の低さである．NDLOCR ver.2.0 以降には，視覚障がい者などに向けた読み上げ機能がある．その処理の 1 つに，読み順の整序処理がある．NDLOCR における読み順整序処理では，余白を利用して文章のひとかたまりを分割する処理を再帰的に行い，いくつかのブロックを形成する．形成されたブロックごとに簡単なルールベースで読み順を決定する．邦字新聞は，1 ページが様々な主題の記事で構成される．記事ごとに文章の配置が異なり，文章の間に図や広告が挟まる複雑な構成をとる．そのため，読み順検出において NDLOCR のような構造的観点のみで決定することが困難である．読み順が正しく検出できず，テキスト出力が文章として成り立たなければ，研究に向けた活用や邦字新聞デジタル・コレクションの全文検索機能に対する活用において不便となる．よって，読み順検出手法の改善が必要である．

以上の課題を踏まえ，本修士論文では NDLOCR に対して手法の改善を行い，邦字新聞に対応した OCR システムとして提案する．邦字新聞画像を用いて NDLOCR との精度比較を行い，改善手法の有効性を検証する．

1 つ目に，レイアウト解析手法の変更を行う．提案手法では，レイアウト解析手法として解像度ピラミッドを適用した CRAFT の手法を採用する．CRAFT は文字領域を抽出する手法の 1 つであり，CNN により文字の中心と文字の連結部分における 2 つのヒートマップを得る．CRAFT に解像度ピラミッドを適用することで，学習モデルの構築にかかる計算資源を削減し，多段組みで構成される近代書籍に対応したモデルが構築されている．邦字新聞に対応したレイアウト解析手法として有効性を検証するため，NDLOCR のレイアウト解析手法と精度を比較する．検証の結果，NDLOCR において精度が低い結果であった邦字新聞画像が，提案手法では大幅な精度向上がみられた．よって，邦字新聞画像におけるレイアウト解析手法として有効であることが示された．更なる精度向上のため，前処理，段組みを区切る横線の除去手法の検討が必要である．

2 つ目に，ルビ除去処理を追加する．ルビ除去手法として，フーリエ記述子と濃度ヒストグラムを用いる．レイアウト解析により得られる行領域には，ルビが含まれる行と含まれない行が混在する．そのため，事前にルビの有無判定を行う．まず，フーリエ記述子により行領域内の文字の輪郭の概形を得る．次に，得られる概形に対して縦方向濃度ヒストグラムを取得する．ヒストグラムの谷となる部分があるものは，ルビが含まれる行であると判定する．最後に，ルビが含まれると判定された行に対して，ヒストグラムの谷となる位置より右側部分をルビとして分離する．ルビ除去処理の有効性を検証するため，NDLOCR と提案手法の文字認識精度を比較する．検証の結果，提案手法が NDLOCR の文字認識率を上回る結果となり，ルビ除去手法が文字認識精度の向上に有効であることが示された．しかし，ルビが含まれない行に対して，誤ってルビ除去処理

を行う例がみられた．誤った位置でルビ除去を行うと，文字認識の低下につながる．ルビの有無判定における精度向上のため，手法の改善が必要である．

3 つ目に，読み順検出手法を変更する．読み順検出を行うため，1 ページが複数の記事で構成される邦字新聞に対し，同一記事推定を行う．邦字新聞は，1 つの記事を構成する文章のブロックが分散的に配置される場合がある．文章の内容を考慮し，同一記事を決定する必要がある．まず，レイアウト解析の出力を利用して余白や線で分けられる文章ブロックを分割する．そして，Okapi-BM25 とコサイン類似度計算により文章ブロック同士の類似度を算出し，類似度の高いブロックを同一記事と推定する．読み順検出手法における検証の結果，1 つの記事を構成する文章ブロックの一部を同一記事として推定することが可能であることが確認された．しかし，同一記事に含まれるすべての文章ブロックを網羅し，抽出することは困難であった．今後は，文章的観点と構造的観点の両方を考慮し，他手法と組み合わせることにより読み順検出精度の向上を目指す．

提案手法は，スタンフォード大学フーバー研究所上田薫教授のもとで実際の運用が期待されている．異なるマシン間でも移植を容易にし，システムを正常に動作させるため，Docker を用いてコンテナ型の仮想環境構築を行う．Docker の利用により，異なるマシンの Ubuntu，Windows にて正常にシステムが動作することを確認する．

キーワード：邦字新聞，OCR，レイアウト解析，ルビ除去

目次

図目次

表目次

第1章　はじめに

明治から昭和初期までの時代に刊行された近代書籍のうち，邦字新聞というメディアがある．邦字新聞は，明治維新前後よりアメリカ大陸，アジアにおいて日本人移民やその第二，第三世代により刊行された海外日系新聞の集合体である．現在，スタンフォード大学フーバー研究所により収集された邦字新聞画像データが，邦字新聞デジタル・コレクション[1]として公開されている．邦字新聞は様々な出版社により刊行され，その性質は，当時の対象購読者や政治状況により，コミュニティ，軍事プロパガンダまで多岐にわたる．これらの情報は，当時の暮らしや政治を知るための重要な資料である．研究に対する利用，邦字新聞デジタル・コレクションの全文検索機能の追加に向け，邦字新聞の全文自動テキスト化が求められている．現在，光学文字認識(Optical Character Recognition，OCR)を用いたテキスト化が進められているが，その成果は未だ十分ではない．邦字新聞は多段組みの構成であり，見出しや図，広告など本文以外の要素も含まれる複雑なレイアウト構成をとる．そのため，一般の書籍に対応する既存の OCR システムでは対応できず，レイアウト解析精度が低い．また，読み順が考慮されずに出力されるため，文字認識の結果が文章として成り立たない場合がある．邦字新聞は，1 ページが様々な主題の記事により構成される．記事ごとにレイアウト構成が異なることから，本文の読み順を簡単なルールベースで決定できない．テキスト化が正しく行えていても，正しい読み順で出力できなければ，研究や全文検索機能に向けた活用の際に不便となる．以上のような問題から，既存の OCR システムにより邦字新聞のテキスト化を十分な精度で行うことは困難であり，邦字新聞に対応した OCR システムの開発が求められる．

日本語文書に対応した OCR システムの 1 つに NDLOCR[2]がある．NDLOCR は，国立国会図書館の所蔵する約 262 万タイトルの書籍資料画像を利用して開発されたものである．2022 年 4 月に ver.1.0，2023 年 4 月に ver.2.0，2023 年 7 月に最新版の ver.2.1 がオープンソースプログラムとして公開されている．NDLOCR は，国立国会図書館の所蔵する幅広い年代の雑誌，図書により学習データセットが構築されている．対応文字種は，ひらがな，カタカナや JIS 第一，第二水準漢字などの基本的な文字種から，欧文，ギリシア文字などがあり，日本語文書に用いられるほとんどの文字に対応している．NDLOCR は幅広い年代，様々な種類の日本語文書に対して高精度なテキスト化が期待できる OCR システムである．

NDLOCR の処理は，以下の通りである．まず，前処理により傾き補正や見開き分割を行う．次に，レイアウト解析において行領域を抽出する．レイアウト解析の要素には本文や見出し，ルビ，図版などが含まれる．そして，抽出される行領域に対してシーケ

ンス認識[3]の手法により行単位で文字認識を行い，テキストを出力する．NDLOCR の学習に使われる文書画像は，その仕様上，1 段組みから 2 段組み程度の単純な構成をとるものが想定されている．邦字新聞のような多段組みかつ，段組み内に図や広告が挟まる複雑なレイアウト構成には当然対応できず，レイアウト解析精度が不十分である．また，邦字新聞の活版印刷による不統一なフォント，サイズにより，NDLOCR は邦字新聞のルビを検出することができない．そのため，ルビの含まれる行において，ルビがノイズとなることにより文字認識精度が低下すると推測される．

NDLOCR ver.2.0 以降では，視覚障がい者用に向けた読み上げ用途にも活用できるよう，読み上げ順序の整序処理が追加されている．余白などの文書における構造的特徴を利用し，再帰的な XY Cut を行うことで読み順を決定する．しかし，邦字新聞で見られる多段組み，複数の記事による構成，文章の間に広告や図が挟まる複雑な構成に対応できない．よって，出力されるテキストが文章として成り立たない場合がある．

本修士論文では，以上のような邦字新聞に対する NDLOCR の課題点を踏まえ，NDLOCR に対して一部手法の変更や追加を行う．次に述べる 3 つの改善を適用して邦字新聞に対応した OCR システムとして提案する．

1 つ目に，レイアウト解析手法の変更を行う．NDLOCR のレイアウト解析手法では，邦字新聞における多段組みの文書に対して精度が不十分である．多段組みかつ多サイズ文字が含まれる近代書籍に対応したレイアウト解析手法として，解像度ピラミッドを適用した CRAFT の手法[4]が提案されている．提案手法では，NDLOCR のレイアウト解析手法を変更し，解像度ピラミッドを適用した CRAFT の手法に置き換える．

2 つ目は，ルビ除去処理の追加である．ルビ除去を行うためにフーリエ記述子と濃度ヒストグラムを用いる．フーリエ記述子により行領域内の文字に対して輪郭の概形を抽出し，縦方向に濃度ヒストグラムをとる．ヒストグラムの谷となる位置より右側にルビがあると推定されるため，その位置で分離する．ヒストグラムの谷がない場合には，ルビが含まれない行であると判定し，ルビ除去は行わない．ノイズとなるルビが除去されることにより，文字認識精度の向上が期待できる．

3 つ目は，読み順検出手法の変更である．邦字新聞は，1 ページが複数の記事で構成される．1 つの記事はいくつかの文章ブロックにより構成されるが，同一記事であっても文章ブロックが分散的に配置されることにより，読み順検出を困難にしている．そこで，読み順検出のために文章内容を考慮した同一記事推定を行う．同一記事を構成する文章ブロック同士は，記事の主題に関連する単語が出現しやすいと推測される．Okapi-BM25[5]による単語の重要度計算と，コサイン類似度計算を利用し，各文章ブロック同士の類似度を算出し，類似度が高いものを同一記事と判定する．

提案手法は，スタンフォード大学フーバー研究所上田薫教授のもとで，実用化に向けた運用が求められている．一般に，マシンや OS が異なる環境下において正しくシステムを動作させるには，多数のライブラリファイルの管理，設定ファイルの配置など，複

雑な作業が必要となる. そこで, 提案手法ではコンテナ型仮想化ソフトウェアの Docker を利用した環境構築を行う．システムの環境をコンテナとしてパッケージ化し，共有することで異なるマシン間でシステムの移植，再現が容易になる．

本修士論文の構成は以下の通りである．第 2 章では，既存手法である NDLOCR について述べる．NDLOCR の手法について詳細に説明し，邦字新聞に対する課題点を述べる．第 3 章では，提案手法におけるレイアウト解析手法，ルビ除去手法，読み順検出手法について説明する．また，Docker を用いた構築環境を述べる．第 4 章では，提案手法の有効性を検証し，結果を述べる．第 5 章では，検証結果を踏まえて考察を述べる．

第2章　NDLOCR

既存手法である NDLOCR[2]は，国立国会図書館が所蔵するデジタル資料の全文テキストデータ作成を目的に開発された OCR システムである．現在，株式会社モルフォ AI ソリューションズにより委託研究開発が行われている．NDLOCR は，国立国会図書館から貸与される約 262 万タイトルの書籍資料画像から構築された学習用データセットが使用されており，日本語資料に対して高精度な OCR を可能とする．しかし，邦字新聞のような多段組みのレイアウト構成に対応できず，その認識精度は未だ不十分である．NDLOCR が邦字新聞に対応し，十分な精度を得るにはいくつかの課題点がある．

本章の構成は以下の通りである．2.1 節では，NDLOCR で用いられる手法について述べる．2.2 節では，NDLOCR が邦字新聞に対応するための課題点について述べる．

2.1　NDLOCR の手法

NDLOCR は，国立国会図書館からの委託により，株式会社モルフォ AI ソリューションズが開発する OCR システムである．2022 年 4 月に ver.1.0 が GitHub にて公開されている[2]．2023 年 4 月には ver.2.0，2023 年 7 月には ver.2.1 が公開され，視覚障がい者用に向けた読み上げ用途にも活用できるよう，読み上げ順序の整序機能が追加されている．また，NDLOCR ver.2.1 では，学習データに 1960 年代以前の資料を含めることにより，文字認識精度の向上が報告されている[6]．

