N-003

画像処理と深層学習を用いた古典籍の整理

Early books understanding base on image processing and deep learning

呂氷†, 冨山宏之†, 孟林†

Bing Lyu, Hiroyuki Tomiyama, Lin Meng

1. はじめに

古典籍は文化遺産として多くの情報が記録されているため、古典籍の整理は政治、歴史、文化などの研究に役立つ。日本には整理されていない古典籍が多く存在し、更なる日本文化の研究と理解のため、日本古典籍の再整理が急がれている。しかし、多くの古典籍はくずし文字により記述されている。くずし文字は文字を簡略化し、迅速に記述できる書体で、「草書」とも言われる。現在、くずし文字はほとんど使われていなく、極わずかな専門家しか解読できないため、古典籍の解読は非常に時間をかかり、難航している。

近年、人工知能技術が邁進し、研究者らは深層学習（Deep Learning）を用いて、くずし文字の自動認識を目指している。しかし、これらの研究では、一つずつの文字を手動で切り取ってから認識を行うため、全文章の自動認識がまだ実現されていない。また、古典籍において、文章と画像を混在するケースが多く存在し、文章を自動解読するのに、予め文章と画像を分離する必要がある。しかし、画像は、文字と同じように、手書きで描かれているため、全文章の自動解読の難しさを増している。

本研究では、画像処理と深層学習を用いて、日本古典籍の画像に対して、文章領域を自動で抽出し（Document Segmentation）、文字の認識(Character Recognition)を実現することを目指し、古典籍を解読し、日本の文化遺産の保存と整理に貢献する。詳細について、立命館大学・ARC 古典籍ポータルデータベースに所蔵されている古典籍画像に対して、ARU-Net[2]を用いて、文章領域を抽出する。そして、その評価結果に基づいて、必要に応じた古典籍画像の鮮明化などを提案した。また、文章領域抽出ミスに対して、LeNet を用いて、文字として認識された絵などを、抽出ミスとして削除する。最後、AlexNet や GoogLeNet を用いて、文字を認識する。本論文は文章の領域抽出を注目する。

2. 深層学習について

画像処理での深層学習について、大きく画像セグメンテーション（領域抽出）、画像識別の二つに分けられている。画像のセグメンテーションでは、U-Net,SegNet がよく知られている。また、ARU-Net は英文の古典籍のセグメンテーションとして使用されている報告もある。しかし、日本古典籍の文章領域のセグメンテーションが未開拓の研究であり、本論文はそれを実現することを目指す。

† 立命館大学 大学院理工学研究科, Graduate School of Science and Engineering, Ritsumeikan University

画像の識別では、簡単な手書き文字認識用の ALexNet から一般物体認識用の AlexNet,GoogLeNet,VGG などの複雑なモデルが提案されている。我々の研究では、AlexNet,GoogLeNet は、拓本など古典文献の認識に有効であることが示されている[3]。しかし、これらの手法の問題点は、予め文字の切り取りが必要である。

さらに、現在はセグメンテーションと画像認識の組み合わせで、画像から、物体領域の抽出と認識を同時に行うモデルも提案されている。その中には、R-Net,YOLO,SSD などが挙げられている。SSD には画像識別モデルである VGG をベースモデルとして使用され、中国古典籍の甲骨文字認識の有効性にも示されている[4]。しかし、これらのモデルが対応できるクラス数が限られているため、数千個のクラスを有する文字に対して、一つの課題となる。

3. 深層学習を古書籍の認識

本プロジェクトは、各深層学習のモデルの強みを生かし、古典籍の文章領域の抽出から、文字の認識までの自動化を目指している。まず、ARU-Net を用いて、文章領域を抽出し、LeNet を用いて、抽出された文章と抽出ミスにより抽出された手書きの絵を分類し、文章の領域抽出ミスを削減する。最後に、AlextNet や GoogLeNet を用いて、文字の認識を行う。本論文では、主に、ARU-Net を用いた日本古典籍の文章領域の抽出を紹介する。

