N-002

深層学習と語彙分析を用いた高精度な拓本文字認識の実現

High-accuracy Rubbing Character Recognition by combining Deep Learning and Lexical Analysis

章 芷毓†, 陳 禹汐†,冨山宏之†, 孟林†

Zhiyu Zhang, Yuxi Chen, Hiroyuki Tomiyama, Lin Meng

1. はじめに

拓本は中国古代文献として大量な潜在的な知識が含まれ、歴史、政治、文化などの研究に対して非常に重要である。しかし、数千年の歴史を持つため、また、石碑の劣化、紛失などにより、拓本文字の認識とそれらの整理が難航している。近年、深層学習は物体認識分野において、大きな貢献を果たすことにより、古代文字の認識に貢献できると考えられる。しかし、テンプレート不足、劣化、多書体などの問題により、拓本文字の整理と解読の難しさを増している。本研究では深層学習と語彙分析を用いて、高精度な拓本文字の認識を目指し、文化遺産の保存と整理に貢献する。

詳細には、データの整理と認識の二つのプロセスで構成する。データの整理において、すでに認識された拓本文字を整理し、それらの訓練により深層学習のモデルを得られる。また、既存の拓本のすべての単語とそれらの出現頻度を計算する。認識について、認識する拓本文字画像に対して、深層学習により文字の候補文字とそれらの確信度を得る。そして、拓本文章中の単語出現頻度と深層学習の認識結果を補正することにより、高精度の認識を目指す。

2. 拓本とその文字認識

二千年前、紙のない時代は人々が動物の骨、石、金属上に重要な記事を刻んで、記録する。紙が発明されても、その習慣を継続している。従って、それらの骨、石などには、たくさんの有益な歴史情報が保存されている。これらの資料を保存する手法は拓本である。拓本は、骨、石、金属に紙を貼って、炭・インクなど塗って、記述された文章を取得する。拓本には、たくさんの歴史情報を保存しているため、研究者らは拓本の整理を目指している。京都大学は、中華民国までのやく 3000 年間の拓本を整理し、文字の自動切り取りを実現できた[1]。Meng らは、深層学習を用いて、拓本文字の認識を目指している。現在、人工知能技術が邁進し、研究者らは深層学習（Deep Learning）を用いて、くずし文字の自動認識を目指している。しかし、これらの研究は、拓本文字の認識および整理が十分と言えない。

近年、深層学習は物体認識分野において、大きな貢献を果し、古代文献の認識に貢献できると考えられる。画像処理での深層学習について、大きくセグメンテーション（領域抽出）と画像分類の二つに分けられている。拓本は、より規則的に刻まれたため、拓本での文章のセグメンテーションを行う必要がない。従って、本研究は深層学習を用いた拓本の文字認識を注目している。しかし、拓本の資源が貧弱で、つまり、すべての文字に対して、深層学習に必要な学習データが揃えていると限られていないため、拓本の語彙分析（単語の出現頻度などの情報）を用いて、文字を推測すること目指している。

† 立命館大学 大学院理工学研究科, Graduate School of Science and Engineering, Ritsumeikan University

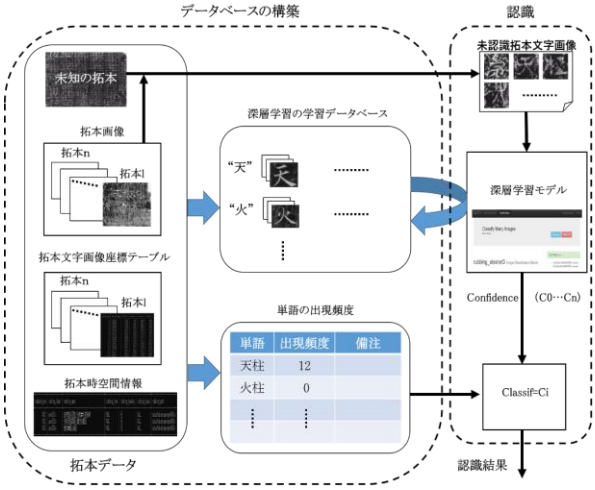

図 1 データベースの構築と文字の認識

3. 深層学習と語彙分析を用いた拓本文字認識

図 1 は本研究の全体を示し、拓本データベースの構築と、それを用いた拓本文字認識に構成されている。拓本データベースの構築では、拓本画像とその既知文字が拓本画像中の座標を記録した拓本文字画像座標テーブルを用いて、自動拓本文字画像を切り取り、深層学習の学習データベースを作成する。また、拓本文字画像座標テーブルは、拓本を文章化され、認識できなかった文字を空となる。それにより単語の出現頻度を得られる。また、拓本時空間情報を持つため、拓本の時空間情報を用いたデータの作成もできる。

3.1 深層学習を用いた拓本文字認識

本研究は深層学習を用いて拓本の文字を認識し拓本文字の確信度を取得する。さらに、ランク 1 （Top predictions）の確信度の文字に対して、閾値を用いて、確信度の高い文字と低い文字を分類する。まず、認識できなかった文字が含まれる拓本を選択し、拓本文字画像を切り取って、深層学習のモデルにより、テスト画像の認識結果とその確信率を獲得する。本研究での予備実験でよい結果を得られた AlexNet を使用する。そして、拓本文字の認識結果の確信度の閾値を決定し、ランク 1 の確信度が閾値以下の文字に対して、語彙分析を用いて、深層学習の認識結果を補正する。