NDLOCR の学習用データセットは，国立国会図書館により貸与された約 262 万点の資料画像データを利用して作成されている．資料画像データから，学習用データセットとして利用できる画像の抽出，選別作業が行われる．学習用の画像は，テキストと図版入り画像，テキストのみの画像，全面外国語画像，図表入り画像などが選別される．画質の悪い画像や，図版のみの画像，楽譜などは学習用の画像から除外される．選別された画像に対して行矩形領域情報や本文文字情報などのアノテーションを行い，学習および評価用データセットが作成される．2023 年 1 月 26 日までに，10,803 件の学習用データセット，3,013 件の評価用データセットの納品完了が報告されている．

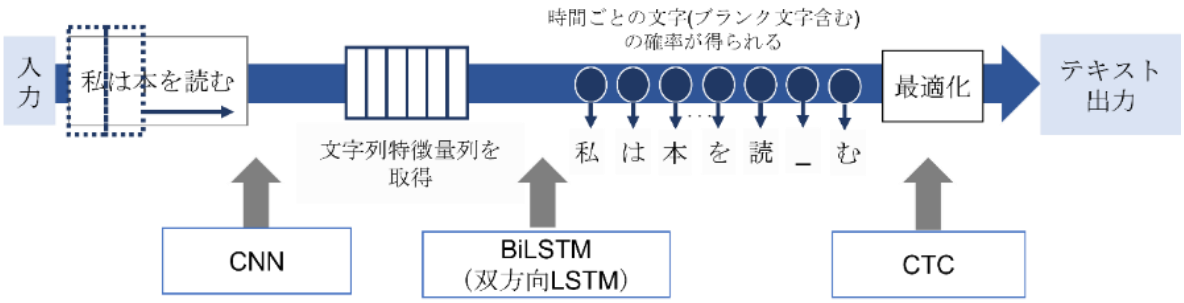
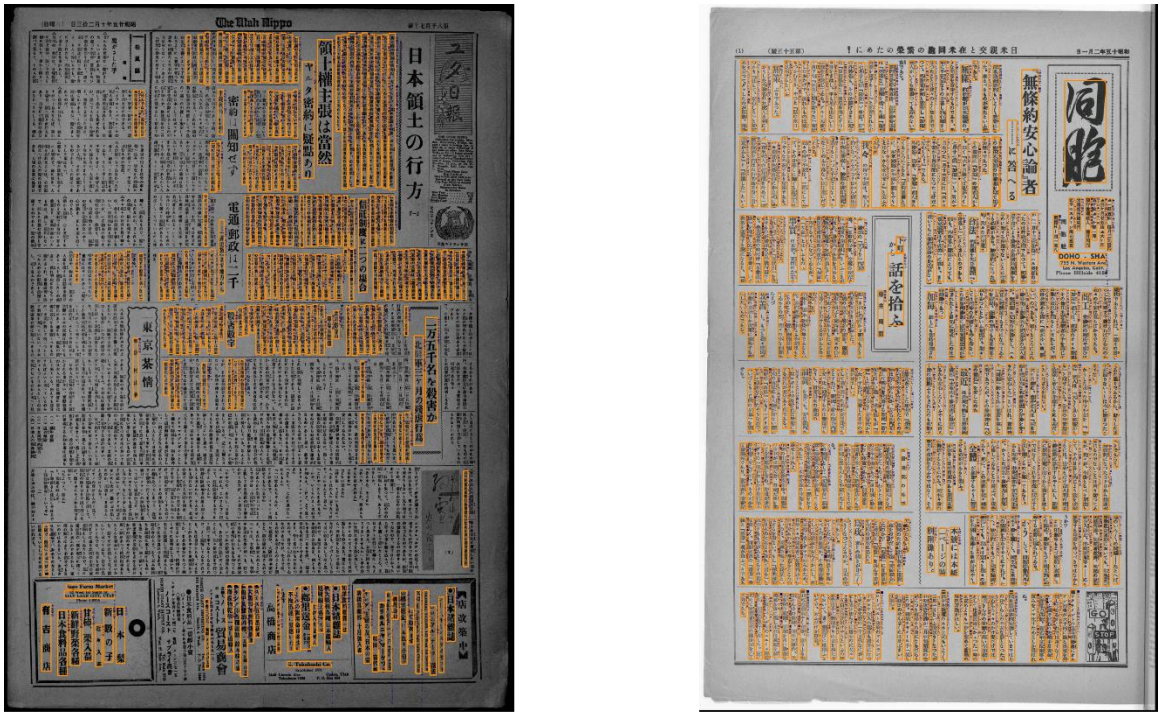


図 2.1 シーケンス認識の処理の流れ

NDLOCR における処理の流れは以下の通りである．まず，画像の前処理を行う．前処理では入力画像に対して見開きの分割, ハフ変換を用いた傾き補正が行われる. 次に，レイアウト解析を行う．レイアウト要素は，図版，表組，ルビ，数式などが含まれるブロック要素と，本文，見出し，キャプション，広告文字が含まれるライン要素により構成される．レイアウト解析の手法には，物体検出手法である Cascade Mask RCNN が検出器として利用される．バックボーンは，ver.1.0 で Resnet[7]が採用されているが，速度劣化を抑えて精度向上が期待できることから，ver.2.1 では ConvNeXt[8]が採用されている．そして，レイアウト解析により得られる行領域に対して文字認識を行う．文字認識の対応文字種は，ひらがな，カタカナ，数字のほか，JIS 第一，第二水準漢字など基本的な文字種である．その他特殊文字として，繰り返し文字，記号，欧文，ギリシア文字が含まれる. 文字認識手法には 1 行の文字列画像を系列データとして扱うシーケンス認識[3]の手法が利用されている．図 2.1 は，シーケンス認識処理の流れを示している．まず，入力される文字列画像に対し，CNN によって文字列特徴量を得る．そして, 双方向 LSTM により文字列画像に対する時間ごとの文字種確立を得る. このとき，文字種にはブランク文字も含む．最後に，最も確率の高い文字種から得られる予測文字列に対して CTC 関数による最適化が行われ，テキストが出力される．

以上の流れで，NDLOCR は 1 行単位での文字列認識を行う．ver.2.1 に限り，文字認識結果に対して「川」と「三」など縦行を横向きで認識する際に問題となる文字に対応するため，ネットワーク内部の縦，横行推定モジュールが導入されている．

NDLOCR ver.2.0，ver.2.1 では，視覚障がい者に向けた読み上げ用途機能が追加され，以下の後処理が行われる．まず，読み順整序のため，ページ余白を利用した再帰的な XY Cut を行い，文章のひとかたまりとなるブロックを分割する．ブロックごとにルールベースを適用することで読み順を決定する．次に，Random Forest による見出しと著者推定を行う. 最後に, 形態素解析ツールの KyTea により漢字の読み推定を行う．



（a）行領域が正しく認識されない例　　（b）行領域が正しく認識される例

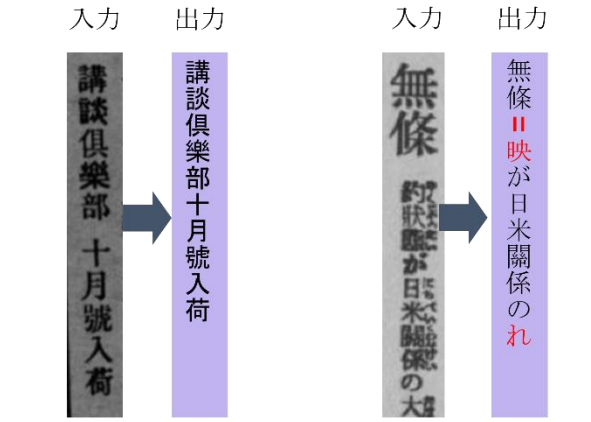
図 2.2 邦字新聞に対する NDLOCR のレイアウト解析結果の例

2.2 邦字新聞に対する NDLOCR の課題点

NDLOCR は，国立国会図書館から貸与された 1870 年代から 1990 年代までの日本語書籍資料画像が学習データセットに利用されている．対応文字種はひらがな，カタカナをはじめ，JIS 第一，第二水準漢字など日本語資料に用いられるほとんどの文字種を網羅している．よって，様々な日本語資料に対して高精度なテキスト化が可能である．

邦字新聞は，明治維新前後より刊行された海外日系新聞の集合体である．邦字新聞の特徴として，多段組みのレイアウト構成が挙げられる．一般的な図書は 1 段から 2 段組みの簡単な構成をとることが多く，NDLOCR では邦字新聞のようなレイアウト構成に委託上の仕様によって対応できない．その他，見出しや図，広告といった様々なレイアウト要素を含むことや，活版印刷文字による不統一なフォントやサイズなど，邦字新聞は一般の図書と異なる文書特徴を持つ． NDLOCR が邦字新聞に対応し，十分な精度を得るには，3 つの課題点がある．

以下より 3 つの課題点について述べる．1 つ目の課題は，邦字新聞に対するレイアウト解析精度の低さである．図 2.2 は， NDLOCR ver.2.1 に邦字新聞を入力したときのレイアウト解析結果を可視化したものである．橙色の枠が，レイアウト解析により認識された行領域を示している．図 2.2 (a)は，行領域が正しく認識されない例である．段



(a) ルビの含まれない行　　(b) ルビの含まれる行

図 2.3 ルビの含まれない行，含まれる行による認識精度の違い

組み数の多い邦字新聞画像に対するレイアウト解析結果は，本文中のほとんどの部分が認識できていない．図 2.2 (b)は，行領域が正しく認識されている例である．段組み数が比較的多いものであっても，解像度が高くノイズの少ない画像であれば，高精度に認識が可能である．しかし，邦字新聞デジタル・コレクション[1]にて提供される邦字新聞画像は，損傷や保存状態が原因でノイズが多く，画像の解像度が高いものは少ない．邦字新聞に対応するには，レイアウト解析手法の改善が必要である．

2 つ目の課題は，ルビが含まれる行に対する文字認識精度の低さである．邦字新聞では，ほとんどの漢字にルビが振られている．レイアウト解析において抽出される文字領域内にルビが含まれると，ルビがノイズとなり，文字認識精度が低下することが報告されている[10][11]．図 2.3 は，NDLOCR ver.2.1 において，抽出行矩形内にルビが含まれる場合と含まれない場合の文字認識精度の違いを示している．図 2.3 (a)，(b)のそれぞれ左側に抽出された行領域，右側に文字認識結果を示している．文字認識結果に対し，誤りである文字を赤色の文字で示している．図 2.3 (a)は，漢字にルビが振られておらず，正しく文字認識が行えている．これに対し，図 2.3 (b)は，行領域内にルビが含まれており，ルビの振られている漢字の文字認識誤りが多く見られる．NDLOCR はレイアウト解析における要素の 1 つにルビが含まれており，本文とルビを識別できる．しかし，邦字新聞は活版印刷により，ルビの位置や文字サイズが規格化されていない．そのため，NDLOCR では邦字新聞のルビを識別することが難しいと推測される．邦字新聞に対応するルビ除去処理を新たに追加する必要がある．

3 つ目の課題は，読み順検出精度である．NDLOCR ver.2.1 では，再帰的な XY Cut により文章のひとかたまりをいくつかのブロックに分け，ルールベースで読み順が決定される．しかし，邦字新聞は多段組みかつ，1 ページに複数の記事が含まれる複雑な構

成を持つ. 同一記事を構成する文章ブロックであっても, 間に図や広告が挟まることで, 分散的に配置される場合がある．これにより，人間が読んでいても順序を決定することが難しい場合がある．同一記事であることは，構造的な観点のみで決定することが困難である．邦字新聞に対応するため，文章内容を考慮した読み順検出手法が必要である．

NDLOCR は幅広い発行年代の日本語文書に対して高精度に認識が可能な OCR システムである．NDLOCR ver.2.1 は，国立国会図書館により貸与された年代別評価データセットによる文字認識性能評価結果において，平均 96.57%の精度を達成していることが報告されている[6]．幅広い日本語文書に対応する OCR システムとして，文字認識精度が非常に高い結果である．NDLOCR に対し，以上に述べた 3 つの課題点を解決することで，邦字新聞に対して十分な認識精度を得ることが期待できる．

