近世古記録翻刻における視覚的曖昧性の解消：Swin Transformer と BERT を統合したマルチモーダル事後補正

吉賀 夏子
大阪大学大学院人文学研究科
yoshiga.natsuko.hmt@osaka-u.ac.jp

概要

近世古記録の翻刻において，字形が酷似する同形異字の判別は依然として課題である．本研究では，小城藩日記を対象に Swin Transformer を用いた画像分類モデルを構築し，視覚情報のみによる認識の限界を分析した．実験の結果，正解率 84.3%を達成したが，構造的曖昧性による精度の壁を確認した．この解決に向け，本稿では視覚スコアと BERT による文脈情報および固有表現情報を統合した二段階修正モデルを提案する．なお，本稿では基礎的検討として，まず視覚と文脈の統合効果に焦点を当てた評価を行う．検証の結果，視覚情報のみでは判別困難な誤りのうち，27.27%が文脈により修正可能であることを示した．本手法は，資源が限られる地域資料のデジタル化における現実的な精度改善策を提示するものである．

1 はじめに

大規模言語モデル (LLM) や高性能 OCR の台頭により，歴史的資料のデジタル化は加速している．特に NDL OCR 等の汎用モデルは，一般的な資料に対して極めて有効である．しかし，地方行政文書などの古記録 (Historical Manuscripts) においては，くずし字の特異性や地域固有の語彙，さらには候文 (*Sourou-bun*) 特有の文法構造により，汎用モデルでは認識誤りが頻発する傾向にある [3].

この課題に対し，本稿ではまず，Swin Transformer を用いたドメイン特化型の画像分類モデルを構築し，特定筆者のくずし字に対する認識性能を検証する．しかし，実験を通じて，字形のみに依存したアプローチには原理的な限界が存在することが明らかとなった．具体的には，八とハのように，特定の書き手において形状が完全に重複する構造的曖昧性を持つ文字ペアの存在である．

本研究では，この視覚的曖昧性を解消するために，画像認識スコアと言語モデル (BERT) による文脈情報を統合したマルチモーダル事後補正アプローチを提案する．本研究の主たる貢献は以下の 2 点である．

1. 一文字画像正解データの構築：特定筆者の書き癖や曖昧性を詳細に分析するため，文字単位の領域情報を付与した高精度な正解データセットを構築した．
2. 視覚と文脈の統合パイプラインの提案：画像分類モデルが出力する上位候補に対し，文脈情報とドメイン知識を用いて再順位付け (Re-ranking) を行う手法を確立した．

2 データセットの構築

本研究では，機械学習モデルがドメイン固有の知識を効率的に学習・評価するために不可欠な，文字レベルで構造化された正解データセットを構築した．

2.1 データソースと収集プロセス

基盤となるデータには，『小城藩日記データベース』（佐賀大学地域学歴史文化研究センター）[1] の翻刻成果を採用した．これは IIIF（International Image Interoperability Framework）準拠システムである『kuzushiji.work [2]』上で，専門家および市民によって翻刻された信頼性の高いデータである [4]．また，人文学オープンデータ共同利用センター (CODH) が公開する日本古典籍くずし字データセット [2] も適宜参照した．

本研究では，これらの翻刻済み画像に対し，江戸期地域資料のくずし字に詳しいボランティアによる厳密なアノテーション作業を実施した（付録図 4）．具体的には，翻刻テキストの一文字ずつに対して正確なバウンディングボックスを付与し，画像と翻刻結果が 1 対 1 で対応する正解データとして整備した．構築した

1） https://crch.dl.saga-u.ac.jp/nikki/
2） https://kuzushiji.work/

表 1　文字認識モデルのアーキテクチャ比較

比較項目	ResNet	ViT	**Swin (本研究)**
特徴抽出	局所的 (CNN)	大域的 (Self-Attn)	**階層的 (Swin)**
計算量	$O(N)$	$O(N^2)$	$O(N)$
空間的連続性	高い	低い (独立)	**高い (窓移動)**