ARU-Net では、歴史的文書におけるテキスト行検出のための深層学習のモデルである。図 1 は ARU-Net の構成を示し、A-Net と RU-Net がベースネットワークとなる。A-Net を用いて、画像中の明るい領域と暗い領域などを区別できる。ARU-Net は、任意のピクセルラベリングを行うことにより、二値化の結果、ページのずれに使用することができる。 ARU-Net は、テキストの各行の開始位置と終了位置のなどを予測することができる。ARU-Net は複雑な文章のレイアウトだけでなく、曲げたりするテキスト行も認識できる。図 1 により詳細の処理の流れは、以下のように示す。原画像に対して、A-Net された結果を softmax し、原画像を RU-Net に経由し、その二つの結果を融合する。また、二

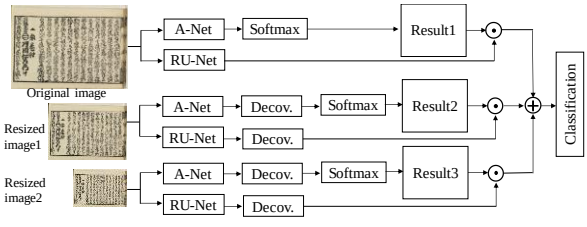

図 1 ARU-Net

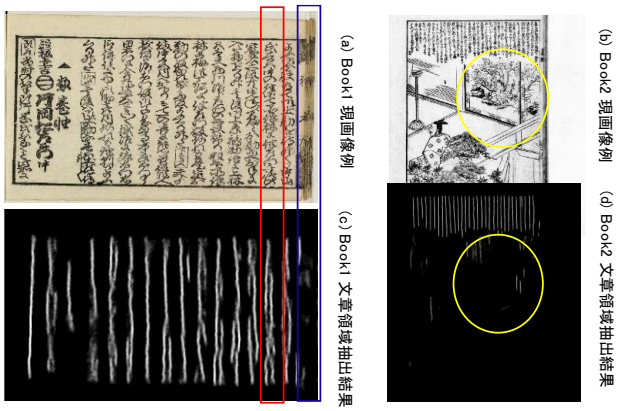

図 2. 処理結果例

つのパターンで縮小された原画像に対して、A-Net された結果を deconvolution と softmax 行い、その圧縮された画像を RU-Net に経由し、deconvolution を行い、その二つの結果を融合する。最後に、A-Net と RU-Net 処理で生成された特徴図の融合結果を用いて、最終的に分類を行う。

4. 実験

4.1 実験条件

本研究は立命館大学アート・リサーチセンター[1]所蔵の「役者神事競」（資料番号：arcBK04-0072、Book1 とする）と「絵本義経一代実記」（資料番号：ナ 04-0752、Book2 とする）を用いて、評価実験を行った。Book1 は文章のみで、Book2 は文章と図両方がある。OS は ubuntu16.04 LTS で、プログラミング言語は Python である。

4.2 実験結果

図 2(a)(b)は、Book1 と Book2 の中の一枚のスキャン写真で、文章領域の抽出結果は図 2(c)(d)を示す。図 2(c),(d)の中の白い線は抽出された文字領域である。この結果によって、画像の中で全部文字領域は正しく抽出されたことが分かった。しかし、図 2(c)は画像の枠のノイズ（青い長方形で示す）と 1 行の文字を 2 行として認識された問題（赤い長方形で示す）がある。また、図 2(d)には、画像のノイズ（黄色の丸で示す）である。

全体評価について、図 3 には、2 冊の本のそれぞれの 10 ページ(P1～P10)の実行後の適合率、再現率と F 値、及びそれらの平均値を示す。再現率はほぼ 1 であるため、全て