第3章　邦字新聞に対応したOCR

邦字新聞は，スタンフォード大学フーバー研究所にて画像データが収集され，邦字新聞デジタル・コレクション[1]として公開されている．邦字新聞は，政治状況やコミュニティ，軍事プロパガンダなど，当時の情報を知るための重要な資料である．現在，OCRシステムによる邦字新聞画像データの全文自動テキスト化が進められている．しかし，一般の文書と異なる様々な特徴を持つ邦字新聞に対して，既存のOCRシステムでは精度が不十分である．よって，邦字新聞に対応可能なOCRシステムが求められている．2.2節では，日本語文書に対応する既存システムであるNDLOCRにおいて，邦字新聞に対応するための課題点を述べた．これらを踏まえ，レイアウト解析，ルビ除去，読み順検出手法の改善を行い，邦字新聞に対応したOCRシステムとして提案する．

本章の構成は以下の通りである．3.1節では，レイアウト解析手法について述べる．3.2節では，ルビ除去手法について述べる．3.3節では，読み順検出手法について述べる．3.4節では，提案手法における処理の流れ，環境構築手法について述べる．

3.1　レイアウト解析手法

3.1.1. 解像度ピラミッドを適用したCRAFT

CRAFT（Character-Region Awareness For Text detection）[9]は，風景画像からテキスト領域を検出するための手法である．CRAFTでは，入力される画像データに対し，CNNによって文字ごとのヒートマップを2種類出力する．1つは，文字の中心部分の予測スコアのRegion Scoreマップである．もう1つは，文字の連結部分の予測スコアのAffinity Scoreマップである．得られた2種類のヒートマップに対して後処理による加算，ラベリング処理などを行うことで，画像から文字領域を抽出できる．

本研究室では多段組構成を持つ近代書籍を対象としたレイアウト解析手法として，解像度ピラミッドを適用したCRAFTの手法[4]を報告している．多段組みかつ様々なレイアウト，文字サイズに対応するCRAFT用モデルを学習するには，学習データに解像度が高い画像データが大量に必要となる．しかし，そのようなデータを扱うためには膨大な計算資源が必要である．そこで，CRAFTと解像度ピラミッドを組み合わせ，不足

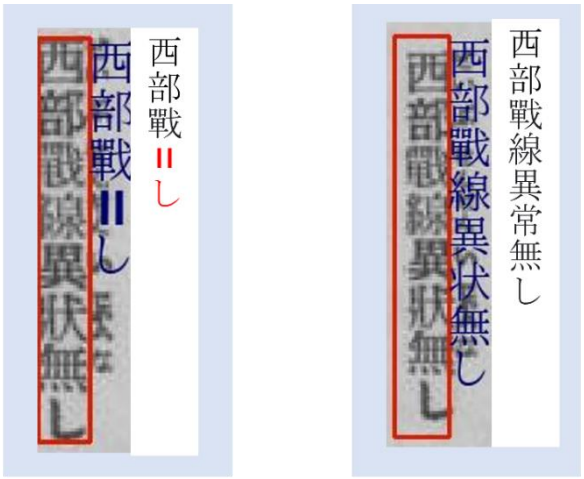


図 3.1 行領域の拡大処理の有無による文字認識結果比較

する計算資源の状況下でも CRAFT の学習を可能としている．多段組み，多サイズ文字が用いられるレイアウト構成に対応した学習モデルが構築されており，同様のレイアウト特徴を持つ邦字新聞に対応可能なレイアウト解析手法である．提案手法では解像度ピラミッドを適用した CRAFT の手法を採用し，行領域の抽出を行う．

3.1.2. 行領域抽出

提案手法による行領域の抽出手法について述べる．まず，入力画像に対して前処理として 2 値化と段組みを区切る横線の除去を行う．そして，前処理画像に対し，3.1.1 項で述べる解像度ピラミッドを適用した CRAFT を用いて 2 種類のヒートマップを得る．

2 種類のヒートマップに対して，以下の後処理を行うことにより行領域を抽出する．まず，2 つのヒートマップに対して閾値処理を行うことで，2 値化画像を得る．2 値化手法には大津の手法を用いる．次に，得られた 2 つの 2 値化画像を加算する．そして，加算画像に対して連結成分のラベリング処理を行う．最後に，それぞれの連結成分に対して外接する長方形領域を取得することにより，行領域を抽出する．

以上の処理により得られる行領域は，文字の端々が切れてしまう場合がある．そこで，得られる行領域に対して，上下方向と左方向に拡大処理を行う．右方向の拡大は，3.2 節で述べるルビ除去を施すため行わない．図 3.1 は，行領域に対して拡大処理を行う前と後の文字認識結果の違いを示している．左側は行領域抽出結果であり，赤色の枠線で示している．右側は文字認識結果であり，赤色の文字は文字認識結果が誤りであることを示している．拡大処理を行わない図 3.1 (a)に対して，図 3.1 (b)のように上下と左方向に拡大処理を行うことで，文字認識結果が改善することが確認できる．

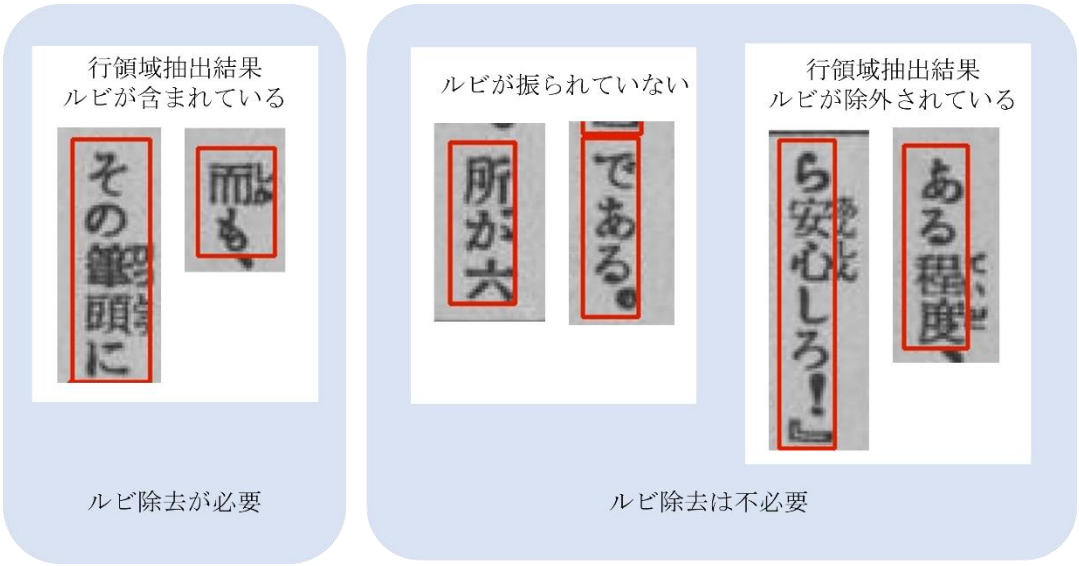


図 3.2 ルビ除去の必要があるものと不必要なものの違い

3.2 ルビ除去手法

邦字新聞の本文内に使われる漢字のほとんどには，ルビが振られている．邦字新聞において漢字にルビが振られる理由の 1 つに，当時の識字率の低さが挙げられる．邦字新聞は，明治から昭和初期までの約 100 年間，アメリカ大陸，アジアにおける日本人移民，またはその第二世代以降により刊行された日系新聞である．第二世代以降の教育に関わる人材，教材不足に伴う識字率の低さのため，ルビが必要であったと考えられる．

邦字新聞には，ルビが振られている行と，そうでない行が混在する．また，3.1 節で述べる行領域抽出手法の結果，ルビが行領域から除外されるものと，ルビが行領域として本文と一緒に含まれるものが混在する．行領域内にルビが含まれる状態で文字認識処理を行う場合，ルビがノイズとなり文字認識精度が低下することが報告されている [10][11]．よって，レイアウト解析の結果行領域内にルビが含まれる場合には，文字認識処理を行う前にルビ除去を行う必要がある．

レイアウト解析により得られる行領域には，ルビが含まれるものと含まれないものが混在する．ルビ除去を行う前に，ルビの有無判定を行う必要がある．図 3.2 は，ルビ除去が必要なものと，そうでないものの違いを示している．図の右側に示すように，赤色の枠で囲まれた行領域にルビが含まれていない場合，ルビ除去の対象から除外する．図の左側に示すようなルビ除去が必要な行領域に対して，ルビの位置を推定し除去を行う．

ルビ除去手法として，濃度ヒストグラムを用いる方法[12]がある．行領域の文字部分と背景部分を 2 値化したものに対して，縦方向に濃度ヒストグラムをとる．濃度ヒスト

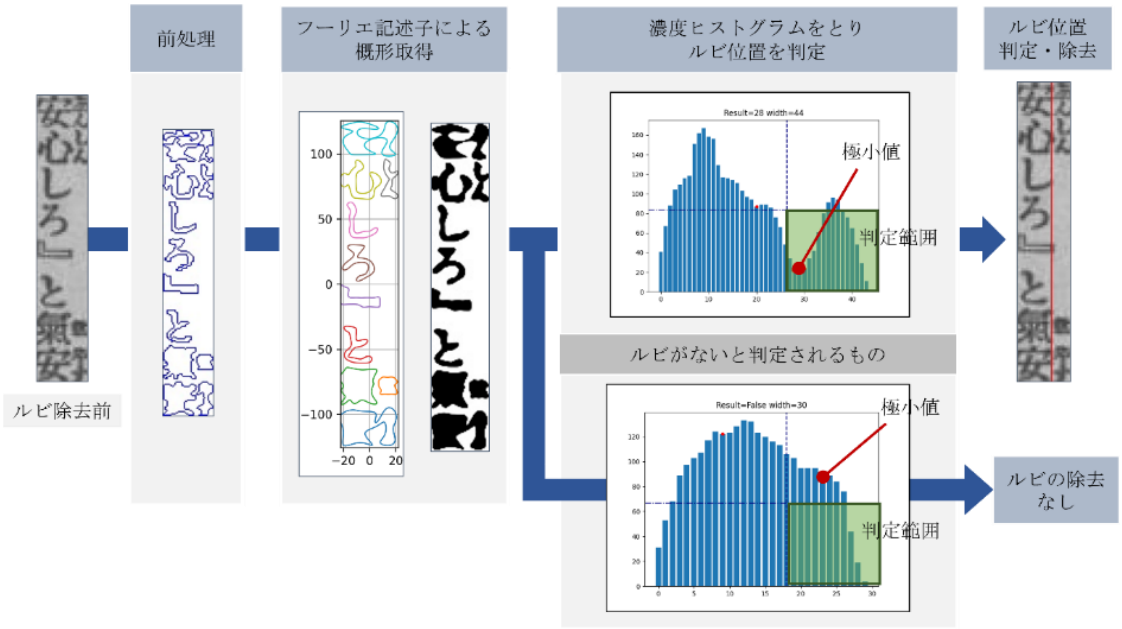


図 3.3 ルビ除去手法

グラムの谷となる部分をルビ位置として，直線的に分離する．濃度ヒストグラムの谷となる部分が存在しない場合に，ルビが行領域内に含まれないと判断できる．

行領域の文字部分と背景部分を分けるため，文字部分の輪郭を取得する．画像の輪郭線を定量的に記述する方法として，楕円フーリエ記述子を用いる手法が提案されている[13]．楕円フーリエ記述子は，閉曲線に対する座標情報を周期関数として捉え，フーリエ級数展開により得られるフーリエ係数から形状を近似するものである．

提案手法では，フーリエ記述子を用いて文字部分の概形を取得し，縦方向に濃度ヒストグラムをとることで，ルビの有無判定とルビ除去を行う．図 3.3 は，ルビ除去手法における処理の流れを示している．まず，行領域画像に対して前処理として 2 値化を行い，連続する領域に対して Canny 法[14]により輪郭点を取得する．次に，楕円フーリエ記述子により輪郭線の近似を行う．フーリエ記述子の展開次数は大きいほど輪郭を微細に記述することができる．提案手法では，最大次数$N = 15$として輪郭の概形を得る．そして，得られた輪郭に対して縦方向に濃度ヒストグラムをとる．濃度ヒストグラムに谷の部分があるかどうかを，ヒストグラムの極小値の有無と位置により判断する．提案手法では，図 3.3 に示す濃度ヒストグラムの判定範囲内に極小値が含まれるとき，ルビが含まれると判定する．具体的な判定範囲は，極小値が縦方向最大濃度値の1/2以下かつ，極小値をとるインデックスが行領域の右側2/5の範囲である．最後に，ルビが含まれると判定されたものに対して，極小値のインデックス位置より右側を行領域内から削除する．以上のようにして，行領域に対してルビの有無判定とルビ除去を行う．