表 2　モデルサイズによる精度の比較

モデル	パラメータ数	Top-1 正解率
Swin-Tiny	28M	84.0%
Swin-Base	88M	84.3%

データセットは[3)]，文字単位で切り出された約 28,000 枚の一文字画像を含んでおり，近世業務記録で用いられた御家流書体の具体例を提示している．

3　視覚情報のみによる文字認識とその限界

3.1　認識結果：モデル規模拡大による検証

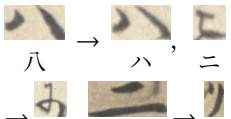
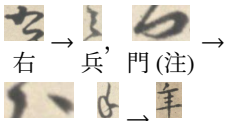
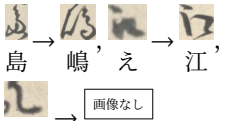
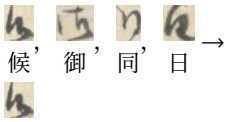
視覚情報の認識には，階層型 Transformer 構造を持つ Swin Transformer[1] を採用した．本モデルの採択理由は，畳み込みニューラルネットワーク（CNN）の代表格である ResNet が持つ局所的な特徴抽出能力と，Vision Transformer (ViT) が持つ大域的な文脈把握能力を階層的なパッチ統合によって両立している点にある．特に，Shifted Window 機構により，計算量を画像サイズに対して線形に抑えつつ，隣接するパッチ間の空間的連続性を保持できる点は，微細な筆致と文字全体の骨格の両方が重要な認識手がかりとなるくずし字認識において極めて有効である．

本研究ではまず，ベースラインとして軽量モデル (Swin-Tiny) を用いて学習を行った結果，検証データにおける Top-1 正解率は 84.0% であった．次に，モデルの表現力不足が精度のボトルネックになっている可能性を検証するため，パラメータ数を約 3 倍に拡張したモデル (Swin-Base) にて再学習を行った．表 1 に各アーキテクチャの特性比較を，表 2 に規模拡大による比較結果を示す．

表 2 に示す通り，モデルサイズを大幅に拡大したにもかかわらず，正解率の向上はわずか 0.3 ポイントに留まった．この結果は，単なるモデルの規模拡大だけでは本タスクの精度を大幅に改善することは困難であることを示唆している．すなわち，くずし字には画像情報のみでは判別が極めて困難な構造的曖昧性が含まれており，これが視覚的な認識精度の壁となっていると考えられる．

3)　https://huggingface.co/datasets/DimV-Ai/kuzushiji-character-dataset-ogihan-v1

表 3　誤認識パターンの分類と事例

カテゴリ	主な要因	代表的なペア（正解 → 予測）
構造的重複	字形の完全な一致	 八 → ハ，ニ → に，二 → リ
字形類似	崩し方や外形の酷似	 右 → 兵，門 (注) → 衛，手 → 年
異体字・変体仮名	歴史的表記の揺れ	 島 → 嶋，え → 江，て → 而
高頻度語の略化	極端な筆画の省略	 候，御，同，日 → 候

（注）  ←「左衛門」等の人名における連綿により，一文字単位の切り出しが困難な事例を含む．

3.2　誤り分析：認識の非ランダム性と構造的要因

視覚モデルによる認識失敗の傾向を詳細に分析したところ，誤認識は全文字に一様に分布しているのではなく，特定の文字ペアに著しく集中していることが明らかになった．分析対象とした全 847 件の誤りのうち，上位 20 種類の誤答パターンが全体の 27.39%（232 件）を占めており，これは認識誤りが「形状の重複」や「歴史的表記の揺れ」といった特定の要因に起因する非ランダムな現象であることを示唆している．

表 3 に，本研究で定義した 4 つの誤認識カテゴリと，それに対応する具体的な誤答ペアを示す．

各カテゴリの特性は以下の通りである．第一に「構造的重複」は，筆致が完全に一致し，字形情報のみでは人間であっても判別が不可能なケースを指す．特に助詞の「ニ」と平仮名の「に」の混同は 104 件と本実験で最多の誤答数となっており，画像情報のみに基づく認識の限界を象徴している．第二に「字形類似」

