2S-08

深層学習を用いた拓本の多書体認識と時空間データベースの作成

岸雅大† 鍋谷美智子‡ 野上佳奈‡ 孟林‡ 山崎勝弘‡

立命館大学大学院　理工学研究科†　　立命館大学　理工学部‡

1. はじめに

古代文献の解読、解析と知識の抽出は、歴史の整理、気候変動、および自然災害の予測対応などの研究に役立つ。図 1 に示す拓本は 1 番長い歴史を持つ重要な古代文献である。しかし、拓本は図 2 のように年代に応じて様々な字体をもち、また、拓本から潜在的な知識抽出の研究があまり存在しない。本研究では、拓本の文字を認識した上で、拓本のキーワードの時空間データベースを作成し、潜在的な知識を発見することにより、様々な分野の研究に役立てることを目標とする。

本研究では、まず、深層学習を用いて拓本内の多書体文字を認識する。次に、拓本から気候、生活、歴史などのキーワードに対して、発生場所、時間とその他の情報をデータベースに登録し、拓本の時空間データベースを作成する。さらに、キーワードの検索により、キーワードを時間順に、地図上で表示して、可視化することにより、潜在的な知識の発見に貢献する。

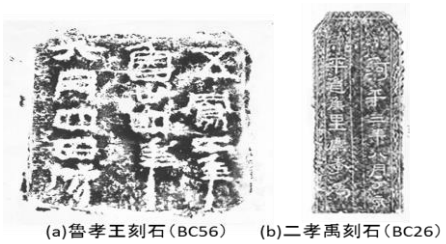

図 1　拓本の例

図 2　拓本内の書体

2. 深層学習を用いた拓本の多書体認識

2.1 拓本データセット

CNN の学習では、ネットワークを学習するための訓練画像と教師画像、学習したネットワークを評価するためのテスト画像を含む大量のデータセットを用意する必要がある。今回は図 3 のように、教師画像に各書体の画像を 1 文字ずつ、設定し、残りの画像から各書体を含む 20 枚の画像をテスト画像とし、それ以外を訓練画像としている。

今回のデータセット構築では京都大学 COE プロジェクト[1]が作成した拓本文字データベースから収集した 100 文字計 105781 枚の画像を使用する。そのうち 500 枚を教師画像、2000 枚をテスト画像、103281 枚を訓練画像とした。教師画像は、各文字についてノイズや断裂が少ない画像で 5 書体を含むように設定している。

図 3　学習用データセットの例

2.2 データ増強

CNN では、一般的に訓練データを大量に用意して学習を行うことで認識率が高くなる。そこで、原画像 1 枚に対して、図 4 のように切り取り、輝度値の変更の処理を行い、学習に使用するデータを増強する。これにより原画像 1 枚に対して、切り取り(17)×輝度値変更(3)＝51 枚の画像を得ることができる。データセット全体では 103281×51=5267331 枚の画像が得られる。

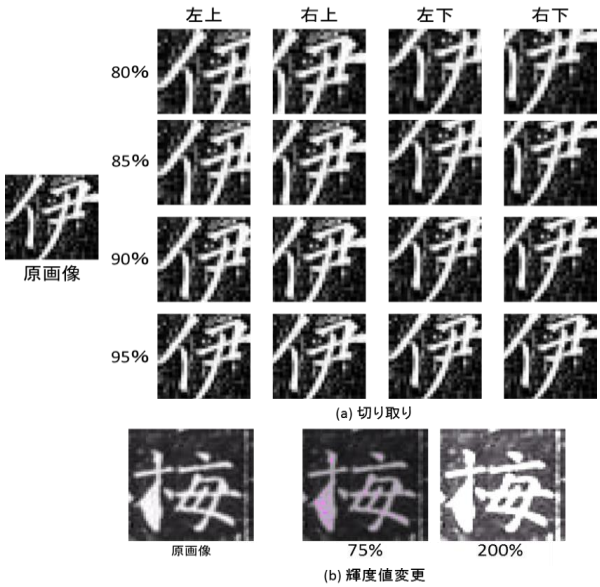

図 4　データ増強

3. 拓本時空間データベースの作成と可視化

本研究では、石刻拓本資料[2]を利用して、図 5(a)のような拓本毎に標題、出土地、年代をまとめた時空間管理番号表を作成する。次に、図 5(b)のように、出土した文字がどの標題の拓本に記されていたかを管理番号でリストアップして、キーワード管理番号表を作成する。

さらに、これら２つの表を用いて図 5(c)の様に時空間データベースを作成する。図 5(c)の情報を図 6 の様に地

Multi-typeface recognition of rubbing using deep learning and creation of spatiotemporal database, Masahiro Kishi† Michiko Nabeya‡ Kana Nogami‡ Lin Meng‡ and Katsuhiro Yamazaki‡ †Graduate School of Science and Engineering ‡Faculty of Science and Engineering, Ritsumeikan University

図上に可視化し、視覚的にわかりやすくする。

図 6 は、雨の時空間データベースを可視化したもので、多く出土している地域上位 3 位までを赤文字で示している。図 6 から、地図の左上の新疆や西藏で雨の文字が多く出土していることが分かる。また、地震という文字で時空間情報の可視化をすれば、地震が起こる周期をある程度予測できるのではないかと考えている。図 6 では 500 年毎に地図に表示しているが、実際は B.C100 年から A.D2000 年まで 100 年毎に表示する。また、登録する文字は自然災害(地震、洪水等)、気候(雨、雷等)に関連した文字を重点的に登録していく。