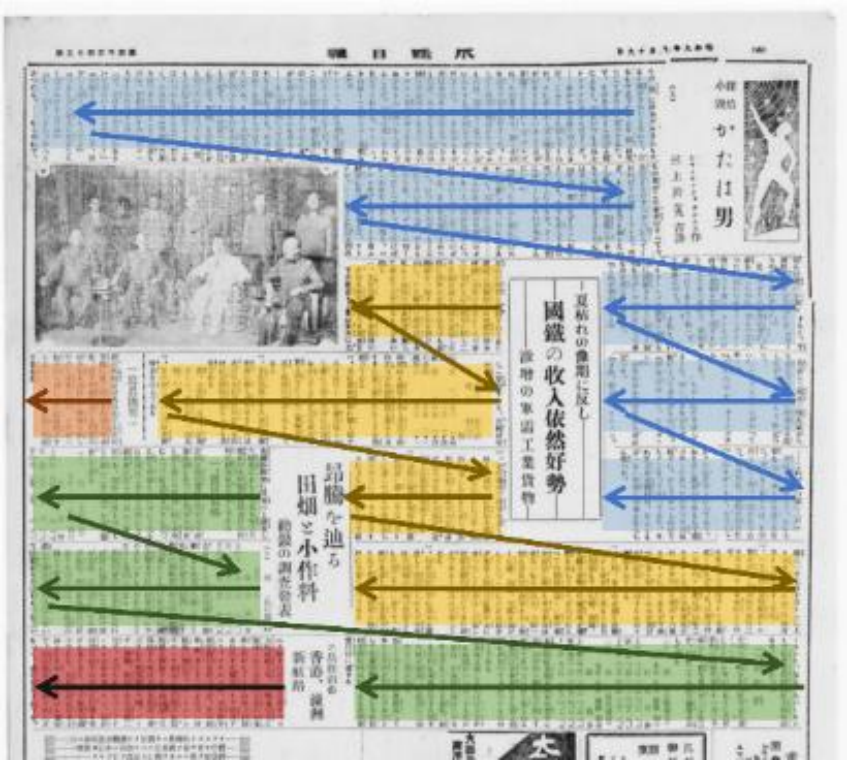


図 3.4 複雑なレイアウト構造で読み順が定めにくい邦字新聞の例

3.3 読み順検出

文書画像に対する OCR システムの主な出力は，レイアウト解析処理によって得られる文字や行，図や写真などの位置を示すレイアウト情報と，文字認識処理によって得られるテキスト情報の 2 つである． 1 段組みから 2 段組み程度の単純なレイアウト構成であれば，OCR システムにより得られる 2 つの出力を利用して，簡単なルールベースを用いた読み順通りの出力が可能である．

邦字新聞のレイアウト構成は多段組みであり，1 ページ内に様々な主題の記事が複数存在する．図 3.4 は，複数記事が含まれることで複雑なレイアウト構造となり，読み順を容易に定められない邦字新聞の例である．余白や線で区切られるある程度の文章のかたまりを，文章ブロックと呼ぶ．同一記事である文章ブロックを，同じ色で示している．記事ごとに文章ブロックの読み順を矢印で示している．図 3.4 のように，同一記事であっても，文章ブロックが分散的に配置される場合がある．このような場合，読み順を構造的観点のみによる簡単なルールベースで定めることは困難である．

OCR システムによる出力が読み順通りでない場合，出力されるテキストが文章として成り立たないという問題が生じる．邦字新聞のテキスト化により，邦字新聞の発行当時に関する研究に向けた活用，邦字新聞デジタル・コレクション[1]における全文検索機

能の追加，近代文語体と現代口語体の相互翻訳手法[15]に向けた活用が期待される．しかし，文章の出力が読み順通りでない場合，これらに対する活用において不便となる．また，OCR システムの後処理として，近代書籍に対応した誤字検出の手法[16]を利用することで更なる精度向上が期待できる．誤字検出手法の利用には，文脈通りの正しい読み順で出力される必要がある．

以上の理由から，邦字新聞に対応する OCR システムには，読み順検出が必要である．読み順を検出するため，提案手法では同一記事に含まれる文章ブロックを推定する．1 つの記事を構成する文章ブロックがページ内に分散配置される構造は，人間が読む場合でも同一記事を決定することが難しい場合がある．よって，同一記事の推定は余白などを利用した構造的な観点で決定するのではなく，文章の内容を考慮する必要がある．同一記事の文章ブロックには，記事の主題に関連する同様の単語が使われることが多い傾向がある．そこで，提案手法では文章ブロックを分割した後，各文章ブロック内容の類似度を求めることにより，同一記事推定を行う．

文章ブロックを分割するため，3.1 節で述べるレイアウト解析手法により得られる Affinity Score マップを用いる．ブロック分割手法は以下の通りである．まず，Affinity Score マップに対して適応的閾値処理によって 2 値画像を得る．次に，2 値画像に対し，モルフォロジー演算のオープニング処理によりノイズ除去を行う．オープニング処理は，2 値画像に対して縮小処理と拡大処理を行うものである．次に，モルフォロジー演算により拡大処理を行う．これは，Affinity Score マップが，文字の連結部分を示しているものであることから，縦方向に拡大することで同一行の文字を連結するためである．さらに，横方向の拡大により，隣接する行を連結する．最後に，それぞれの連結部分に対し，外接する長方形領域を抽出することにより文章ブロックの分割を行う．

図 3.5 は，文章ブロック分割による Affinity Score マップの拡大処理後の出力例である．図 3.5 に示す入力画像に対して，解像度ピラミッドを適用した CRAFT[4]により中央に示す Affinity Score マップの出力が得られる．そして，文章ブロック分割処理における拡大処理により，右側に示す画像が得られる．Affinity Score マップに縦，横方向の拡大処理などを行うことで，余白などで区切られる文章ブロックが抽出される．

同一記事推定を行うため，文章ブロック同士における文章内容の類似度を計算する．文書の類似度を求める代表的な方法に，TF-IDF とコサイン類似度計算を用いる方法がある．TF-IDF は，文書に含まれる各単語の重要度を評価する手法である．TF-IDF 値は，ある文書集合に対し，それぞれの文書内に含まれる各単語の出現頻度を表す TF（Term Frequency）値と，ある単語が文書集合中にどれだけ含まれているかを表す IDF

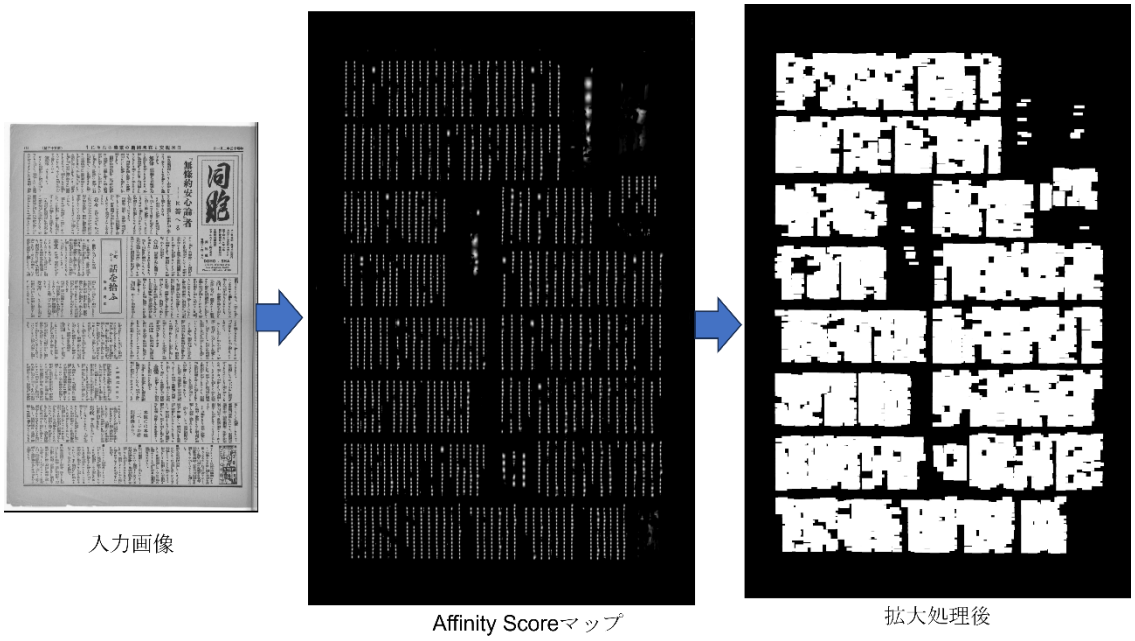


図 3.5 文章ブロック分割における拡大処理出力例

（Inverse Document Frequency）値を掛け合わせることで算出される．IDF 値は，文書集合に対し，ある単語が含まれる文書の割合の逆数をとることで算出される．それぞれの文書における TF-IDF 値に対し，コサイン類似度を求めることで，1 に近いほど文書同士の類似度が高く，-1 に近いほど低いと推定される．ただし，TF-IDF を用いた文書の類似度計算手法は，文書が長いほど単語の出現頻度を表す TF 値が大きくなりやすく，それぞれの文書の長さや単語数を考量する必要がある．

TF-IDF の改善手法として，Okapi-BM25[5]がある．TF-IDF の算出に用いる TF 値と IDF 値に加え，文章内に含まれる総単語数を表す DL（Document Length）値を用いる．文書全体の DL 値の平均値が，単語の重要度計算に含まれる．平均値よりも DL 値の多い文書に含まれる単語の重要度が低く算出される．これにより，TF-IDF 手法の欠点である文書の長さの差による算出値に対する影響が軽減される．TF-IDF 値の時と同様にして，Okapi-BM25 による単語の重要度算出結果に対してコサイン類似度を求めることにより，類似性の高い文書を推定することができる．

類似度計算を用いた同一記事推定による，読み順検出処理の流れを図 3.6 に示す．まず，レイアウト解析結果を利用し，邦字新聞の本文領域に対してブロック分割を行う．次に，文字認識の結果を利用し，各文章ブロックに対して，形態素解析エンジンの

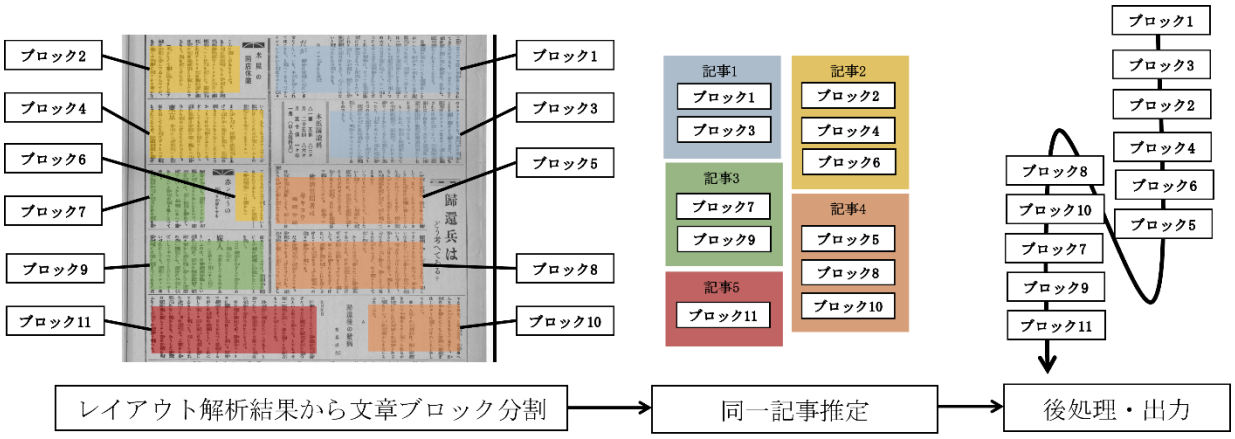


図 3.6 読み順検出処理の流れ

MeCab[17]を用いて文章内の名詞を抽出する．そして，Okapi-BM25 とコサイン類似度計算により，文章ブロック同士の類似度を算出し，類似度が高いもの同士を同一記事とする．最後に，同一記事であると推定される各文章ブロックを，上から下，右から左に配置される文章ブロックの順に並べ替え，テキストを出力する．