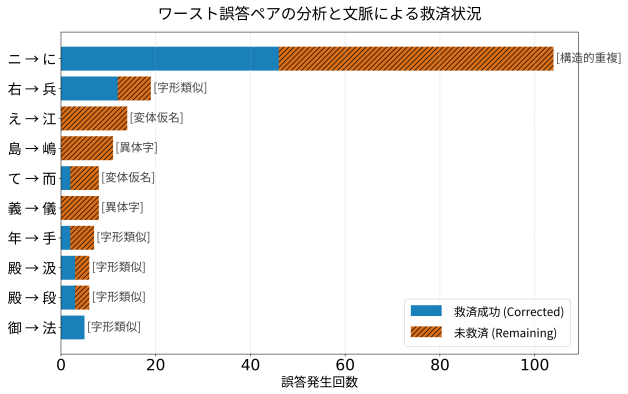


図 1　ワースト誤答ペアの分析と文脈による救済状況．棒グラフの長さは総発生回数を，青色部分は提案手法により正解に修正された件数を示す．

は，崩し方や画構成が酷似しているために OCR が外形を誤認するケースである．第三に「異体字・変体仮名」は，同一字種における歴史的な表記バリエーションに起因する．第四に「高頻度語の略化」は，候文において頻出する「候」や「御」などが極端に簡略化され，画像的特徴が極めて乏しくなるケースである．

特に人名（「左衛門」等）に頻出する「衛門」の表記において，「門」が「衛」と識別されるケースが散見された．これは，筆者がこれらの文字をひと繋がりのストローク（連綿）として記述するため，一文字単位の領域切り出し（Segmentation）において物理的な境界が曖昧になり，後続の「門」の一部が「衛」の終筆として，あるいはその逆として認識モデルに解釈されることに起因すると考えられる．このような連綿に由来する誤認識は，単一文字の形状認識の枠組みを超えた，構造的な課題であるといえる．

図 1 は，ワースト誤答ペアに対する文脈補正の効果を可視化したものである．

図 1 が示す通り，視覚情報のみでは判別が困難な「構造的重複」や「字形類似」の誤りの多くが，Stage 2 における言語モデルの事後補正によって正解へと導かれている．以上の分析から，認識精度のさらなる向上には，個別の文字形状を精緻化するアプローチよりも，文脈情報やドメイン知識を統合するマルチモーダルな補正が不可欠であると結論付けられる．

4　提案手法：マルチモーダル二段階修正

小規模なドメイン特化データセットを用いて最大の補正効果を得るため，本研究では図 2 に示すマルチモーダル二段階修正モデルを提案する．

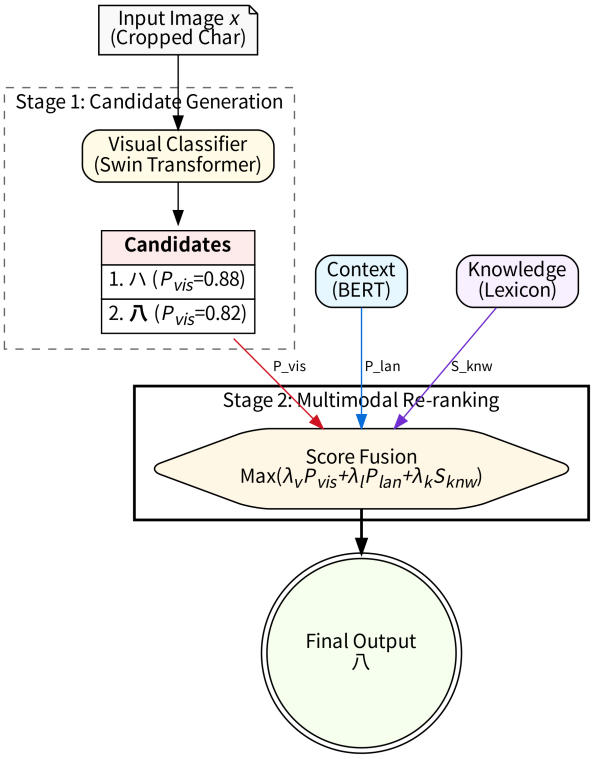


図 2　提案手法の全体概要図．あらかじめ切り出された文字画像に対し，Stage 1 で視覚特徴に基づく候補生成を行い，Stage 2 で視覚・文脈・知識情報を統合したマルチモーダル補正を行うプロセスを示す．

