木簡画像データセット
－公開と活用－

1　はじめに

AI（人工知能）による文字認識等の研究成果の増加とともに、自由に利活用できるオープンデータセットの需要も高まっている。そこで、奈良文化財研究所では、木簡等研究資源のオープンデータ化を推進すべく、2020年11月に「スマートフォン撮影による木簡画像処理実験用データセット Ver.1.0」（以下「木簡画像データセット」）を、奈良文化財研究所学術情報リポジトリで公開した[1)]。

本稿では、木簡画像データセットの概要と作成方法、そして研究活用の事例について報告する。

2　AIによる文字認識研究

近年、人文学オープンデータ共同利用センター（CODH）と国文学研究資料館の共同研究で、くずし字に関する大規模な機械学習データセット「日本古典籍くずし字データセット」[2)]が広く世界に公開され、AIによるくずし字認識（OCR）の研究開発が促進されている。

とりわけ、2019年10月に開催されたKaggleコンペティション「くずし字認識」では、293チームが参加し、トップスコアは95％という高い認識結果が示された。

このような情勢をうけ、人文情報学を中心とする研究領域においては、機械学習を活用した文字字形認識研究の有効性が確信されるに至った。

3　木簡画像データセットの概要

公開目的　深層学習による古代手書き文字認識の研究での活用を念頭に置いて作成するとともに、幅広い分野で利活用を想定し、汎用性の高いデータ形式を採用した。

データ概要及び作成方法　本データセットは、画像変換ニューラルネットワークの学習及び提案手法の性能評価への利用のために作成・整備した。データセットは200点の保存処理済み木簡画像と、それぞれの墨痕に対するアノテーションを含む。

各画像は以下の手順で作成した。まず、保存処理済み木簡一点を白色布上に置き、スマートフォン（Android SO-02K）カメラで撮影した。撮影時の光源環境には特段の管理をおこなっていない。

次に、撮影により得られたカラー画像中の墨痕領域に手動でアノテーションを付与した。同時に、木簡が置かれている白色布の領域を背景領域として除去した。

そして、図64に例示するように、木簡のカラー画像と、付与されたアノテーションを別のレイヤーに配置し、マルチレイヤTIFF 形式の画像ファイルとして保存した。

利用条件　クリエイティブ・コモンズ 表示- 継承 4.0 国際 ライセンス（CC BY-SA）相当の条件で提供している。利用に際しては所蔵機関の許諾を必要とせず、奈良文化財研究所の所蔵史料であることを明示しさえすればよいという条件に設定し、活発な利活用を期待した。

4　データセット活用による研究成果

「木簡画像データセット」を機械学習の学習データとして活用し、木簡画像における木目等のノイズをAIで自動除去する画像処理プログラム「深層学習による木簡実測図の自動作成」を、埼玉工業大学・大山航氏を中心に開発した[3)]。ここでは、本プログラムの概要や目的、処理結果例等を簡単に紹介する。

プログラムの概要　本木簡画像処理プログラムは、深層学習技術を活用して、デジタル撮影された木簡写真から、木簡の形状と墨痕を正確に転写した木簡実測図を自動作成するものである。

開発目的　開発の目的は以下三点である。

（1）木簡整理作業の省力化に寄与し、それにより質が高く多様な視点からの資料情報収集を可能にするため。

（2）木簡に書かれた文字の視認性向上に寄与し、「図像」としての文字研究を実現するため。

（3）木簡文字自動認識の前処理として活用するため。

Webアプリケーションの公開　「深層学習による木簡実測図の自動作成」プログラムは、Webアプリケーションツールとして公開している[4)]。

このツールは、木簡実測図の自動作成処理をWebブラウザ上で実行できる。現在は、スマートフォンやタブレットで撮影した木簡写真を処理するバージョン、すでに撮影された木簡のデジタル画像をアップロードするバージョンの二通りがある。前者は、木簡整理作業中に、手元で、簡便に、処理結果を確認できるツールとして、将来的な実用化に向け、研究を継続している。

図64 「木簡画像データセット」の例

図65 処理結果例

処理結果及び評価　本プログラム（アップロード版）により自動作成された処理結果例を図65に示す。この結果から、木材の色調や木目の状態によらず、木簡形状と墨痕を正確に認識していることが確認できる。

現段階での処理精度は、ファイルアップロード版でおよそ９割、カメラ撮影版でおよそ４～５割である。とりわけカメラ撮影版では、光源環境等の撮影条件の影響により処理結果がばらつく傾向がある。

学習データのブラッシュアップ　現在、「深層学習による木簡実測図の自動作成」プログラムの処理結果で、誤認識あるいは認識が甘い箇所を手動で修正したデータを作成し、学習データとしてフィードバックする作業を進めている。この手法はすでに多くの研究で有効性が示されており、今後の精度向上が大いに見込まれる。

５　展望―文字画像データセットの公開

近年、奈文研では木簡のボーンデジタル画像の整備・活用に積極的に取り組んでいる。このデジタル画像をもとに、木簡上の各文字座標を取得し、単文字画像を作成している。作成された単文字画像は、現在「史的文字データベース連携検索システム」サイト上で、オープンデータとして高精細画像及びメタデータを公開しており[5]、2021年３月時点で、約13,000件に達している。

本データ群は、手書き文字画像認識や運筆情報取得のための機械学習の学習データセットとして、2021年度内の公開を目標としている。国内外問わず様々な分野での利活用を広く呼びかけていきたいと考えている。

本稿は科研費基盤（S）「木簡等の研究資源オープンデータ化を通じた参加誘発型研究スキーム確立による知の展開」（課題番号18H05221）等の成果を含む。

（畑野吉則・馬場　基・桑田訓也・高田祐一）

註

1）https://repository.Nabunken.go.jp/dspace/handle/11177/7944

2）http://codh.rois.ac.jp/char-shape/
国文学研究資料館所蔵で日本古典籍データセットにて公開する古典籍、および国文学研究資料館の関係機関が公開する古典籍から切り取った、くずし字4,328文字種の字形データ1,086,326文字（2019年11月時点）。

3）大山航・畑野吉則・馬場基「深層学習による木簡実測図の自動作成」『じんもんこん2020』。

4）http://imedia.sit.ac.jp/MokumeDemo/

5）畑野吉則・馬場基・桑田訓也・高田祐一「史的文字データベース連携検索ポータルサイトの公開」『紀要 2020』48-49頁。