3.4 邦字新聞に対応した OCR の実装

3.4.1. システムの処理の流れ

提案手法の処理の流れを述べる．まず，画像の入力に対し，前処理として適応的閾値処理による 2 値化と，段を区切る横方向の線の除去を行う．画像から直線を検出する手法として，漸進的確率的ハフ変換による手法 [18] があるが，画像内の線の長さや太さを指定しながら検出する必要がある．邦字新聞において除去の対象となる横方向の線は，様々な太さや長さがあり，ゆがみが多い．よって，直線を検出する漸進的確率的ハフ変換による手法は不適切である．提案手法では，モルフォロジー演算により横方向の線の除去を行う．縦方向のサイズを 1 としたカーネルを用いて，画像の収縮と膨張処理を繰り返す手法である．様々な長さやゆがみのある線であっても，線の検出ができる．

次に，レイアウト解析により行領域を取得する．3.1 節で述べる解像度ピラミッドを適用した CRAFT[4]により得られる Region Score マップと Affinity Score マップに後処理を施すことにより，行領域を得る．そして，得られる各行領域画像に対して 3.2 節で述べるフーリエ記述子と濃度ヒストグラムを用いたルビ除去処理を行う．以上までのレイアウト解析の処理により得られる行領域に対して，NDLOCR と同様にして，シーケンス認識手法[3]により文字認識を行う．

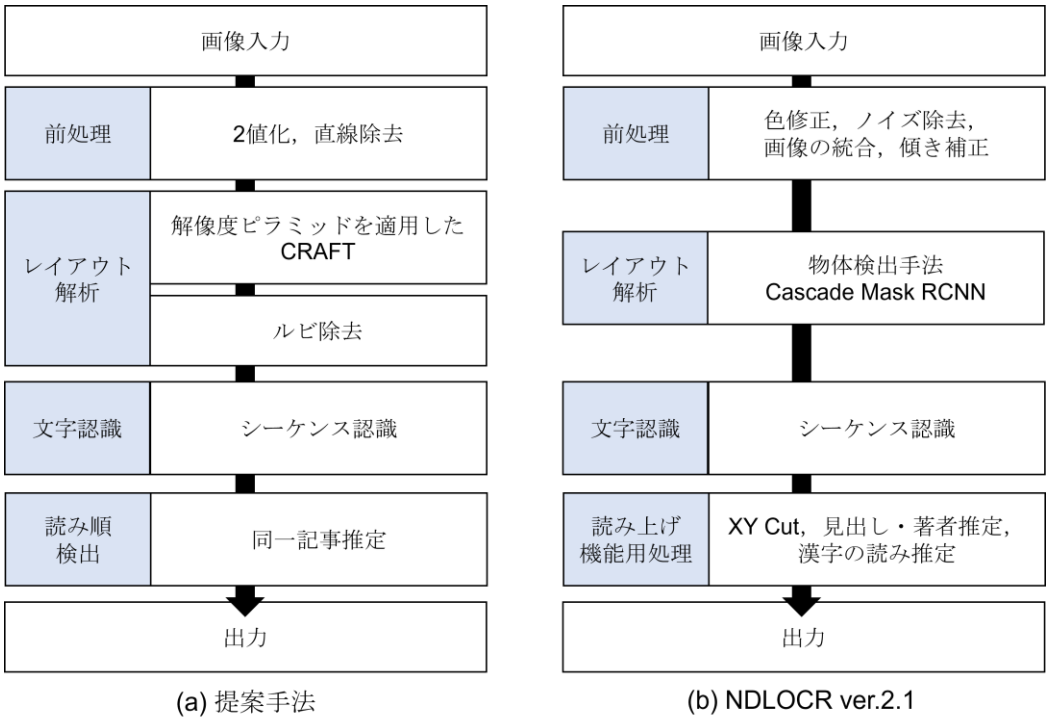


図 3.7 提案手法と NDLOCR ver.2.1 の処理の比較

最後に，読み順検出を行う．NDLOCR ver.2.1 では，読み上げ機能用の処理として XY Cut によって画像を構造的に分割することで読み順を整序する処理が含まれる．提案手法は，文章内容を考慮した同一記事推定手法により読み順を検出するため，3.3 節で述べる文章の類似度計算を用いる．図 3.7 は，提案手法と NDLOCR ver.2.1 におけるそれぞれの処理の流れと，手法の違いを示している．提案手法は邦字新聞に特化するため，NDLOCR と異なる処理を行う．

3.4.2. Docker による環境構築

提案手法による OCR システムは，現在 Linux 環境の下で研究，開発が行われている．今後，スタンフォード大学フーバー研究所上田薫教授のもとで，邦字新聞に対応する OCR システムの運用が求められている．開発されるシステムを別マシンで正常に動作させるには，多数のライブラリファイルの管理，バージョンによる依存関係の管理，設定ファイルの配置など，複雑な手順を踏む必要がある．

そこで，コンテナ型仮想化ソフトウェアの Docker[19]による環境の構築を行う．コンテナ技術とは，アプリケーションの実行に必要なライブラリやツールをまとめて分離させ，1 つの OS 環境下で複数の実行環境を構築するための技術である．Docker を利

表 3-1 動作確認済みの環境

OS	Ubuntu 22.04.3	Windows 11 22H2
GPU	NVIDIA AD102 [GeForce RTX 4090]	NVIDIA TU106 [GeForce RTX 2060 SUPER]
NVIDIA Driver	525.125.06	535.129.03
CUDA Toolkit	11.8.89	11.7.0

用することで，開発環境をコンテナとしてパッケージ化できる．また，機械学習を行うGPU マシンの開発環境においては，NVIDIA Container Toolkit[20]を用いることで，Docker 環境内で GPU を自動検出できる．これにより，機械学習を利用するシステムの依存関係の管理が容易となり，再現性，移植性が向上する．

Linux および Windows において，Docker により GPU 実行環境を構築する．表 3-1 は動作確認済みのマシン環境を示している．Linux ディストリビューションの Ubuntu と Windows の 2 つの OS で動作確認を行っている. Windows 環境下の場合, Windows11 または Windows10 ver.1903 以上が必要である．また，NVIDIA 社が提供するグラフィックボード用ドライバの NVIDIA Driver と，NVIDIA Container Toolkit のインストールが必要である．表 3-1 に，それぞれの動作確認済みのバージョンを示している．

開発環境における Docker のコンテナ型仮想化実装は，Linux のコンテナ技術に由来する．Windows 上で Linux ディストリビューションを実行する機能として，Linux 用 Windows サブシステム（Windows Subsystem for Linux ver.2，WSL2）[21]がある．WSL2 をバックエンドとして Docker の環境構築を行うことで, Windows 上でも Linux のコンテナが実行可能となる．動作確認では，Windows 上で Docker 環境を構築および利用するためのツールである Docker Desktop を利用している．WSL2 上で Docker Desktop を実行することで，Linux のコンテナが実行可能であることを確認している．

第4章　邦字新聞 OCR の検証

4.1　レイアウト解析の検証

4.1.1. レイアウト解析手法の有効性検証方法

提案手法におけるレイアウト解析処理に関して，NDLOCR との精度の比較を行い，手法の有効性を検証する. 比較に用いる NDLOCR は現在最新版の ver.2.1 を利用する.

検証用画像として邦字新聞デジタル・コレクション[1]にて提供される邦字新聞画像を用意する．邦字新聞の発行年代，場所により，フォントやレイアウト特徴が様々であるため，それぞれ発行年代や発行場所の異なる邦字新聞画像を 8 枚用いる．折り目やにじみ, 傷などによるノイズが多く, 人間でも読むことが困難なものは対象から除外する.

レイアウト解析精度の評価方法について述べる．提案手法と NDLOCR は，レイアウト解析において行領域を得る．邦字新聞の本文領域において，1 行が正しく切り出されたものを正解の出力とする．1 行の途中で切れてしまい，行領域抽出結果に文字の漏れがあるものは不正解とする．提案手法では，1 行内に 2 つ以上の領域が抽出される場合がある．2 つ以上の領域が抽出される場合でも，文字の漏れなく 1 行をすべて切り出せているものであれば，正解とする．本文内の行の総数のうち，正しく行領域が切り出せているものの割合を算出し，レイアウト解析の認識率として精度を評価する.

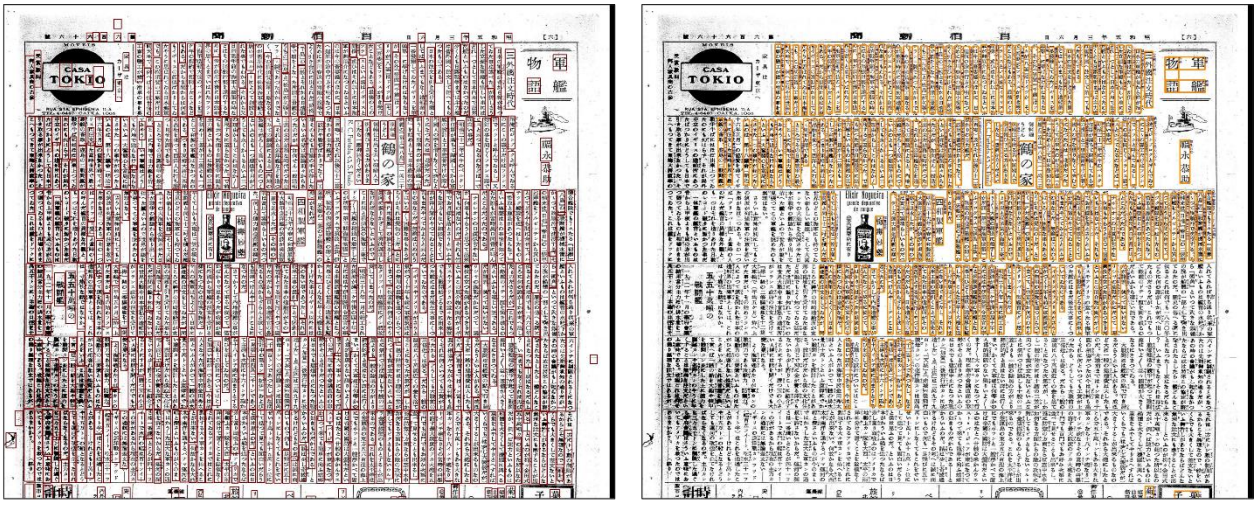
4.1.2. レイアウト解析手法の有効性検証結果

8 枚の邦字新聞画像に対し, 提案手法と NDLOCR によりそれぞれレイアウト解析を行う. 表 4-1 は, 提案手法と NDLOCR のレイアウト解析によるそれぞれの認識率と平均認識率を示している．提案手法によって得られる認識率は最高で 97.69%，最低で 71.73%であり，平均認識率は 87.60%である.

画像 8 枚中 4 枚の画像は，NDLOCR に対して提案手法の認識率が高い．図 4.1 は提案手法と NDLOCR のレイアウト解析結果の例を示している. 行領域として認識された部分を，赤色および橙色の枠で示している．図 4.1 の左側に示す提案手法の認識率は 87.32%であり，右側に示す NDLOCR の認識率は 49.26%である．NDLOCR では半分以上の行を認識できておらず，提案手法との精度の差が大きい.

表 4-1 提案手法と NDLOCR による各画像のレイアウト解析認識率と平均認識率

	画像 A	画像 B	画像 C	画像 D	画像 E	画像 F	画像 G	平均
提案手法	95.81%	88.26%	92.52%	87.32%	79.89%	97.69%	71.73%	87.60%
NDLOCR	96.34%	89.39%	71.02%	49.26%	57.22%	98.46%	29.62%	70.19%



提案手法
認識率 87.32%

NDLOCR ver.2.1
認識率 49.26%

図 4.1 提案手法と NDLOCR のレイアウト解析結果の例

このように，NDLOCR に対して提案手法の認識率がより高い場合，NDLOCR との認識率の差が大きい結果となった．最も認識率の差が大きいもので，42.11%の差がある．提案手法の認識率が NDLOCR に対し低下するものに関しては，最大でも 1.13 %の認識率の差であり，その差は小さい．

一般に，実用される OCR システムにおいて，レイアウト解析精度は 99.9%以上であることが求められる. 実用に向け, 提案手法のレイアウト解析精度の向上が必要である. 提案手法は，NDLOCR によるレイアウト解析の認識率が著しく低いものに対し，大幅な精度向上がみられた．よって，提案手法は邦字新聞に特化した OCR システムとして有効なレイアウト解析手法であることが示された.

