CH-013

古文書翻刻初学者に向けた 文字認識処理と単語辞書を用いた翻刻候補文字列の提示手法
For Beginners in Old Documents a Method of Presenting Candidate Character Strings for Decipherment Using Character Recognition Process and Character String Dictionary

片山 歩希[1)]　　松尾 賢一[1)]
Ayuki Katayama　Ken-ichi Matsuo

1　はじめに

日本には過去の歴史資料として，約 10 億冊の古文書・古記録が残っている.[1] この古文書は歴史上の出来事や日常的な事務処理などに対する歴史認識における重要な文献資料の一種である．しかし，古文書を解読できる人々 (以降，翻刻者) の数は日本人口の 0.01%にも満たない.

古文書の翻刻者が少ない理由は，古文書中のひらがなを字母である漢字のくずし字で書いてあることや文と文がつながっていることなどによる古文書自体の判読性の低さがあげられる.

これに対して，翻刻初学者のために現在，漢字のくずし字を学ぶツール「KuLA」[2] やくずし字解読サービス「みんなで翻刻」[3] などが提案・開発されている.「みんなで翻刻」は，デジタルデータになっている歴史文献資料（古文書）を市民参加型で翻刻をするサービスである．その中のツールの一つに，くずし字認識 AI（単一文字認識処理）[4] がある．くずし字認識 AI は，デジタルデータ上の翻刻初学者の判読性が低い単一文字パターンに対して，現代文字を類似度順に提示する.

このように既存のツールは，単一文字レベルでの翻刻を支援しており，単語レベルの翻刻には文字列分の各単一文字画像をその都度入力し，単語であるかの判断は翻刻初学者にゆだねられていた.

しかしながら，古文書を翻刻するには，単一文字レベルではなく，単語レベルの意味理解が重要であるにも関わらず，単語レベルでの支援ができるツールやシステムは存在せず，単語の画像セットなども現時点では公開されていない.

本研究では，翻刻初学者に対して，単一文字レベルの支援ではなく単語レベルの支援するために，既存の単一文字認識処理を用いながら，後処理として文字列の生成や単語辞書による比較処理をする手法を提案する．これらの処理により，翻刻初学者に対して単語レベルでの提示を可能にし，古文書に対する翻刻初学者への翻刻支援を目指す.

本手法の有効性は自らが用意したテストデータセットを提案手法に入力することで，提示される第 n 位までの順位と提示される第 10 位までの単語候補の中に含まれるかを判断する提示率を明らかにする.

2　提案手法

本研究の提案手法は，翻刻初学者に対して単語レベルでの提示を可能にするために，まず翻刻初学者の判読性が低い文字列分の単一画像に対して，単一文字認識処理をかけることで，文字列分の現代文字の候補を取得する．次に，取得した現代文字の候補を網羅的に連結させることで文字列を生成する．最後に生成した文字列と古文書の単語辞書を比較することで，意味を持つ文字列である単語を提示する.

以上のことを実現するために，提案手法の全体像を次の図 1 で示し，より具体的な内容と定義を以下で述べる.

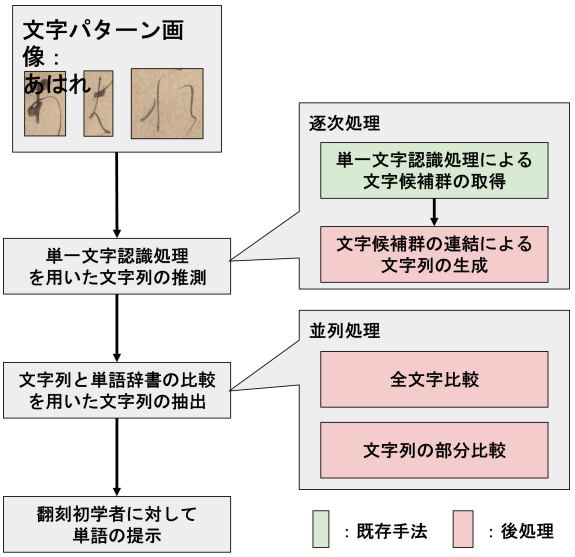

図 1　提案手法の全体像

翻刻初学者の判読性が低い文字列分の単一文字画像を文字パターン画像と定義する．文字パターン画像を単一文字認識処理にかけることで各単一文字画像に類似する現代文字の複数の候補を取得する．このとき，現代文字の複数の候補を文字候補群と定義する．単一文字認識処理だけでは，翻刻初学者に対して文字レベルの提示しかできない.

