N-013

近代公文書データセットと日本古典籍くずし字データセットを用いた手書き文字認識に関する比較

Comparison of Handwritten Character Features in Modern Official Document Dataset and Kuzushiji Dataset

宮川 裕貴† 山田 雅之† 中 貴俊† 兼松 篤子†
Yuki Miyagawa Masashi Yamada Takatoshi Naka Atsuko Kanematsu
宮崎 慎也† 長谷川 純一‡
Shinya Miyazaki Junichi Hasegawa

1. はじめに

近代公文書には当時の情勢を知ることができるなど史料的価値があるが，近世古文書の流れを汲む文書のため，解読には専門知識が必要である．

我々の研究グループでは，近代公文書の自動解読システムの開発を目指し，台湾総督府文書を題材に，現在までに約 114 万の手書き文字データを有するデータセット（以降，公文書データセット）を作成した．しかし，字種ごとのサンプル数に偏りがあり，本データセットで訓練した手書き文字認識モデルでは，サンプル数の少ない字種について，認識率が低い傾向にある．サンプル数を十分に確保する方法として，他のデータセットからの文字データの流用が考えられるが，その字形の特徴を事前に把握する必要がある．

本稿では，公文書データセットと同様に日本の歴史的文書を題材としている日本古典籍くずし字データセット[1]（以降，古典籍データセット）を比較対象として取り上げ，両データセットにおける文字データの字形特徴に関する類似性を，単文字認識実験を通じて考察する．

2. 関連研究

Clauwat らは古典籍データセットから派生した KMNIST データセット[2]に対する単文字認識実験を実施しており，MNIST の文字に比べ同条件における KMNIST の文字の認識率は低いという結果を報告している[3]．

3. 近代公文書データセット

台湾総督府文書を題材にして作成している公文書データセットは，現時点で 3,847 字種 1,140,295 文字分のデータを有している．図 1 に文字のサンプルを示す．

サンプルが 1 つしかない字種は 539 あり，10 未満の字種は 1,631 ある．また，サンプル数の大きい 958 字種のサンプルが全体の 95%を占める．

図 1 公文書データセットの文字例

4. 日本古典籍くずし字データセット

古典籍データセットは，人文学オープンデータ共同利用センターが公開しており，データセットは，4,328 字種，1,086,326 文字分のデータを有する．図 2 に文字のサンプルを示す．

サンプル数が 1 の字種は 790 あり，10 未満の字種は 2,233 字種ある．また，サンプル数が大きい 784 字種のサンプルが全体の 95%を占める．

† 中京大学 Chukyo University
‡ 中京大学人工知能高等研究所 Institute for Advanced Studies in Artificial Intelligence

図 2 古典籍データセットの文字例

5. 単文字認識実験

5.1. 実験方法

両データセットにおいて，サンプル数が 100 以上の字種を対象とした実験用データセットを作成した。本データセットでは，1 字種あたりのサンプル数の上限を 350 とし，350 を超える字種については、重複なしで 350 サンプルを選択した．作成したデータセットは訓練用とテスト用に 8：2 で分割した．また，訓練時のデータ拡張には画像の切り抜き、変換画像の複数結合，アフィン変換，射影変換，色合いの変化，部分削除を適用した．モデルは ResNet101[4]を使用し，画像サイズ 32×32，バッチサイズ 1,024，最適化関数 RAdam，損失関数として交差エントロピー誤差関数を用いて訓練を実施した．

5.2. 実験結果

それぞれのデータセットで訓練したモデルの両テストデータに対する損失と正解率を表 1 と表 2 に示す．

表 1 公文書データセットによる訓練モデル

テストデータ	損失	正解率（%）
公文書	0.358	92.0
古典籍	1.811	62.4

表 2 古典籍データセットによる訓練モデル

テストデータ	損失	正解率（%）
公文書	1.931	59.5
古典籍	0.312	92.7

公文書データセットで訓練したモデルの古典籍データセットに対する正解率は 62.4%であり，古典籍データセットで訓練したモデルの公文書データセット対する正解率は 59.5%であるため，両データセットでは 60%程の文字について，似た特徴を持つと考えられる．

6. まとめ

本稿では、2 種類の歴史的文書データセットをとりあげ，それらの基礎的統計量と単文字認識実験について述べた．両者ともに総サンプル数が 100 万を超えるデータセットであるが，字種毎でサンプル数に偏りが見られた．また，単文字認識実験により，両データセットにおいて異なる特徴を持つ手書き文字が一定数存在することが判明した．

今後は字種毎の認識を通して，特徴が類似する字種を調査するとともに，両データセットを合わせて訓練したモデルによる単文字認識実験を行い，データセットを併用の効果を検証する．

謝辞

本研究は JSPS 科研費 JP20H01304、および中京大学戦略的研究「デジタル・ヒューマニティズプロジェクト：日本近代公文書自動解読システムの開発」の助成を受けた．

参考文献

[1] “日本古典籍くずし字データセット”, doi:10.20676/00000341，http://codh.rois.ac.jp/char-shape/

[2] “KMNIST データセット”，http://codh.rois.ac.jp/kmnist/

[3] Tarin Clanuwat et al. ,“Deep Learning for Classical Japanese Literature”, doi:10.20676/00000341，Neural IPS, Machine Learning for Creativity and Design Workshop（2018）

[4] Kaiming He et al. , “Deep Residual Learning for Image Recognition”, doi:10.48550/arXiv.1512.03385,CVPR, pp.770-778（2016）