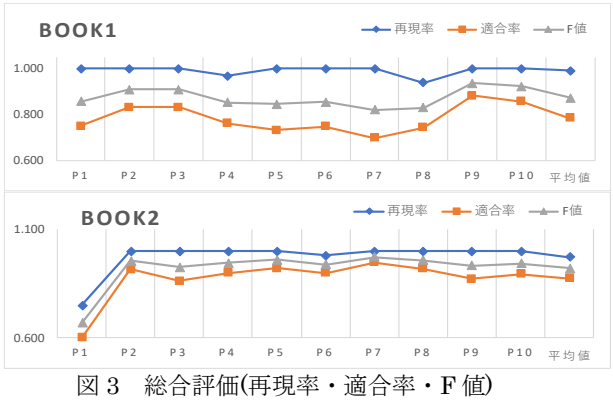

図 3 総合評価(再現率・適合率・F 値)

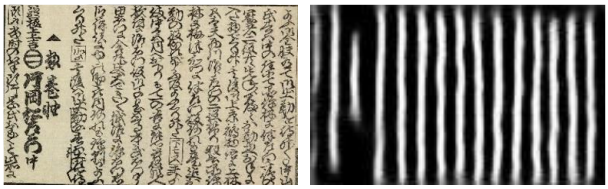

図 4.枠ノイズ除去とその結果

の文章領域が正しく抽出できたとわかった。しかし、ノイズ及び、手書きの絵の一部が文章として認識されたため、適合率と F 値は 1 に達成することができなかった。ノイズの除去と手書きの絵と文字の分類により、文章領域抽出の改善案が今後の課題となる。

4.3 ノイズ除去による精度の向上

Book1 の枠ノイズ除去による精度の向上について説明する。まず、図 2(a)に対して、大津法により閾値を決定し、二値化処理を行う。そして、縦軸と横軸での黒ピクセルヒストグラム（文字とノイズ部）をカウントする。両軸の黒ピクセルのヒストグラムについて、枠のラインでは非常に高いとの特性がある。従って、両軸のピクセルのピーク値(3 つ以上がある)を探し、隣のピーク値間で微分演算を行う。各軸において、大きさが上位二つの微分結果を持つ座標は、枠となる。最後に、両軸で見つける 4 つの上位座標によって画像をカットする。図 4(a)は図 2(a)の枠ノイズ除去後の結果で、図 4(b)は文章抽出の結果である。図 4(a)では枠の削除ができ、図 4(b)では、枠ノイズの影響をなくなり、1 行の文字を 2 行として認識された問題も解決された。

5. おわりに

本論文では、ARU-Net を用いて日本の古典籍の文章の自動的抽出を紹介し、その有効性性能を示した。評価において、再現率は 1 に達成できたが、適合率と F 値は 1 に達成できなかった。つまり、文章の部分が正しく抽出されたが、手書きの絵が文章部として判定されたケースが存在する。それにより、簡単なモデルを用いて、手書きの絵と文章を分類することにより、文章領域抽出のミスを削減できると見込まれ、今後の課題となる。更なる実験の追加と、文章抽出後の文字の自動的な認識も今後の課題となる。

謝辞

本研究は立命館大学アート・リサーチセンターの助成を受けたものです。また、助言などを頂いた立命館大学の赤間亮先生と金子 貴昭先生に感謝します。

参考文献

[1] 古典籍ポータルデータベース，立命館大学アート・リサーチセンター：https://www.arc.ritsumei.ac.jp/database.html (2019.4.6.)

[2] T. Grüning, et al. “A Two-Stage Method for Text Line Detection in Historical Documents,” CVPR2018, 2018

[3] L. Meng, et al., “Ancient Asian Character Recognition for Literature Preservation and Understanding,” EuroMed2018 Int. Conf. on DIGITAL HERITAGE, 2018.

[4] L. Meng, et al.,“Recognition of Oracle Bone Inscriptions Using Deep Learning based on Data Augmentation,” 2018 IEEE Int. Conf. on Metrology for Archaeology and Cultural Heritage, 2018.