画像特徴を用いた難読古典籍画像に対する内容解析と提示方法の開発

公立はこだて未来大学　寺沢 憲吾

研究の背景・目的

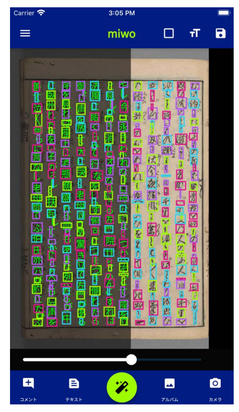
近年では、CODHが開発したAIくずし字認識モデル「KuroNet」や、そのスマホアプリ「みを」に代表されるように、AI技術を用いて古典籍画像から文字認識（テキストデータ化）を行うことが現実になりつつあります。文字認識ができれば、全文検索だけでなく、自然言語処理（NLP）技法を用いた内容解析など、さまざまな方面から人文学研究に役立つことが期待されます。

一方で、現状では一定以上の精度で文字認識ができるものは膨大な古典籍の一部に限られていることも指摘されています。

本研究は、現在発展めざましい「画像→テキスト化→自然言語処理」という流れとは別のアプローチとして、「テキスト化を経ずに、画像処理技術を駆使してできることを探る」方向性で、AI技術の人文学研究への活用法を模索するものです。

「KuroNet」で翻刻された「徒然草」
使用した画像 :http://kotenseki.nijl.ac.jp/biblio/200019959/viewer/2

「KuroNet」 による認識結果
（「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」ニューズレター ふみ 第13号（2020年1月）より



「みを」の紹介画像
http://codh.rois.ac.jp/miwo/ より

文字認識できるのはすばらしい。
でも、高精度で文字認識できる文献ばかりではない？

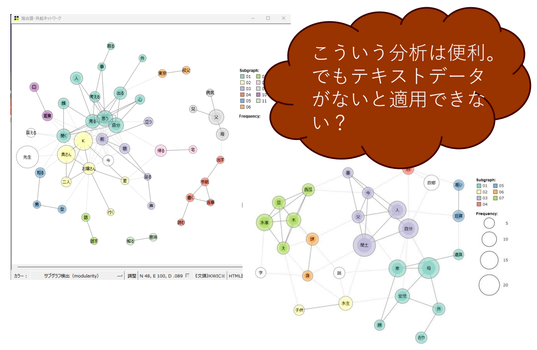
国会図書館デジタルコレクションでは約10万点の古典籍資料を閲覧できるが、すべてがテキスト化されているわけではない

研究のねらい

「テキスト化（文字認識）」以外のやり方で、画像技術の人文学研究への貢献を図る具体的な方法として、本研究では主に2つの研究テーマ（内容解析と提示方法）を掲げています。

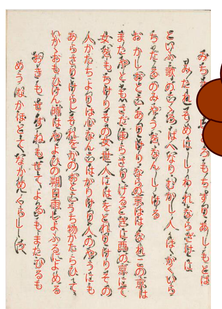
第一に、**まだテキスト化されていない**古典籍画像をターゲットとして、テキスト化を経ずに、画像のままで、自然言語処理のようなことをできるようにしよう、というのが「内容解析」です。本研究では、古典籍画像を対象に、字形の類似性等に着目して作成する「**画像特徴に基づく擬似コード**」を作成し、これに基づいて内容解析を行う新たな技術を開発します。

もう一つのテーマは、テキスト化の結果を取り入れつつ、古典籍原典に現れる文字の字形を、原典ならではの味わいを極力保持しつつ、できる限り読みやすいところまで変形する、画像と認識結果の新たな提示方法を開発するものです。これにより、専門教育を受けていない一般の人が古典籍の魅力に触れる機会を増やすとともに、初学者の学習に役立つツールの開発にも結びつくことを期待しています。



「KH Coder」 による共起解析の例。本研究は、「テキスト化されていない文書画像」に対しても、このような分析が可能となることを目指す。

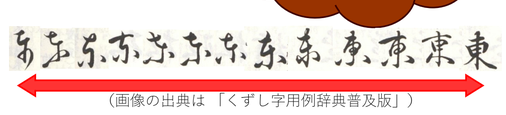
左側は夏目漱石「こころ」、右側は魯迅「故郷」を対象にした共起ネットワークの例。
左側の例では「K」と「お嬢さん」と「奥さん」が、右側の例では、「自分」と「閏土」が、それぞれ共起性が高いということが可視化されている