4.2　ルビ除去手法の検証

4.2.1. ルビ除去手法の有効性検証方法

提案手法と NDLOCR の文字認識手法は，両者ともに 2.1 節で述べるシーケンス認識手法[3]である．そのため，文字認識手法による精度の変化はないと推測できる.提案手法では文字認識の精度向上のため，ルビ除去処理を行う．ルビ除去手法の有効性を検証するため，レイアウト解析結果からルビ除去を行う提案手法による文字認識の精度と，NDLOCR による文字認識の精度を比較する．

文字認識精度の算出方法を述べる．邦字新聞画像 1 ページの本文領域において，レイアウト解析により 1 行が正しく抽出された行領域の総数に対し，1 行が正しく文字認識できている行領域の割合を算出する．正しく文字認識ができている条件は，文字認識で得られる 1 行すべてのテキストと正解データを照らし合わせ, 誤りなく文字認識ができているものとする．1 行の文字認識の出力結果内に 1 文字以上の誤りがある場合，不正解の出力結果として扱う．ただし，1 行内にかすれやにじみ，ノイズにより判読不能の文字が含まれる場合，その文字の正誤は文字認識の正誤判断には用いない．

4.2.2. ルビ除去手法の有効性検証結果

文字認識率の精度を比較する際，レイアウト解析精度に大きな差があると，認識率の比較が正しく行えないため，レイアウト解析精度が同程度の画像を用いるのが望ましい．文字認識精度の比較には，4.1.2 項のレイアウト解析精度結果から，レイアウト解析精度の差が比較的少ない画像 A，画像 B，画像 F の 3 枚の画像を用いる．

3 枚の画像に対する提案手法と NDLOCR の文字認識精度の比較結果を述べる．表 4-2 は，提案手法と NDLOCR による文字認識率の比較を示している．提案手法において，最も文字認識率の高いものが画像 F の 73.23%，次いで画像 A の 60.10%，最も認識率の低いものが画像 B の 45.57%である．いずれも NDLOCR における文字認識率を上回る結果である．

表 4-2 提案手法と NDLOCR の文字認識精度比較

	画像 A	画像 B	画像 F
提案手法	60.10%	45.57%	73.23%
NDLOCR	42.30%	26.25%	34.77%

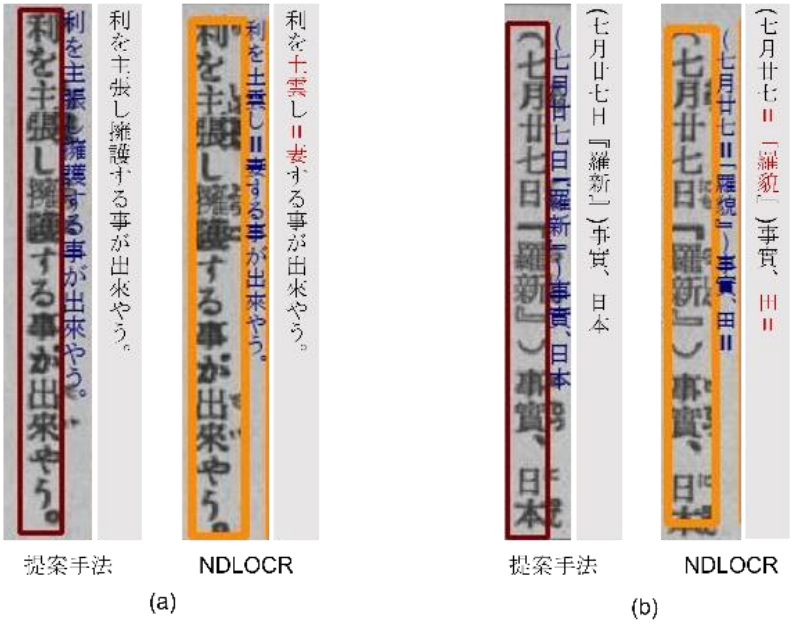


図 4.2 提案手法により文字認識精度が改善する例

図 4.2 は，提案手法においてルビ除去処理が行われることにより，文字認識精度が改善する例を示している．各図の左側は，レイアウト解析により得られる行領域と文字認識の結果を可視化したものである．レイアウト解析により得られる行領域を赤色，橙色の枠で示している．各図の右側は，文字認識結果である．赤色の文字は，文字認識結果が誤りであるものを示している．図 4.2 (a)に示す NDLOCR において文字認識結果が誤りとなる漢字が，提案手法ではルビ除去により改善されている．図 4.2 (b)では，NDLOCR の出力結果に対し，提案手法によるルビ除去処理による改善に加え，拡大処理を行うことによって文字認識結果が改善している．

検証により，NDLOCR に対して提案手法の文字認識精度が向上しており，ルビ除去処理が文字認識精度の改善に有効であることが示された．ただし，ルビが含まれない行に対してルビがあると誤判定し，不必要にルビ除去処理が行われ，文字認識に失敗する例がある．精度の向上のため，ルビの有無判定の改善が求められる．

4.3　読み順検出手法の検証

4.3.1. 同一記事推定手法の有効性検証方法

提案手法では，読み順検出のため同一記事推定を行う．提案手法による同一記事の判定条件について述べる．3.3 節で述べた類似度計算による同一記事推定を行い，類似度が 0.2 以上であるブロック同士を同一記事であると判定する．

同一記事であると判定されるブロックが複数ある場合，重複するブロックすべてを同一記事と判断する．例えば，ブロック A がブロック B，C と同一記事と判定され，ブロック B においてブロック A，D が同一記事であると判定される場合，同一記事であるブロック A，B に対して，それぞれ同一記事であるブロック C，D 同士も類似性が高いと推測し，ブロック A，B，C，D 全てが同一記事であると判定する．

読み順推定手法の有効性を検証するため，邦字新聞の画像を 5 枚用意する．提案手法による同一記事推定手法は，レイアウト解析，文字認識ともに精度の高い出力結果が得られることを前提としている．現段階では，ノイズがあり解像度の低い画像は，レイアウト解析や文字認識の精度が十分でない．検証用の邦字新聞画像には，レイアウト解析において 90%以上の精度が得られる鮮明な画像を用意する．

NDLOCR ver.2.1 では構造的な観点で読み順検出を行っているが，提案手法のように内容的な観点における読み順検出は行っていない．そのため，検証には NDLOCR との比較は行わず，提案手法の精度から手法の有効性を検証する．

4.3.2. 同一記事推定手法の有効性検証結果

同一記事推定手法の有効性検証結果について述べる．図 4.3 は，5 枚の検証用画像に対する同一記事推定結果を可視化したものである．白色の枠で囲まれた部分は正しい同一記事の文章ブロック集合を示している．同一記事であると推定された文章ブロック同士を同じ色で示している．無色の部分は，同一記事の該当なしと推定されるものである．

検証の結果，全ての画像において，同一記事の一部分を抽出できた箇所があり，部分的に同一記事推定ができている．しかし，同一記事内すべてのブロックを網羅して抽出できているものは 1 つもないことが確認できる．

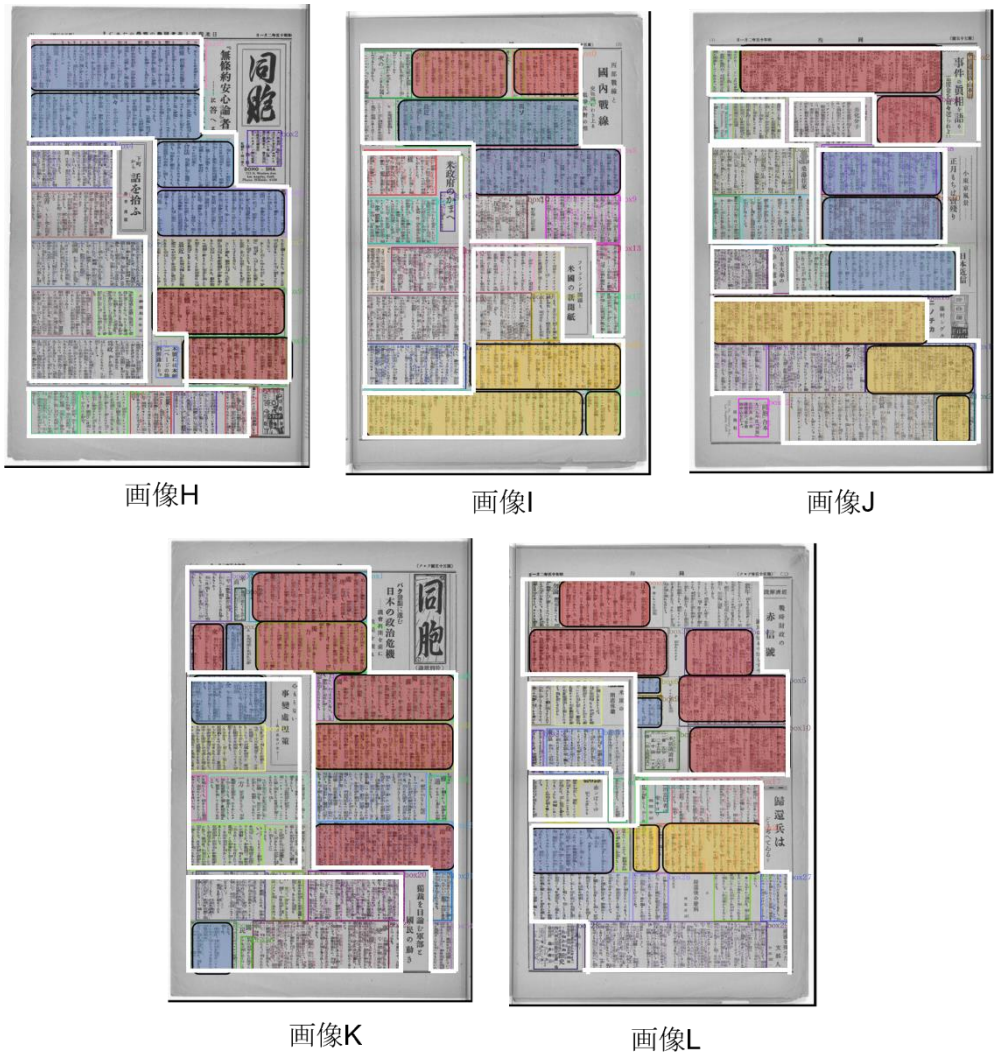


図 4.3 各検証画像における同一記事推定の検証結果

図 4.3 画像 J，画像 K，画像 L には同一記事であると推定されるものに誤りがある．画像 K，L における青色のブロックは，大きさが極端に小さいものがある．ブロックが小さく単語数の少ないブロックで使われる単語は，Okapi-BM25 の特性により重要度が高くなりやすい．他のブロックにおいて青色のブロックと同じ単語が使われた際に影響を与えやすくなり，推定結果に誤りが生じていると推測される．また，画像 J では前処理として行う形態素解析に誤りがあり，誤推定が生じている．

検証の結果，提案手法により推定できるのは部分的な記事の推定であり，現段階では読み順を正しく出力するのが困難であることが判明した．提案手法では，文章内容のみで同一記事を推定する．文書構造的に同一記事ではない事が明らかであるものは最初から候補として外すなど，精度の改善に向けた手法の検討が必要である．

