

情報センター・画像センター共催研究会報告

## 歴史情報研究の展開

### —文理融合研究のかたち—

二〇二〇年一〇月一二日、史料編纂所前近代日本史情報国際センターと画像史料解析センターとの共催で、Zoom によるオンライン研究会を開催した。

史料編纂所では、日本史の各種情報をデータベース化するだけでなく、データ処理に情報学の最新の手法を用いて、データの整理や探索過程の省力化・正確化をはかる方法、あるいは歴史的思考過程に、多様な視覚や情報を外部から自動的に提示して研究を支援するシステムの開発などに取り組んでいる。従来より、情報センターに人文情報学の専任教員を配置し、また画像センタープロジェクトにおける共同研究や科学研究費補助金による基盤研究を通して、その研究開発を進めてきたが、さらに二〇一九年、画像センターに機械学習による画像分析などを専門とする大山航氏を客員教授に迎え、今年度、情報センターにも人文情報学の中村寛、考古学を専門としながら他分野を交えた文理融合研究を推進している渋谷綾子をあらたに教員に加えた。そこで、この機会に史料編纂所が取り組んでいる歴史情報研究の具体的な内容を共有し、今後の課題などを考えるために、研究会を設けたものである。

研究会には、五三名の所員が参加し、人文・歴史情報研究の現在の状況を把握するとともに、研究素材としてのデータや連携データの拡充、データの信頼性・再現性や公開性の確保などがなお課題であることを知ることができた。また歴史情報と理系分野との協働のあり方に、ひととおりでない幅広いものがあることも実感できた。ここに研究会の概要を掲載する。今後も文理双方で意見を交わしながら歴史情報学の基礎部分をしっかりと構築し、発展させていきたい。

（高橋敏子／画像史料解析センター長）

## 画像でつなぐ歴史学と情報学

### —深層学習を活用した画像史料解析の事例紹介—

大山 航

#### はじめに

近年、深層学習を代表とする機械学習によるデータ解析技術が様々な分野で活用されている。歴史学分野も例外ではなく、大量データを援用した種々の試みが行われている。一方で、少量かつ不均衡で、重複するデータを多く含む画像史料は、最新の機械学習技術にとっても大変チャレンジングな題材である。

筆者は、パターン認識や機械学習、画像解析を本業とする研究者であり、これまで古文書の花押や筆跡の画像データ、木簡画像の解析といった共同研究を行ってきた。本報告ではこれらの共同研究から、いくつかの研究成果を紹介する。

#### 一 深層学習の概要

深層学習は、現在の人工知能、機械学習ブームの立役者である。二〇一二年に開催された国際的な画像認識コンペティション ILSVRC において、従来を遥かに凌ぐ高い性能を実現したことで注目を集めた。機械学習、パターン認識技術による画像認識では、まず、その画像をよりよく表現できる特徴量を抽出する必要がある。従来手法では、この特徴量の抽出手順を開発者が手作りで設計する必要がある。これが性能のボトルネックとなるケースが多かった。これに対して、深層学習では、大量データを使った End-to-End 学習により、深層学習モデル自身が認識対象に効果的な特徴抽出を学習できる。

一方で、深層学習モデルが獲得できる特徴表現は、学習用に与えられる画像データに依存する。一般的に、大量データが与えられた場合は十分な性能

を獲得できても、次のようなケースへの対応は困難である。

- ・ 少量：主に学習用データの総量が不足する場合や、正解付け作業のコスト（時間・費用）が高い場合など。少量データに対しては過学習の発生により性能が低下しやすい。

- ・ 不均衡：一部のクラスが未知、または極めて小さいなどデータがアンバランスである。このような不均衡データに対しては、最悪の場合、小さいクラスが黙殺される場合もある。

- ・ 重複：複数クラスがオーバーラップ（あいまい）する。クラス内の変動が大きい、またはクラス間の違いが小さい場合。文字認識における類似文字や双子の顔認識などが一例である。認識エラーの原因となる。