3.2 語彙分析

深層学習により確信度の低い拓本文字に対して、語彙分析結果により再評価を行う。詳細について、拓本文字の所

属する年代のコーパスを利用し、深層学習の認識結果により可能であれる文字と単語に基づいて、可能性がある単語を推測する。そして、単語出現頻度を統計し、候補文字の全体確信度を計算する。式(1)により、再評価を行う。深層学習から出力した確信度を$confDL$、単語出現頻度を$confLE$と定義し、再評価された候補文字の確信度は$confidence$となる。これにより全体確信度が最も高い候補文字は最終的な認識結果と判断される。

$$confidence = confDL \times confLE \quad (1)$$

4. 実験

4.1 実験条件

本研究は京都大学拓本文字データベースの拓本[1]を用いて、拓本データベースを新たに構築し、実験を行った。図 2 は実験に使用した８文字の拓本である。この拓本は南北朝時代の拓本「北齊天柱山」の一部分で、劣化により非常に認識しにくい。わかりやすくするために、画像の右に現代文字を追加している。深層学習の学習データベースでは、テストに使用する拓本文字と字形に類似する文字が含まれる 100 種拓本文字をし、それらの 100 種文字は 45995 枚画像を学習データとして使用する。データベースの構築には、OS が ubuntu16.04 LTS で、SQL が mysql8.0.16 で、プログラミング言語が Python である。深層学習の実装には caffe を使用した。入力画像サイズは 224x224、学習率を 0.01 とし、教師画像の認識率が収束するまで学習を行う。実験に使用した GPU マシンは CPU:Xeon E5-1620v3,GPU: TITAN X(Pascal) x2,メモリ:64GB である。

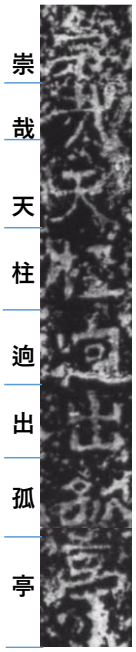

図 2 拓本例

4.2 実験結果

表１は深層学習を用いた拓本文字の認識結果である。認識対象文字に対して、上位５位と類似する文字とそれらの確信度を示す。８文字中、正解はランク１と認識された文字が６個となり、深層学習の認識有効性を示されている。そのうち、確信度は 90%を超えたが、ランク１は正解ではなかった場合もある。従って、ランク１の確信度閾値は 99%を決定し、確信度は 99%以上の５文字を確定した。残りの３文字に対して語彙分析を行った。

表 2 は「柱」「逈」「孤」３文字が南北朝のコーパスを用いて、相関性がある単語出現数量を検索し、候補文字の全体確信度を計算した語彙分析の結果である。そのうち、全体確信度が最も高い単語は「天柱」「逈出」「孤亭」（各表一番上の単語を示す）で、３文字が「柱」「逈」「孤」と再認識できた。この結果から、深層学習により確信度の低い拓本文字に対して、語彙分析の有効性を示された。

表１深層学習の認識結果

対象	認識結果				
	ランク１	ランク２	ランク３	ランク４	ランク５
崇	崇(99.40)	崇(0.46)	宗(0.1)	茅(0.0)	堂(0.0)
哉	哉(99.83)	裁(0.09)	天(0.05)	災(0.02)	戌(0.01)
天	天(100)	大(0.0)	更(0.0)	戸(0.0)	未(0.0)
柱	怒(90.45)	柱(3.75)	妃(1.89)	桂(1.49)	掾(0.92)
逈	避(66.58)	堂(7.54)	怛(6.08)	逈(2.93)	哩(2.61)
出	出(99.97)	山(0.02)	土(0.01)	之(0.0)	二(0.0)
孤	孤(67.72)	弧(32.18)	豫(0.04)	軫(0.01)	銷(0.01)
亭	亭(99.88)	戸(0.03)	粟(0.03)	茅(0.02)	宗(0.01)

表 2 語彙分析の結果

単語	数量	頻度	確信度
天柱	12	100	3.75
天妃	0	0	0
天怒	0	0	0
天桂	0	0	0
天掾	0	0	0

単語	数量	頻度	確信度
逈出	1	100	2.93
避出	0	0	0
堂出	0	0	0
怛出	0	0	0
哩出	0	0	0

単語	数量	頻度	確信度
出孤	0	0	0
出弧	0	0	0
出豫	0	0	0
出軫	0	0	0
出銷	0	0	0

単語	数量	頻度	確信度
孤亭	1	100	67.72
弧亭	0	0	0
豫亭	0	0	0
軫亭	0	0	0
銷亭	0	0	0

5. おわりに

本論文では、深層学習と語彙分析を用いて拓本文字を認識することを紹介し、その有効性を示した。評価において、有効性を示したが、８文字の拓本しかないならテストデータ量が不足で実験結果の信頼性が高くなかった。今後の課題として、深層学習の訓練データに対してデータを増強し、拓本文字の種類を増加し、深層学習の認識率を向上することが期待する。それにより、更に認識できなかった拓本文字を追加すると、この手法を利用し、高精度の認識を実現できることを目指している。

謝辞

本研究は JSPS 科研費 18K18337 の助成を受けたものです。また、データを提供して頂いた、京都大学人文科学研究所と安岡孝一先生に感謝いたします。

参考文献

[1] 拓本文字データベース，京都大学・東アジア人文情報学研究センター：coe21.zinbun.kyoto-u.ac.jp/djvuchar (2019.4.6.)

[2] L. Meng, et al.,“Ancient Asian Character Recognition for Literature Preservation and Understanding,” EuroMed2018 Int. Conf. on DIGITAL HERITAGE, 2018.

[3] L. Meng, et al.,“Recognition of Oracle Bone Inscriptions Using Deep Learning based on Data Augmentation,” 2018 IEEE Int. Conf. on Metrology for Archaeology and Cultural Heritage, 2018.