4.1　Stage 1: 候補生成

第一段階では，入力された一文字画像 x に対し，Swin Transformer を用いて文字候補のリストを生成する．本段階では，単一の認識結果のみならず，上位 k 個の候補文字集合 $C = \{c_1, c_2, \ldots, c_k\}$ とその確信度 $P_{vis}(c_i|x)$ を出力させる．

4.2　Stage 2: 候補再順位付け

Stage 1 で識別された文字候補に対し，以下の評価関数 $Score(c_i)$ を最大化する文字 $\hat{c}$ を最終出力とする．

$$\hat{c} = \underset{c_i \in C}{\operatorname{argmax}} \left(\lambda_v P_{vis}(c_i|x) + \lambda_l P_{lan}(c_i|T) + \lambda_k S_{knw}(c_i) \right) \tag{1}$$

ここで，P_{vis} は視覚スコア，P_{lan} は言語スコア，S_{knw} は人名・地名など固有表現に基づく知識スコアである．言語スコア P_{lan} の算出には，事前学習済みモデルとして cl-tohoku/bert-base-japanese-v3 を用いてファインチューニングを行っている．なお，本実験においては，深層学習モデルによる視覚・文脈統合の効果を純粋に検証するため，知識スコアの重みを $\lambda_k = 0$

表 4　誤り訂正実験の結果（厳しい条件下）

指標	件数	割合
対象誤り総数	847	100.00%
リカバリー成功（訂正）	231	**27.27%**
リカバリー失敗（未訂正）	616	72.73%

表 5　提案手法による救済事例のバリエーション

カテゴリ	文脈（抜粋）	修正内容	P_{lan}(正解)	P_{lan}(誤り)
人名 (Name)	留守八郎 [MASK] 衛門	兵 → **左**	0.994	0.000
敬称 (Honorific)	千鶴 [MASK] 御事	孫 → **様**	0.830	0.000
助動詞 (Auxiliary)	御膳 [MASK] 差上候付	々 → **被**	0.854	0.000
助詞 (Particle)	御屋敷之義 [MASK] 付	二 → **ニ**	0.775	0.001
動詞 (Verb)	御鷹三連被進 [MASK] 事	之 → **候**	0.611	0.005

と設定し，辞書情報は使用していない．

5　実験と考察：文脈情報の効果

5.1　定量評価：文脈によるリカバリー率

提案手法の効果を検証するため，OCR が誤認識した文字に高い確信度 ($P_{vis} = 0.8$) を，正解に低い確信度 ($P_{vis} = 0.4$) を仮定する厳しい条件 (Strict Condition) を設け，計 847 箇所の誤り箇所を対象に実験を行った．なお，前述の通り本実験では固有表現辞書 S_{knw} は適用せず，Swin Transformer と BERT の統合のみでどこまで補正が可能かを評価した．

表 4 に示す通り，提案手法は全誤りの 27.27% を正解へと修正した．これは，モデルの規模拡大による精度向上がわずか 0.3 ポイントに留まった Stage 1 の結果（表 2）と比較して対照的である．すなわち，画像認識のみでは突破困難であった構造的曖昧性の壁が，BERT による言語文脈の導入によって効果的に補完され，認識精度が実質的に底上げされたことを意味している．