単語レベルでの提示のために，文字パターン画像の文字候補群をそれぞれ網羅的に連結させることで文字列を生成する．生成された文字列は文字列群と定義され，文字列群には単語的に意味を持つ文字列と単語的に意味を持たない文字列が存在する．翻刻初学者に対して単語レベルでの提示をするには，単語的に意味を持つ文字列を文字列群から抽出したい.

そこで，文字列群と単語辞書を比較することで単語的に意味を持つ文字列を抽出する．最後に，単語的に意味を持つ文字列を翻刻初学者に対して単語として提示する.

1)　奈良工業高等専門学校 National Institute of Technology, Nara College

以下に，各処理の技術に関する説明を述べる．

2.1　単一文字認識処理を用いた文字列の推測

翻刻初学者に対して単語レベルでの提示するために，文字列を生成する．そこで，単一文字認識処理によって，文字列を生成するための文字候補群を取得したい．

文字候補群は，単一文字認識処理によって取得される各文字パターン画像の現代文字の複数の候補である．次に，文字候補群をそれぞれ連結させることで文字列を生成する．

単一文字認識処理のデータセットは「みんなで翻刻」のくずし字 AI（単一文字認識処理）で実際に利用された CODH （人文学オープンデータ共同利用センター）が公開する「日本古典籍くずし字データセット」[5] を用いて機械学習で自作する．

自作した単一文字認識処理を用いることで，次の 2 つのステップで単一文字認識処理から文字列の生成が実現できる．

- 単一文字認識処理による文字候補群の取得
- 文字候補群の連結による文字列生成

2.1.1　単一文字認識処理による文字候補群の取得

単一文字認識処理によって文字候補群の取得をする．図 2 のように，入力された各単一文字画像に対しての現代文字の文字候補を類似度順に取得することができ，文字パターン画像分の文字候補を文字候補群として扱う．

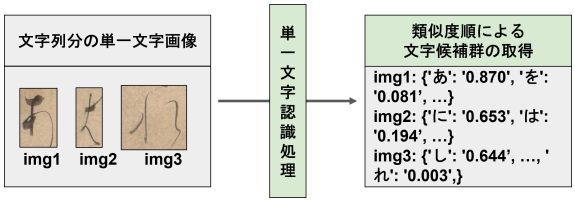

図 2　文字候補群取得の流れ

2.1.2　文字候補群の連結による文字列の生成

単一文字認識処理では，単一文字レベルでしか翻刻初学者に提示することができず，翻刻初学者にとって単語レベルでの判断ができない．そこで，網羅的に各現代文字の候補群を連結させることで文字同士の前後関係を疑似的に発生させる．文字候補群の連結による文字列の生成は，図 3 のように 2.1.1 節で取得できた各単一文字画像の現代文字の候補群に対して，網羅的に連結させることで文字列を生成する．

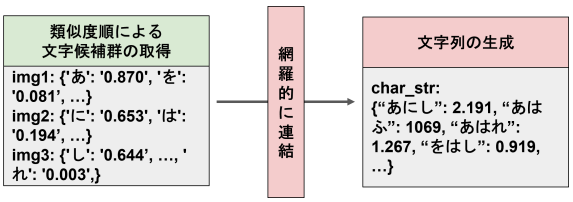

図 3　文字列の生成

生成された文字列は網羅的に連結させているために単語的に意味を持つ文字列と単語的に意味を持たない文字列が生成される．この二つの文字列を文字列群とする．

2.2　文字列と単語辞書の比較を用いた単語の生成

2.1.2 節で生成された文字列群から単語的に意味を持つ文字列のみを翻刻初学者に提示するために，生成された文字列と単語辞書を比較することで単語的に意味を持つ文字列を抽出する．

単語辞書と比較する方法は，図 4 のように次の 2 つのステップで実現できる．また，利用する単語辞書は東京堂出版の「くずし字用例辞典　普及版」[6] を利用する．