第5章　考察

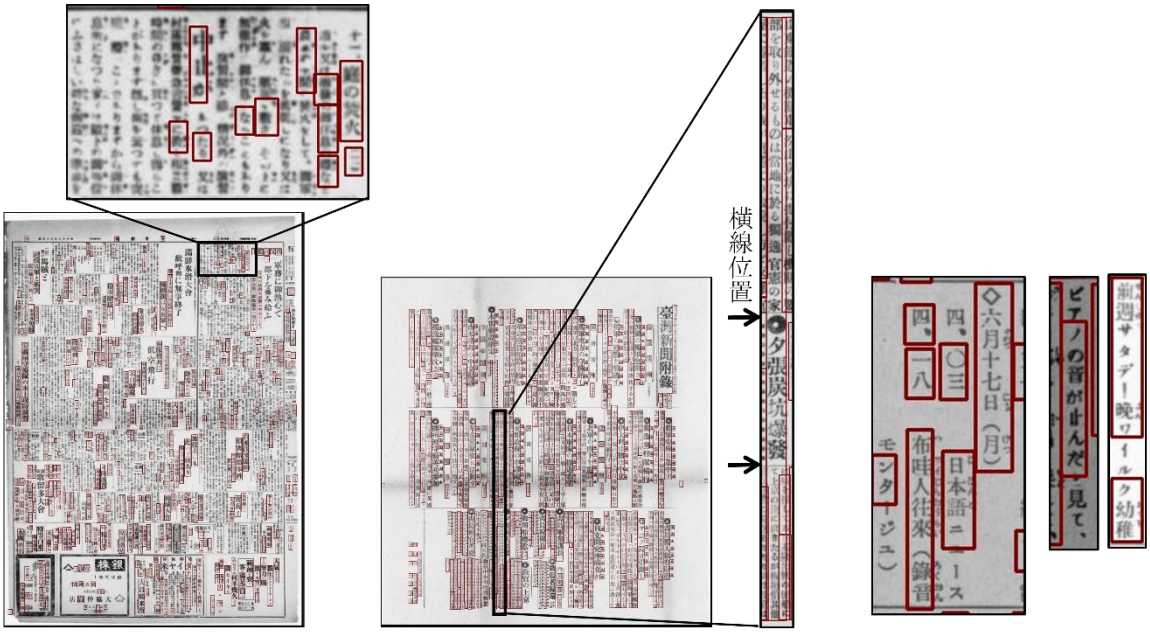
本修士論文では，邦字新聞に対応する OCR システムの手法を提案する．提案手法では, 既存手法である NDLOCR に対して邦字新聞に対応するための 3 つの手法の改善を行う．第 4 章では，改善手法であるレイアウト解析，ルビ除去，読み順検出の手法について有効性を検証した．検証の結果，実用に向けて様々な課題があることが判明した．本章では，各手法の考察を述べ，今後の研究に向けた課題点について説明する．

5.1　レイアウト解析の考察と課題

提案手法のレイアウト解析において，検証によりいくつかの課題点が挙げられた．以下より，行認識に失敗する例について説明し，今後の課題を述べる．

解像度が低い画像における誤認識について述べる．図 5.1 (a)のように，解像度が著しく低く文字が全体的にぼやけている画像は，レイアウト解析精度が低下する．邦字新聞デジタル・コレクション[1]によって提供される画像は，発行年代や新聞の種類により，解像度が低いものがある．人間の目では文字や行として認識できる解像度のものであっても，提案手法のレイアウト解析では正しく認識することができない場合が多い．今後，解像度の低い画像に対応する方法を検討する必要がある．

段組みを上下に区切る横線を，行領域として誤認識する場合について述べる．提案手法では，前処理の時点で横線の検出と除去を行っている．しかし，横線にゆがみや途切れがあると，正しく除去されない場合がある．図 5.1 (b)は，横線が除去されておらず，段を上下にまたいで 1 つの行領域として認識される例を示している．拡大画像に示す矢印が段を区切る横線の位置を示しており，本来は段ごとに行領域が抽出されなければならない．このような場合，鮮明な画像であっても行領域の抽出に失敗する．邦字新聞画像は，ゆがみやノイズが原因でほとんどの場合横線が直線的に引かれておらず，太さにばらつきがある．提案手法では前処理としてモルフォロジー演算による横線の除去を行っているが，線にかすれや途切れがある場合に対応できておらず，改善が必要である．



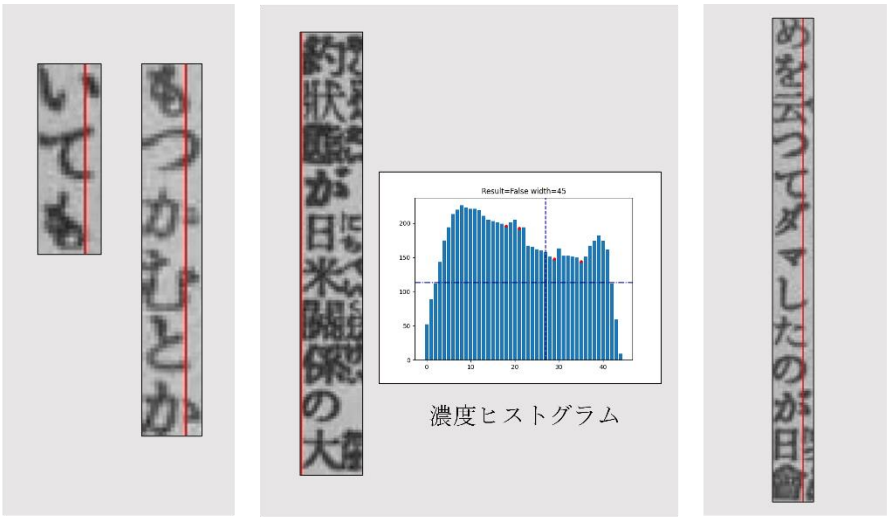
(a) 全体の解像度が低い　(b) 横線の検出が不足している　(c) カタカナや句読点

図 5.1 レイアウト解析の失敗例

カタカナや句読点が含まれる行の誤認識について述べる．図 5.1 (c)に示すように，行内に連続してカタカナが含まれる場合や，間に句読点が挟まる場合に，抽出される行領域が途中で切れてしまう場合がある．これは，カタカナや句読点のような画数の少ない文字がノイズと同様に認識されているためであると推測される．CRAFT[9]の後処理において，ノイズであると判定された小さな領域は削除される．よって，カタカナや句読点が同時に削除される場合がある．同様の理由で，文字がかすんでいるものに対しても，行領域として認識されない場合がある．CRAFT における後処理手法の改善と，前処理によるノイズ除去手法を検討する必要がある．

5.2 ルビ除去手法の考察と課題

提案手法におけるルビ除去手法は，邦字新聞のような活版印刷により規格化されたフォントが用いられていない文書に対応する手法である．レイアウト解析で得られる行領域には，ルビの含まれる行と含まれない行が混在するため，ルビの有無判定を行っている．検証により，ルビ除去によって文字認識の精度が向上することが示された．しかし，ルビの有無判定を誤ることにより，正しくルビ除去が行えない場合がある．特に，ルビの含まれない行に対してルビの有無判定を誤り，不必要にルビ除去処理を行う場合，文字認識率の低下につながるため，改善が必要である．以下よりルビ除去に失敗する例について説明し，今後の課題を述べる．



(a) 分離する要素を持つひらがなを含む行　(b) 漢字を多く含む行　(c)ひらがなと漢字が混合する行

図 5.2 ルビ除去の失敗例

ルビが含まれない行に対してルビの有無判定を誤る例について述べる．図 5.2 (a)は，「い」や「か」のように分離する要素を持つひらがなが含まれる行領域に対して，ルビの有無判定を誤る例である．画像に示す赤色の線より右側部分がルビであると誤判定され，ひらがなの一部が行領域から除かれている．特に，ひらがなだけで構成される行に対して，このような誤りが多く見られる．

ルビが含まれる行に対してルビの有無判定を誤る例について述べる．図 5.2 (b)は，ルビが含まれる行領域であるが，ルビ除去が行われていない例である．右側に，ルビの有無判定に用いられる縦方向濃度ヒストグラムを示している．ヒストグラムに谷となる部分が表れていないことから，行領域にルビは含まれていないと判定されている．これは，行のほとんどがルビの振られた漢字で構成されるときに，漢字とルビとの間隔が狭いために，縦方向濃度の差がでないことが原因である．

ルビ除去によりひらがなの一部が除去される失敗例について述べる．図 5.2 (c)は，赤線より右の位置でルビが除去され，ひらがなの一部分がルビと同時に除去される例である．邦字新聞が活版印刷であり，文字サイズや位置が規格化されていないことから，ルビとひらがなが重なっている場合がある．ひらがなの一部がルビと同時に行領域から除外されるため，文字認識の低下につながる．

ルビ除去手法の課題について述べる．ルビが含まれない行に対する誤判定は，誤った位置でルビ除去を行うこととなり，文字認識の精度低下につながるため，改善が必要である．提案手法では，フーリエ記述子による文字の輪郭の概形で得る縦方向濃度ヒストグラムのみをルビの有無判定に用いる．行領域における文字の密度は，行領域に漢字やひらがながどれだけ含まれるかにより異なる．ひらがなのみで構成される行は，漢字が

含まれる行に対して文字の密度が低いと推測される. 文字の密度を利用してひらがなのみで構成される行をルビ除去対象から除外するといった, 濃度ヒストグラム以外の方法を用いたルビの有無判定方法を検討する必要がある.

5.3 読み順検出手法の考察と課題

提案手法では, 複数記事で構成される邦字新聞に対応した読み順検出のため, 本文に書かれている内容的観点から同一記事推定を行うことで読み順を検出する. 検証の結果から, 同一記事推定手法の考察と課題について述べる.

形態素解析に誤りがあることから推定を失敗する例について述べる. 検証の結果, 形態素解析による名詞抽出において,「ハツキリ」という単語が「ハツ」と「キリ」に分けられ, 名詞に分類されるという結果が見られた.「ハツキリ」というのは, 現代口語体における「はっきり」という副詞であり, 本来名詞の抽出からは除外される対象の単語である. 邦字新聞は, 明治維新前後の時代に発行されており, 文体が近代文語体である. しかし, 提案手法に用いられる形態素解析エンジン MeCab[17]のシステム辞書は近代文語体に対応していないため, 名詞抽出に誤りが生じる場合がある. 形態素解析に誤りがあると, 同一記事推定における単語の重要度計算に必要のない要素が加わるため, 対応が必要である. 名詞を漢字のみに限定するなどの工夫をする他, 近代文語体に対応する形態素解析の手法を検討する必要がある.

提案手法における同一記事推定の結果は, 文章ブロック同士における内容の類似度計算結果に依存する. 遠い位置関係にある文章ブロック同士で, 同一記事である可能性が低いものであっても, 類似度計算の結果によっては同一記事であると判定される. 構造的な観点における同一記事推定手法と, 文章的な観点における同一記事推定手法を組み合わせることで, 精度の向上が期待できる.

5.4 今後の研究に向けて

本修士論文では, 邦字新聞に対応する OCR システムの手法を提案した. 一般に, OCR システムの実用には, 精度が 99.9%以上であることが望まれる. 特に, OCR システムにおけるレイアウト解析精度は, システム全体の認識率に影響する. 今後は, 提案手法によるレイアウト解析において, 99.9%以上の精度の達成を目指す. レイアウト解析精度向上のためには, 前処理による 2 値化とノイズの除去方法の検討が必要である. その他, 多段組みで構成される邦字新聞において, 段組みを区切る横線の除去が必要である. 提案手法では, モルフォロジー演算によるオープニング処理を用いて横線を検出し, 除去を行っている. しかし, かすれにより横線が途切れてしまう場合に対応できない. 太さ, 長さが異なり, かすれがある横線に対応可能な検出方法を検討する必要がある.