5.2　定性評価とケーススタディ

救済事例を分析したところ，単なる共起を超えた，多角的な言語特徴が補正に寄与していることが確認された（表 5）．

具体的には，人名パターン（「左衛門」等）や敬称（「様」）において，BERT が先行する文字列から後続する語彙のカテゴリを強く規定し，視覚的には確信度が低かった正解候補の順位を引き上げている．図 3 のスコア変動が示す通り，特に (a) の人名事例では，視覚スコアでは「兵」が優勢であったが，文脈導入後は「左」の連接確率が高まり，逆転現象が起きている．

また，助詞の「ニ」と「に」の判別においても，先行する名詞（「屋敷之義」）との接続関係や，候文特有の定型句としての機能をモデルが検知している点は注目に値する．一方で，依然として修正が困難な事例として，文脈の後半が欠落している場合や，複数の助詞が連続する箇所の曖昧性が挙げられる．これらは単一の言語モデルのみならず，より広範な文書構造の把握が必要であることを示唆している．

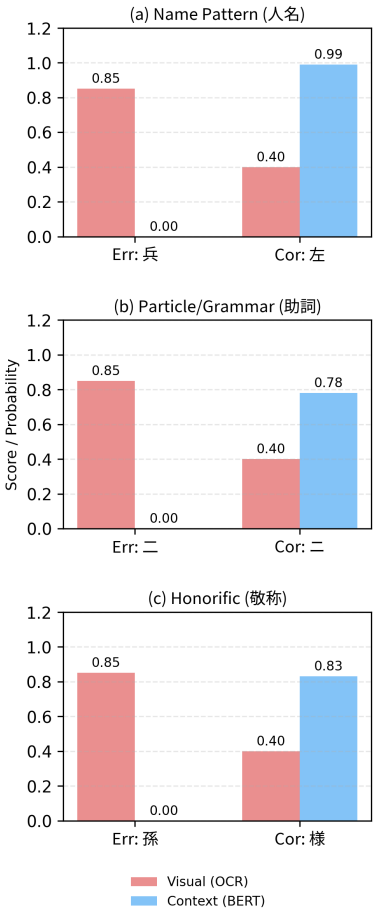


図 3　ケーススタディにおける各候補のスコア変動プロット（(a) 人名，(b) 助詞，(c) 敬称）．文脈情報の導入により正解文字の順位が逆転する様子を示す．

6　おわりに

本研究では，近世古記録翻刻における視覚的な認識限界を定量的に示し，BERT を用いた事後補正の有効性を確認した．本手法は，予算や人員が限られる地域資料のデジタル化において，小規模なデータ資源と既存の言語モデルを組み合わせた現実的な精度改善手法を提示するものである．今後の課題として，文脈による救済が困難であった残りの誤りに対し，生成モデルを用いたより広範な文脈理解や，地域資料に特化したドメイン適応の深化が挙げられる．

謝辞

本研究は JSPS 科研費（課題番号：JP22K18149）の助成を受けたものである．また，データセット構築および実験の実務は，リサーチ・アシスタントの Dimitra Vogatza 氏の多大なる貢献による．

参考文献

[1] Liu, Z., et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV, pp.10012–10022, 2021.
[2] CODH. 日本古典籍くずし字データセット, 国文研ほか所蔵／ CODH 加工, doi:10.20676/00000340.
[3] 吉賀夏子, ほか. 郷土に残存する江戸期古記録の機械可読化を目的とした市民参加および機械学習による固有表現抽出. 情報処理学会論文誌, Vol.63, No.2, pp.310–323, 2022.
[4] 吉賀夏子, 橋本雄太. 多様なくずし字画像に対応するアノテーションデータセット収集システムの試作. 研究報告人文科学とコンピュータ（CH）, Vol.2023-CH-131, pp.1–8, 2023.
[5] He, K., et al. Deep Residual Learning for Image Recognition. CVPR, pp.770–778, 2016.
[6] Kolesnikov, A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR, 2021.