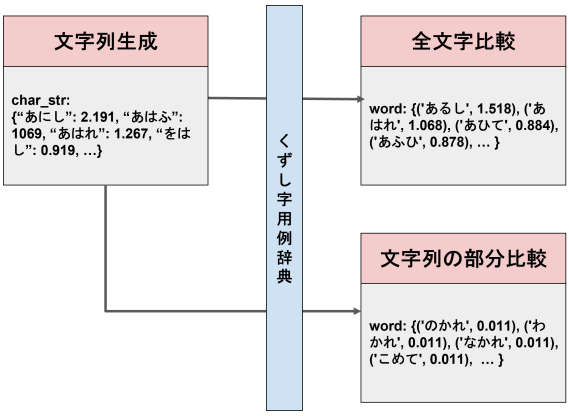

図 4　文字列生成の流れ

2.2.1　全文字比較

2.1.2 節で生成された文字列群と古文書の単語辞書を比較し，単語辞書に存在する文字列を抽出することを全文字比較と定義する．全文字比較の探索方法は探す文字列が単語辞書にヒットした文字列は単語的に意味をもつ文字列として保持し，単語辞書の最後まで探索する．

2.2.2　文字列の部分比較

2.1.2 節で生成された文字列群は，2.1.1 節の単一文字認識処理の候補から構成されているために，単一文字認識処理による誤認識が発生すると正解の現代文字が含まれない文字列群が生成される可能性がある．

そこで，文字列の部分比較は，2.1.1 節で取得した文字候補群に正解の現代文字が存在しない場合を考慮して，文字列の一部の文字を隠しながら探索することで文字認識処理による誤認識を補う．

隠す文字は，生成された文字列の中で，単一文字の類似度が式 1 で得られる文字列の平均類似度より低い文字で単語の候補を探索する．探索方法は探す文字列が単語辞書にヒットした文字列は単語的に意味をもつ文字列として保持し，単語辞書の最後まで探索する．

$$\text{文字列の平均類似度} = \frac{\text{各単一文字の類似度の総和}}{\text{単一文字の文字数}} \quad (1)$$

例えば，文字列 abc の各単一文字の類似度が a=0.9, b=0.6, c=0.7 のとき, 次の式 2 になる．

$$\begin{aligned}\text{文字列の平均類似度} &= (0.9 + 0.6 + 0.7)/3 \\ &= 0.733 \end{aligned} \quad (2)$$

このとき，文字列の平均類似度より低い類似度である文字 b,c で単語群の候補を取得する．

2.3　翻刻初学者に対して推論した単語の提示

本処理は，2.2 節によって生成された単語的に意味を持つ文字列の 2 つの候補を組み合わせることで提示する．提示方法は，どちらにも含まれている文字列は誤認識である可能性が低いので順位を上げるために各文字列の類似度を足し合わせる．

残りの文字列の候補は各文字列の類似度順で並び替える．

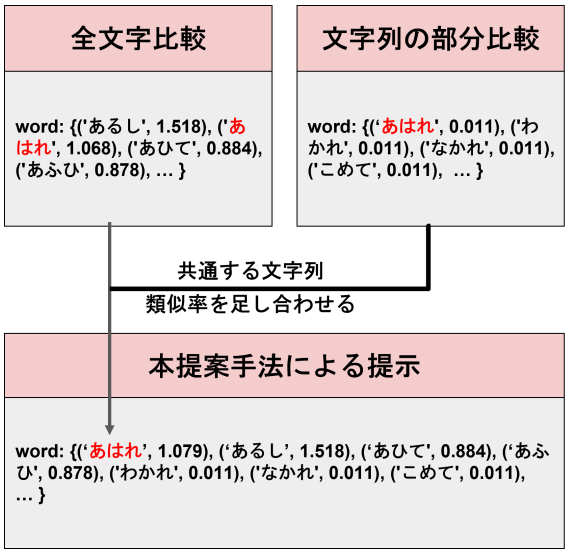

図 5　文字列における候補の出力方法

3　提示手法の実験方法と結果

本章では，提示手法を実現するための実験方法と結果について述べる．

3.1　テストデータセット

テストデータセットは CODH が公開する日本古典籍データセット [5] から「源氏物語 あけまき（総角）」のデジタルデータを利用する．図 6 のようにデジタルデータである源氏物語の文献からひらがなで書かれている単一文字画像に対して，手動で画像切り抜きし，1 単語分が 1 セットになるデータセットを 115 セット用意した．