前述の通り、画像史料データはこのような少量、不均衡、重複の特性を有する典型的データであり、最新機械学習技術にとってもチャレンジングなテストケースと言える。

## 二 深層学習による花押画像の解読

二〇一五年度より、花押データベースに対して、画像をクエリとして検索できる機能を実現するための機械学習の適用に関する研究プロジェクトを進めている。本節ではこの研究プロジェクトの成果を紹介する。

### 1 畳み込み自己符号化器による花押形状特徴の抽出

機械学習、深層学習からみた花押画像データには、次のような難しさが存在する。

- ・ 形状変動の大きさに対して、花押筆者（クラス）あたりの画像数が少なく、変動の正確なモデル化が困難である。そのため画像分類手法の適用は難しい。
- ・ 筆者ごとの花押画像数のばらつきが大きく、画像数の多い筆者を重視してしまう。

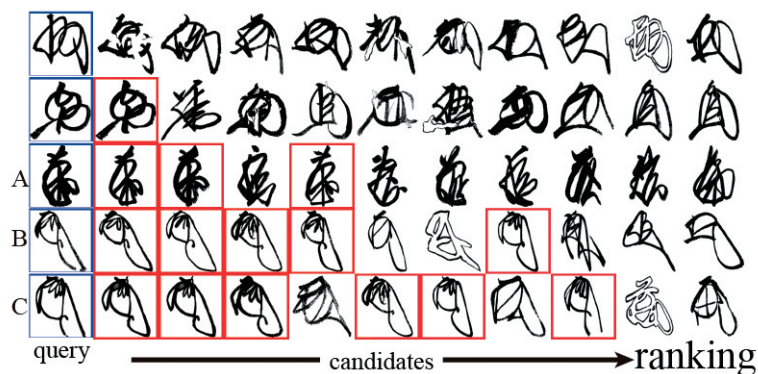
一方で、花押画像データ全体では画像数は比較的大きく、深層学習により

データセット全体の変動を表現するコンパクトな形状特徴表現の獲得が期待できる。本研究では画像に対して用いられる標準的な深層学習技術である畳み込み自己符号化器を用いて、花押形状特徴の抽出を試みた。

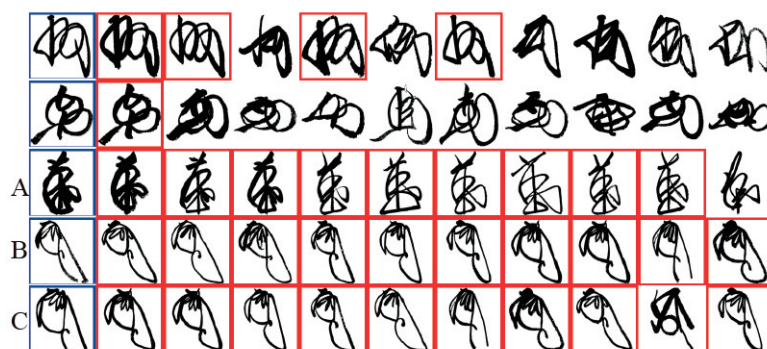
図一に、学習が完了した畳み込み自己符号化器により抽出された花押の形状特徴の例を示す。最上段が花押画像、下四段が抽出された形状特徴である。解析の結果、花押形状の特徴が3種類と、筆跡の太さを表現する特徴量（再下段）が抽出されていること、本手法により抽出された特徴量を用いて、従来型の開発者が設計した形状特徴と同程度の画像検索性能が実現できることがわかった。



図一 畳み込み自己符号化器により抽出された形状特徴



形状特徴単独で学習した場合の検索結果



形状特徴と類似度尺度を同時に学習した場合の検索結果

図二 形状特徴と類似度尺度の同時学習による花押画像検索の性能向

2 花押形状検索への深層計量学習の導入

前項で述べた畳み込み自己符号化器による形状特徴抽出は、筆者クラスごとのデータ不足と偏りを、データセット全体を学習することでカバーする手法であった。これに対して、花押画像データが持つ、少量、不均衡、重複データ問題に直接アプローチする手法として、形状特徴と類似度尺度を同時に学習する手法を開発した。

形状特徴と類似度尺度の同時学習により、画像検索の性能が向上した例を図二に示す。図中上段、下段がそれぞれ、前節で述べた形状特徴を単独で学習した場合、形状特徴と類似度尺度を同時に学習した場合の検索結果例であ