7　付録

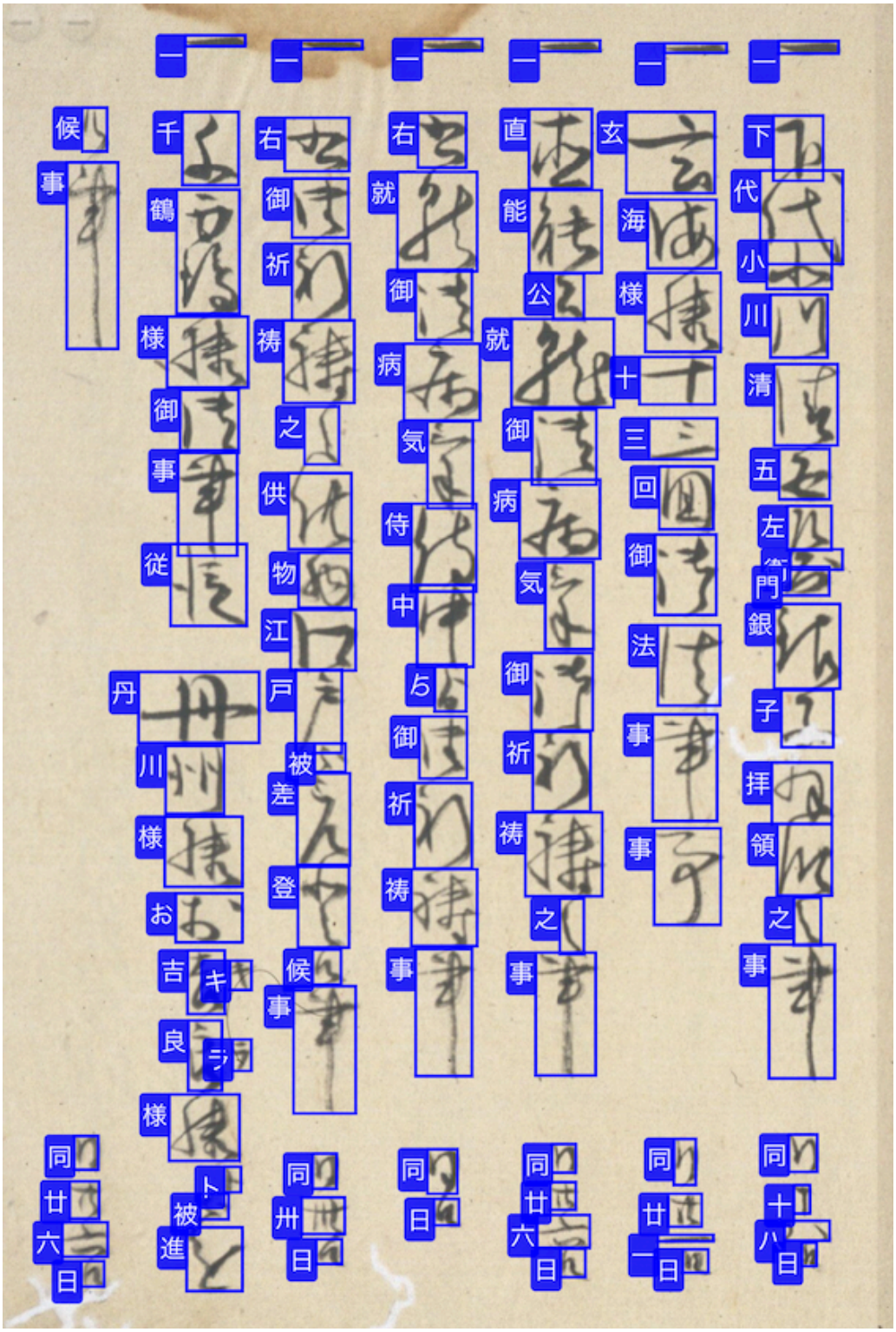


図 4　kuzushiji.work でアノテーションされた『日記目録』の例（佐賀大学地域学歴史文化研究センター）．『日記目録』とは，『日記』と呼ばれる江戸時代の諸藩や武家等で作成された日々の業務記録をさらに箇条書きの目録に整理したものであり，およそ 120 年分が現在も紙資料および『小城藩日記データベース』に蓄積されている．