3.2　自作する単一文字認識処理の実験方法と結果

本節では，2 章 1 節を実現するための実験方法について述べる．自作する単一文字認識処理は，機械学習の中で画像分類のアーキテクチャである ResNet34[7] を利用する．機械学習の方法は，torchvision で提供される学習済みの ResNet34 に対して，ファインチューニングする．

ファインチューニング用の学習データは 64x64 ピクセルに正規化した 49 クラスのひらがな文字画像 675,263 枚を利用する．訓練データは学習データの 70%の画像を利用し，学習におけるパラメータはエポック数 50，バッチサイズ 64 とする．検証データは，学習データの 30%を利用して，結果は約 95%の認識率を示した．

3.3　テストデータセットによる単一文字認識処理の結果

本節では，3.2 節で自作した単一文字認識処理に対して，テストデータセットである文字パターン画像分の 360 文字の単一文字画像を入力した結果を述べる．

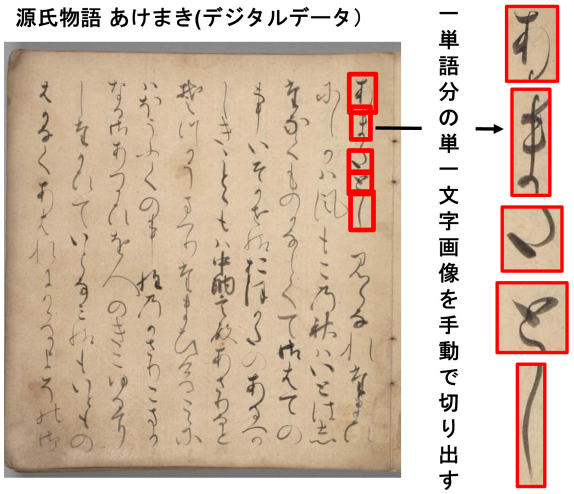

図 6　データセット作成の流れ

テストデータセットによる自作した単一文字認識処理の結果は，文字パターン画像分の 360 文字に対して，281 文字であった．このことから，認識率は約 78%となり，3 章 2 節における認識率 95%から激しく落ちた．

3.4　テストデータセットによる提示手法の結果

実験によって得られる成果は，自らが用意したテストデータセットを提案手法に入力することで，提示される第 10 位までの単語的に意味をもつ文字列の中に含まれるかを判断する提示率を明らかにする．

本提示手法で得られる提示率は以下の 2 つである．

- 再現した単一文字認識処理による翻刻初学者への提示率
- 単一文字認識処理後に文字列辞書を用いた場合による翻刻初学者への提示率

再現した単一文字認識による提示では文字候補が提示され，文字列辞書を用いた場合では，単語的に意味をもつ文字列が提示される．各提示の順位は類似度を昇順に出力し，上位 10 位まで翻刻初学者に提示する．結果比較では，各候補の出力の平均順位と提示率から評価する．

3.4.1　各処理における平均順位と提示率の結果

単一文字認識処理と単語辞書を用いた場合の処理に対してテストデータ 115 セットを入力したときの各候補の平均順位と提示率の結果を以下に示す．

表 1　単一文字認識処理と単語辞書を用いた場合の結果

	単一文字認識処理	単語辞書を用いた場合の結果
平均順位	3.5 位	3.9 位
提示率	40.8%(47/115 セット)	56.5%(65/115 セット)

結果より，単一文字認識処理による提示率は 40.8%となり，提示される平均順位は 3.5 位（総順位 168 位/47 セット）であった．また，単語辞書を用いた場合の結果は 56.5%となり，提示される平均順位は 3.9 位（総順位 257 位/65 セット）であった．

4　考察

本稿では，3 章の結果を元に提案手法の考察する．

4.1　単一文字認識処理の認識率

3 章 3 節の結果より，自作した単一文字認識処理に対して，テストデータセットの認識率は約 78%となり，CODH が公開するデータセットの認識率は 95%である．

このような自作のテストデータセットによる認識率が 17%に落ち込んだ理由として，考えられるのは図 7 のような単一文字画像の切り取り方の問題があげられる．

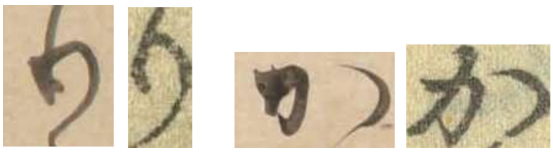

図 7　自作と CODH が公開する画像

自作したデータセットでは認識する単一文字に対して余白が多い画像であるが，CODH が公開する画像は認識する単一文字に対して余白が少ない画像である．このことから，どちらの画像も正規化をするが自作したデータセットの画像は情報量が欠落するので，十分な認識結果を得られなかったと考える．