図三 深層学習による木簡実測図自動作成の処理結果例



る。それぞれの図において、最も左の花押画像がクエリ画像であり、クエリ画像から右側へ向かって順位の高い順に検索結果が並んでいる。検索結果の赤枠はクエリと同じ人物クラスの花押であることを示す。同時学習の導入により、検索性能が大きく向上したことが確認できる。

### 三 深層学習による木簡実測図の自動生成

木簡研究の成果を管理し再利用を促進するために、デジタルアーカイブの整理と集積が行われている。木簡のデジタル写真データを各種のデータベースに収録するためには、釈読やデータへのタグづけ作業が必要となる。この作業の大部分が研究者の手作業で行われており、作業の省力化、効率化を実現し、正確性を向上する支援ツールの充実が望まれている。

急速に性能が向上している深層学習技術は、分類、認識だけでなく画像生成や画像変換にも活用されており、木簡に対する画像処理の高精度化、自動化の実現が期待される。深層学習技術は、古文書研究では徐々に活用されつつあるものの、木簡研究では十分に活用されていない。



本研究において筆者らは、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手法を提案した。提案手法により自動作成された木簡実測図の例を図三に示す。図中では木簡画像と、その画像に対して提案手法が作成した木簡実測図を左右に並べて配置した。これらの結果から、木材の色調や木目の状態によらず、安定して木簡形状と墨痕を正確に捉えた実測図を作成可能であることが確認できる。

本研究で実現された木簡実測図の自動生成は、木簡研究に以下の貢献をもたらす。

(1) 研究者の省力化に寄与し、それにより質が高く多様な視点からの歴史情報収集を可能にする。

(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可能にする。

(3) 木簡文字自動認識の前処理として活用できる。

#### おわりに

本稿では、深層学習技術の、画像史料解析への応用事例として、

・花押画像の機械解読

・深層学習による木簡見取図の自動生成

の研究事例を紹介した。前述した通り、画像史料は少量・不均衡・重複データの典型例であり、機械学習の適用対象としても非常に興味深い研究題材である。今後も画像史料研究に適用できる、画像解析、機械学習研究を進めていく所存である。

付記 本報告では主に史料編纂所「花押データベース」、及び、奈良文化財研究所「スマートフォン撮影による木簡画像処理実験用データベース」を用いた。

(埼玉工業大学／画像史料解析センター客員教授)

## データ駆動型歴史情報研究基盤の構築に向けて

中村 覚

#### はじめに

東京大学史料編纂所は明治以来一五〇年に渡り蓄積してきた目録、画像、本文、文字など大量かつ多様な人文科学データを有している、具体的には、『大日本史料』など一二〇〇冊近い基幹史料集の編纂・刊行実績、および歴史情報の公開・発信でも四〇種五六〇万件のデータベースと史料画像二〇〇〇万件のデジタルアーカイブを有している。このビッグデータを活用することにより、他の機関では真似できない、より良質なデータ駆動型の歴史情報研究基盤を構築することができる。

本稿では、この「データ駆動型歴史情報研究基盤」の構築に向けた取り組みについて報告する。「データ駆動」が意味するところは一つに定まらないが、本研究では、何らかのタスクに対して、人の要求を待たずして計算機が情報を提示すること、とする。例えば検索タスクにおいて特定の情報に辿り着いた際、その情報に関連する情報をシステムが自動で提示する、といったことが考えられる。これにより、利用者が見落とししていた情報、思いもよらなかった情報、などを提示し、当該タスクの高度化・効率化を目指す。

#### 一 データ駆動型歴史情報研究基盤の構築方法

データ駆動型歴史情報研究基盤の構築は、表現は異なるが、国内外で積極的に取り組まれているテーマである。ヨーロッパでは「タイムマシンプロジェクト」<sup>1)</sup>が進められており、国内では、例えばRIS／人文学オープンデータ共同利用センター（以下、CODH）<sup>2)</sup>が歴史ビッグデータ研究を進めている。「タイムマシンプロジェクト」は、ヨーロッパ史に関する大規模なシミュレータを構築し、文化施設が有する膨大なコレクションをデジタル情報システ