CODH，KogumaNetくずし字認識ビューア

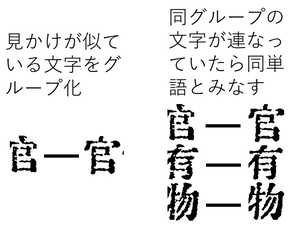
機械に読んでもらったのではなく、「自分で読んだ」気持ちを味わってみたい？

どのへんまでなら無理なく読める？

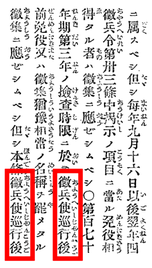


（画像の出典は 「くずし字用例辞典普及版」）

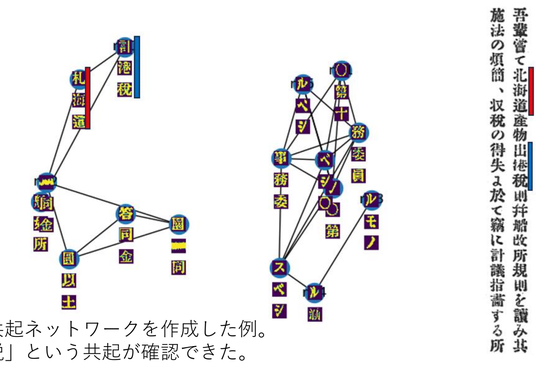
画像のままで行う内容解析



このグループ化のプロセスは、一般的な文字認識のプロセスと異なり、訓練データを事前に大量に用意する等のプロセスが不要である。



擬似コードの例。何と書いてあるかはわからないが、同一単語の検出は可能



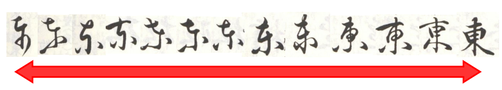
擬似コードに基づき共起ネットワークを作成した例。
「北海道」と「出港税」という共起が確認できた。

（主な発表実績）
Sora Ito and Kengo Terasawa, "Extraction of Distinctive Keywords and Articles from Untranscribed Historical Newspaper Images,“IWAIT2020, Jan. 5-7, 2020.

可読性を高めた提示方法

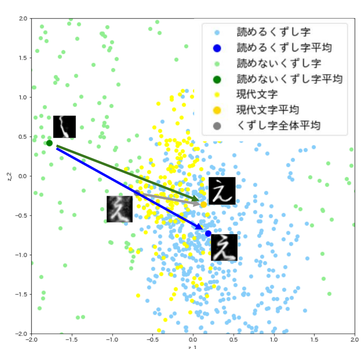
このテーマでは、古典籍に現れる難読な文字について、もともとの字形の特徴をできるだけ維持しつつ、できる限り読みやすい字形に近づけていく、新しい提示方法を開発しています。これにより、専門教育を受けていない一般の人が古典籍の魅力に触れる機会を増やすとともに、初学者の学習に役立つツールの開発にも結びつくことを期待しています。

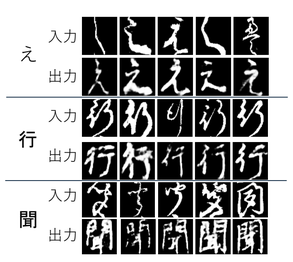
手法としては、人工知能技術（画像生成AI）の様々な手法を取り入れています。この分野もより新しく高性能な方法が次々に公開されているため、それらを本手法にも取り入れ、より自然な見た目となる変形画像の生成法を追求して、研究開発を進めています。



【補間のイメージ図】
原典が著しくくずした草書の場合も、
段階的に楷書に近づけることができる
（画像の出典は「くずし字用例辞典普及版」）

DDIMを用いてランダムに生成した画像に対する補間画像の作成結果。
訓練に用いたのはETLデータセット。





VAEを用いて、潜在空間における「可読性向上ベクトル」を定義し、実際に現れるくずし字画像の可読性を向上させるアイディア（荒木2022）
→実験協力者に評価してもらい、変形による可読性の向上を実際に確認

	変換前正解率	変換後正解率	筆跡保持に関する実験の正解率
平仮名	23.3% (3.5/15)	66.7% (10.0/15)	75.0% (3.8/5)
単純な構造の漢字	48.3% (7.0/15)	91.7% (13.8/15)	65.0% (3.3/5)
複雑な構造の漢字	25.0% (3.8/15)	86.7% (13.0/15)	50.0% (2.5/5)
すべての文字	32.2% (14.5/45)	81.7% (36.8/45)	63.3% (9.5/15)

【手法の流れ】

既存手法を用いて、原本画像に現れる文字の字種・座標を取得

認識結果に基づいて、zi2ziによる、比較的読みやすい文字を生成

原本画像から文字を消去し、その上に生成した文字を配置。
この際、文字サイズや文字数などを用いて自然な配置となるように工夫する

フォント自動生成手法として提案されたzi2ziモデルを用いて、現代文字にくずし字の書体の特徴を与えて、読みやすさを向上させた文字の生成を行うアイディア（蠟崎2022）
→実験協力者に評価してもらい、古典籍の原典画像、既存の翻刻結果表示手法、提案手法を比較評価した結果、提案手法が可読性と鑑賞性について両者バランスよく保持できているという結果が示された

（主な発表実績）
荒木亮介，寺沢憲吾，芸術性と可読性を備えたくずし字の生成，情報処理学会第84回全国大会，愛媛大学（ハイブリッド）2022年3月3日-5日，学生奨励賞受賞
Kosuke Kakizaki and Kengo Terasawa, "A Novel Historical Manuscripts Displaying Method That Improves Their Readability While Preserving Their Appreciation,“ International Workshop on Frontiers of Computer Vision, IWFCV2022, Hiroshima, Japan (Online), Feb. 21-22, 2022. Best Presentation Award 受賞
蠟崎公亮，寺沢憲吾：古典籍の可読性を向上しつつ鑑賞性を維持する翻刻結果の新たな表示法，人文科学とコンピュータシンポジウム　じんもんこん2022（2022 年12 月10 日発表）

お問合せ先：
公立はこだて未来大学 システム情報科学部
情報アーキテクチャ学科 准教授 寺沢 憲吾
kterasaw@fun.ac.jp