また別の理由として，図 8 のように一つのひらがなクラスに対して複数の字母が存在するために，学習の時点で一部のクラスのみが落ち込んでいる可能性があるが，3 章 2 節の結果からは発見できなかった．

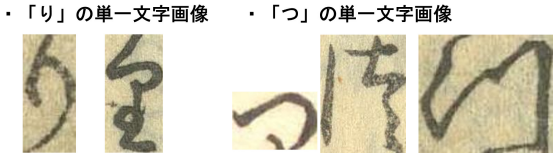

図 8　文字クラスに対する複数の字母

4.2　単語辞書を用いた場合による単語の提示率と平均順位

本実験では，3 章 2 節に作成した文字認識処理の認識率が落ち込んでしまったために，本提案手法が CODH の公開するデータセットで学習した文字認識処理の提示率 95%に達しなかった．

しかし，事前に用意したテストデータセットに対しての文字認識処理の提示率が表 1 より 40.8%に対し，単語辞書を用いた場合による単語の提示率は 56.5%の結果を示し，15.7%向上した．平均順位では，表 1 より，文字認識処理の平均順位が 3.5 位に対し，単語辞書を用いた場合による平均順位は 3.9 位の結果となった．

このことから，本提案手法は単語レベルでの提示を可能とするために単語辞書を用いたことで，単一文字レベルから単語レベルで提示することが可能となり，さらに単一文字レベルよりも提示率の優位性が見られた．

しかし，提示される平均順位は単一文字レベルが単語レベルよりも低いため，平均順位の優位性は見られなかった．

5　おわりに

本研究では，翻刻初学者に対して従来の単一文字レベルでの提示ではなく，単語レベルでの提示を可能にし，古文書に対する翻刻初学者への翻刻支援を目指すことを目的とした．

本研究の提案手法は，従来の単一文字認識処理に加えて，後処理として文字列の生成や単語辞書による比較処理による単語レベルでの翻刻初学者への提示である．また，本提案手法の優位性として，単語辞書を用いることで必ず単語的に意味を持つ文字列を翻刻初学者に対して提示する．さらに，単一文字ではなく，文字列で見るために単一文字認識処理による誤認識でも提示率を保つことや提示による候補の順位を下げないと考えた．

本研究の実験結果から，本提案手法による提示率の優位性は単一文字認識処理の提示率よりも高いためにあった．しかし，候補の順位を下げないと考えた優位性は単一文字認識処理の提示率よりも高いためになかった．

本研究の課題としては，本事件では単一文字認識処理の認識率がテストデータセットの入力による結果は低下した．その理由として，用意したテストデータセットの単一文字画像の余白が多いことがあげられる．

単一文字認識処理に入力する前に前処理として，単一文字を画像から検出する処理や余白を削減する処理を入れることで単一文字認識処理の認識率が低下することを防げると考える．

参考文献

[1] 橋本雄太. Ai 文字認識とクラウドソーシングを組み合わせた歴史資料の大規模テキスト化. 2020.

[2] 雄太 橋本, 行雄 久田, 知世 有澤, ダニエル 小林ベター, and 洋一 飯倉. くずし字学習支援アプリケーションの開発. Technical Report 5, 京都大学大学院文学研究科博士後期課程, 大阪大学大学院文学研究科博士後期課程, 大阪大学大学院文学研究科博士後期課程, 大阪大学大学院文学研究科博士後期課程, 大阪大学大学院文学研究科, may 2016.

[3] 雄太 橋本. 市民参加と ai—「みんなで翻刻」開発者の立場から.

[4] 北本朝展，カラーヌワット・タリン，宮崎智，山本和明. 文字データの分析——機械学習によるくずし字認識の可能性とそのインパクト——. **電子情報通信学会誌**, 102(6), 2019-06.

[5] 『日本古典籍くずし字データセット』（国文研所蔵／ codh 加工） doi:10.20676/00000340.

[6] 児玉 幸多 編. くずし字用例辞典　普及版. **東京堂出版**, pages 1237–1294, 1998.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.