提案手法は，邦字新聞の本文を文字に起こすことを目的としている．しかし，邦字新聞には見出しや図，広告などの本文以外の要素が含まれる．提案手法におけるレイアウト解析では，本文以外の要素を区別する手法は備わっていない．そのため，図や広告に含まれる文字が本文の要素として抽出される場合がある．今後，本文以外の要素を区別する手法を追加し，更なる精度の向上を目指す．

本研究における最終的な目標は，邦字新聞に対する自動テキスト化ならびに現代口語体への翻訳システムの構築である．特に，邦字新聞デジタル・コレクションの全文検索において，現代口語体によるキーワードでも検索が行え，邦字新聞の情報に手軽にアクセスできる環境を構築することを目的としている．

邦字新聞デジタル・コレクションの翻訳システム構築に向けた近代文語体と現代口語体の相互翻訳手法が提案されている[15]．相互翻訳手法では，手作業による文字起こしで作成された近代文語体，現代口語体が対となった学習データが利用されている．相互翻訳手法の利用には，正しい読み順で出力されたテキストデータが必要不可欠である．提案手法では，1 ページが複数の記事で構成される邦字新聞に対応するため，本文内容を考慮した同一記事推定を行っている．しかし，検証により精度は不十分であることが示されており，今後は他手法との組み合わせを検討する必要がある．

文章の話題ごとに適切な区切りをつけるテキストセグメンテーションの手法がある．テキストセグメンテーションは，文書の検索や要約，抽出に応用される．テキストセグメンテーションの手法の 1 つに，Jeonghwan Lee らによって提案される 3 つの分類層を用いたマルチタスク学習技術を適用する手法[22]がある．3 つの分類層をマルチタスク方式で組み合わせることで，相互補完的に 3 つのレイヤーの精度が向上し，高精度なテキストセグメンテーションを行うことができる．複数の記事により構成される邦字新聞において，文章をトピックにより抽出することに活用が可能である．今後，マルチタスク学習技術を適用する手法により，同一記事推定手法の精度改善を試みる．

第6章　おわりに

本修士論文では，邦字新聞に対応したOCRシステムを提案する．システムの開発のため，既存システムであるNDLOCRに用いられる一部手法の置き換え，追加を行う．

邦字新聞とは，明治維新前後よりアメリカ大陸，アジアにおいて日本人移民により発行された海外日系新聞の集合体である．現在，スタンフォード大学フーバー研究所にて収集された邦字新聞画像データが，邦字新聞デジタル・コレクションとして公開されている．邦字新聞は様々な出版社により刊行され，その性質は，コミュニティや軍事プロパガンダなど多岐にわたる．邦字新聞テキストデータの研究活用，邦字新聞デジタル・コレクションの全文検索機能の追加へ向け，全文自動テキスト化が求められている．

日本語文書に対応したOCRシステムの1つに，NDLOCRがある．NDLOCRは，国立国会図書館の所蔵する約 262 万タイトルの書籍資料画像が学習データの構築に利用されている．対応文字種はひらがな，カタカナやJIS第一，第二水準漢字などであり，日本語文書に用いられるほとんどの文字種に対応している．よって，様々な年代，種類の日本語文書に対して高精度なテキスト化が期待できる．

しかし，邦字新聞に対するNDLOCRの精度は，未だ不十分である．NDLOCRが邦字新聞に対応し，十分な精度を得るための3つの課題点を挙げる．1つ目の課題は，レイアウト解析精度の低さである．邦字新聞のような多段組みの構成をとるレイアウトに対して，NDLOCRのレイアウト解析手法では対応できず，精度が不十分である．2つ目の課題は，ルビの含まれる行に対する文字認識精度の低さである．邦字新聞にはほとんどの漢字にルビが振られている．NDLOCRでは邦字新聞のルビを検出することができず，ルビがノイズとなり文字認識の精度低下の原因となると推測される．3つ目の課題は，読み順検出精度の低さである．NDLOCRでは，余白を利用して文章のブロック分けを行い，簡単なルールベースにより並び替えることで読み順を検出する．しかし，複数の記事により構成され，記事ごとにレイアウトが異なる邦字新聞においては，構造的観点のみで読み順を決定することは困難である．読み順が検出できなければ出力が文章として成り立たず，研究に対する活用や全文検索機能への活用において不便となる．本修士論文では，以上の3つの課題点を踏まえ，NDLOCRにおける手法の一部を改善し，邦字新聞に対応したOCRシステムとして提案する．

提案手法では，NDLOCRに対して以下の3つの改善を行う．1つ目に，レイアウト解析手法の変更を行う．レイアウト解析の手法に，解像度ピラミッドを利用したCRAFTの手法[4]を採用する．CRAFT[9]は，CNNにより文字の中心，および文字間のヒートマップを出力するレイアウト解析手法である．CRAFTに解像度ピラミッドを組み合わ

せることで, 学習にかかる計算資源を削減して多段組みかつ多サイズ文字により構成される近代書籍に対応したモデルを構築している.

2 つ目に，ルビ除去処理の追加を行う．レイアウト解析により抽出される行領域内には，ルビが含まれる行と，含まれない行が混在する．ルビが含まれない行に対してルビ除去処理を行うと，文字認識精度低下の原因となる．そのため，ルビの有無判定を行った後に，ルビ除去処理を行う．まず，フーリエ記述子により行領域内の文字の輪郭の概形を得る．次に，得られる概形に対して縦方向に濃度ヒストグラムをとる．ヒストグラムの谷となる部分があるものを，ルビが含まれる行と判定する．最後に，ルビが含まれると判定された行に対して，濃度ヒストグラムの谷となる位置でルビを分離する．

3 つ目に，読み順検出手法の変更を行う．読み順検出のため，複数の記事で構成される邦字新聞に対し，同一記事推定を行う．まず，レイアウト解析の出力を利用し，余白や線で区切られる文章ブロックを分割する．そして，文字認識により得られた出力に対して形態素解析を行い，名詞を抽出する．最後に，Okapi-BM25[5]とコサイン類似度計算により文章ブロック同士の類似度を算出し, 類似性が高いと算出された文章ブロックの集合を同一記事であると判定する.

改善に用いられる 3 つの手法における有効性を検証する. レイアウト解析手法の有効性を検証するため，発行年代，新聞名の異なる 8 枚の邦字新聞画像を用いる．検証の結果，提案手法によって得られるレイアウト解析の認識率は最高で 97.69%，最低で 71.73%であり，平均認識率は 87.60%となった．NDLOCR ver.2.1 によるレイアウト解析の結果と比較し，最大で 42.11%の差で精度が向上した．検証の結果，レイアウト解析手法が邦字新聞に対して有効であることが示された．誤認識として，段組みを区切る横線が除去されておらず，文字として認識されることで，段を上下にまたいだ 1 行として抽出される例が見られた．レイアウト解析の更なる精度の向上のため，段組みを区切る横線の除去手法など，画像の前処理手法を検討する必要がある.

ルビ除去処理手法における有効性を検証するため, 提案手法と NDLOCR の文字認識精度を比較する．3 枚の邦字新聞画像を用いた検証の結果，提案手法において最も文字認識率の高いもので 73.23%，次いで 60.10%，最も低いもので 45.57%となった．いずれも NDLOCR による文字認識率を上回る結果となり, ルビ除去手法の有効性が示された.「い」や「か」といった分離する要素を含むひらがなのみで構成される行において，誤ってルビ除去処理を行う例がみられた. 誤った位置でのルビ除去は文字認識の精度低下に繋がるため，ルビの有無判定手法を改善する必要がある.

読み順検出の手法を検証するため，5 枚の邦字新聞画像を用いて検証を行う．検証の結果，提案手法により 1 つの記事を構成する一部の文章ブロックが抽出された．しかし, 同一記事を構成するすべての文章ブロックを網羅して抽出できているものは 1 つもないことが確認された．推定結果の誤りには，形態素解析の際に用いられるシステム辞書が近代文語体に対応していないことが原因で, 名詞の抽出が正しく行われていない例

が見られた．提案手法では，同一記事推定結果が文章ブロックの類似度計算結果に依存する．余白や文章ブロックの位置関係を利用した構造的観点と，提案手法のような文章的観点の両手法を組み合わせた同一記事推定手法の検討が必要である．

邦字新聞に対応したOCRシステムは，スタンフォード大学フーバー研究所上田薫教授のもとで実際の運用が求められている．提案手法によるOCRシステムは，異なるマシン間でもシステムを容易に移植し，正常に動作させるため，Dockerを用いてコンテナ型の仮想環境構築を行っている．Dockerの利用により，異なるマシンでのUbuntu，Windows環境下においてシステムが正常に動作することを確認している．

本研究における最終目標は，邦字新聞に対する自動テキスト化ならびに現代口語体への相互翻訳システムの構築である．提案手法により，既存手法のNDLOCRに対してレイアウト解析精度の改善，ルビ除去処理による文字認識精度の改善が見られた．実用に向けて，OCRシステム全体として99.9%以上の精度を達成する必要がある．今後，前処理手法によるノイズの除去や2値化手法，段組みを区切る横線の除去方法を検討し，レイアウト解析精度の向上を目指す．また，近代文語体，現代口語体への相互翻訳手法の適用には，読み順通りの出力が必須である．邦字新聞における同一記事推定手法を高精度なものにするため，他手法との組み合わせを検討していく．

謝辞

本研究を行うにあたり，指導教官である城和貴教授には，丁寧なご指導と的確な助言をいただき，大変お世話になりました．副教官である髙田雅美先生には，研究への助言だけでなく，あらゆる面でサポートして頂きました．心からお礼を申し上げます．

城研究室の皆様には，研究にあたり，様々な助言，サポートをいただきました．この場を借りて感謝の意を示させていただきます．ありがとうございました．最後に，進学を応援し，いつも支えてくれた家族に感謝いたします．ありがとうございました．

参考文献

[1] 邦字新聞デジタル・コレクション: https://hojishinbun.hoover.org/?l=ja（参照 2023-11-11）

[2] NDLOCR: https://github.com/ndl-lab/ndlocr_cli（参照 2023-11-11）

[3] B. Shi, X. Bai, and C. Yao. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 2298/2304, (2017).

[4] 飯田紗也香, 竹本有紀, 石川由羽, 髙田雅美, 城和貴. 多段組多サイズ見出しで構成される近代書籍のレイアウト解析. 情報処理学会論文誌数理モデル化と応用. (2023).

[5] Robertson, S. E. and Zaragoza, H.. The Probabilis- tic Relevance Framework: BM25 and Beyond, Found. Trends Inf. Retr., Vol. 3, No. 4, pp. 333–389, (2009).

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition: https://arxiv.org/abs/1512.03385（参照 2023-11-21）

[7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie. A ConvNet for the 2020s: https://arxiv.org/abs/2201.03545（参照 2023-11-21）

[8] Baek, Youngmin, et al." Character region awareness for text detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019).

[9] 粟津妙華, 髙田雅美, 城和貴:遺伝的プログラミングを用いた近代書籍からのルビ除 去, 情報処理学会論文誌. 数理モデル化と応用(TOM), Vol. 6, No. 2, pp.53-62(2013).

[10] 粟津妙華, 髙田雅美, 城和貴: 活字データを用いた近代書籍からのルビ除去, 情報処 理学会論文誌. 数理モデル化と応用, Vol.8, No.1, pp.72-79 (2015).

[11] 令和 4 年度 NDLOCR 追加開発事業及び同事業成果に対する改善作業: https://lab.ndl.go.jp/data_set/r4ocr/r4_software/（参照 2023-12-01）

[12] Fukuo, M., Enomoto, Y., Yoshii, N., Takata, M., Kimesawa, T. and Joe K.: Evaluation of the SVM based Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books, Proc. 2011 International Conference

Parallel and Distributed Processing Technologies and Applications (PDPTA 2011), Vol. Ⅱ, pp.727-732. (2011).

[13] Komatsubara, M., Ishikawa, C., Takata, M., Kamo, H., Nide, N., and Joe, K.: Auto Classification of Feces for Health Condition Analysis, PDPTA, (2007).

[14] J. Canny, “A Computational Approach to Edge Detection,” IEEE Trans.Pattern Analysis and Machine Intelligence, vol. 8, pp. 679-698. (1986).

[15] Honoka Nishikawa, Yuki Takemoto, Sayaka Iida, Yu Ishikawa, Masami Takata, Kaoru Ueda and Kazuki Joe: Translating Early-modern Written Style into Current Colloquial Style in Hoji Shinbun, Advances in Parallel & Distributed Processing, and Applications (PDPTA ‘22).

[16] 福元春奈，竹本有紀，石川由羽， 髙田 雅美，城和貴，近代書籍文字認識に対応した誤字検出，情報処理学会研究報告，vol.2022 MPS 141 No.21 pp.16, (2022-12).

[17] MeCab: https://github.com/taku910/mecab（参照 2023-12-01）

[18] J. Matas, C. Galambos and J.V. Kittler: “Robust Detection of Lines Using the Progressive Probabilistic Hough Transform,” CVIU, vol.78, no.1, pp.119–137, 2000.

[19] Docker: https://www.docker.com/ja-jp/（参照 2024-01-22）

[20] NVIDIA Container Toolkit: https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/latest/index.html（参照 2024-01-22）

[21] Windows Subsystem for Linux: https://learn.microsoft.com/ja-jp/windows/wsl/（参照 2024-01-22）

[22] Jeonghwan Lee, Jiyeong Han, Sunghoon Baek, Min Song. Topic Segmentation Model Focusing on Local Context: https://arxiv.org/abs/2301.01935（参照 2023-12-01）

研究業績

国際研究集会（口頭発表のみ）

1. 熊谷 もも，邦字新聞 OCR の概要と設置，スタンフォード大学上田研究室セミナー，スタンフォード大学フーバー研究所，2023 年 10 月．

国内学会（口頭・査読無）

1. 熊谷 もも，古磯 則江，髙田 雅美，上田 薫，城 和貴，多段組みで構成される近代書籍の読み順推定手法の検討，研究報告数理モデル化と問題解決（MPS），Vol.2023-MPS-143，No.21，pp.1-6，2023 年 6 月．