(a)時空間管理番号表

管理番号	標題	出土地	年代
NAN0001X	趙泰武殿前戯絞柱石孔	吉林省	趙建武4年(338)
NAN0002X	前秦鄭能修_太尉祠碑	吉林省	建元3年(367)6月
NAN0003A	秦廣武將軍口産碑	吉林省	建元4年(368)
NAN0003B	秦廣武將軍口産碑(碑陰)	吉林省	建元4年(368)
・	・	・	・

(b)キーワード管理番号表

文字種	雨	雷	雲	雪	・・・
管理番号	NAN0001X	GIS0018X	TOU1831X	GEN0267A	・
	KAN0012X	GIS0021X	TOU1832X	GEN0267B	・
	KAN0026A	NAN0101X	TOU1833X	GEN0268X	・
	KAN0029X	NAN0105X	TOU1835X	GEN0269X	
	・	・	・	・	

(c)時空間データベース

文字種	年代	出土地
雪	256	河南

文字種	年代	出土地
雷	396	山西

文字種	年代	出土地
雨	338	吉林省
	368	山東省
	399	山東省
	・	・

図 5　時空間データベース

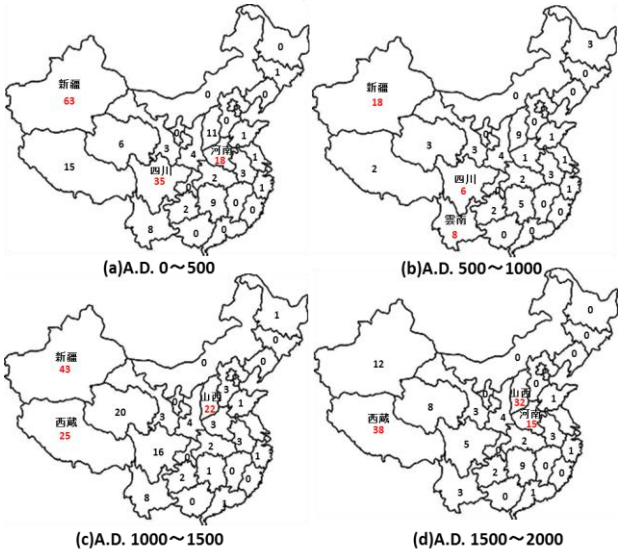

図 6　雨による時空間データベースの可視化

4.　実験

4.1　実験内容

拓本文字データベース[1]から収集し、データ増強を行った 100 文字計 5267331 枚の画像を用いて、Alexnet と GoogLenet で学習、認識を行った。また、61 文字計 3794592 枚の画像を用いて学習、認識した際の認識率の比較を行った。CNN の実装には Caffe を使用し、入力画像のサイズは 224×224、学習回数は 30epoc 行った。
実験に使用した GPU マシンは、CPU：Xeon Ef-16200 v4, GPU:GTX1080Ti(3584 コア)×4, メモリ:64GB である。

4.2　実験結果・考察

表 1 に 61 文字種と 100 文字種の認識結果を示す。また、図 7 に認識した画像と誤認識した画像を示す。今回の実験では、61 文字種の実験では GoogLenet で認識率 96%、100 文字種の実験では Alexnet で認識率 92%を達成した。

図 7 の認識した画像では、華の様にノイズにより文字が多少潰れている、毛のように文字が薄れている、氣のようにノイズが多い場合でも認識できていることが分かる。

図 7 の誤認識した画像では、寫は鳥、木は朱、牛は午、と誤認識された。ノイズの位置、大きさによって他の文字と類似してしまい、誤認識していると考えられる。

また、100 文字種では、61 文字種より認識率が低下している。文字種を追加したことにより、似ている文字が増え、誤認識をしたのではないかと考えられる。

さらに、61 種の実験に比べて、GoogLent の認識率が Alexnet より低下しており、GoogLenet の方が類似した文字をより多く誤認識している。

これらの解決策として、データ増強手法を増やす、ネットワークの構造、パラメータを最適化することがあげられる。

表 1　61 文字と 100 文字の学習時間と認識率

文字種	ネットワーク	学習時間	正解数	認識率
61	Alexnet	9時間34分	1147/1220	94%
100		13時間17分	1847/2000	92%
61	GoogLenet	1日4時間	1172/1220	96%
100		2日15時間	1782/2000	89%

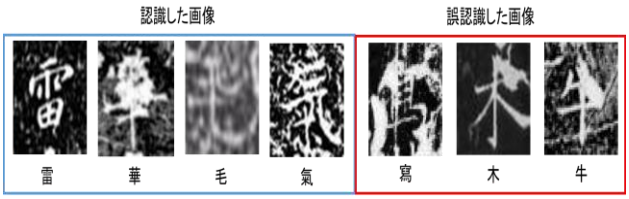

図 7　認識・誤認識した画像

5.　おわりに

本稿では、深層学習を用いた拓本内の多書体認識において、61 文字種で 96%、100 文字種で 92%の認識率を達成できることを述べた。また、雨,雷など気候に関する文字の時空間データベースを作成し、地図上に可視化した。今後の課題として、拓本内の文字の認識率向上、時空間データベースへの文字の追加、可視化した時空間データベースからの知識の抽出などが挙げられる。

参考文献

[1] 拓本文字データベース
http://coe21.zinbun.kyoto-u.ac.jp/djvuchar
[2] 京都大学人文科学研究所所蔵　石刻拓本資料
http://kanji.zinbun.kyoto-u.ac.jp/db-machine/imgsrv/takuhon/