4ZK-02

深層学習を用いた削り屑木簡のグループ化と再構成

山田　海人†　畑野　吉則‡　馬場　基‡　大山　航†
埼玉工業大学†　　奈良文化財研究所‡

はじめに

木簡は，文字が記された木片古文書である．表面を削り取り再利用されていたため，出土する木簡のほとんどが削り屑の状態である．この削り屑木簡を再構成し解読すれば，当時の物流や行政の様子がわかると期待される．図 1 に出土した削り屑木簡の例を示す．削り屑木簡に墨痕が認められるが，内容の読解は困難である．

削り屑木簡再構成の実現には様々な課題がある．主な課題に，削り屑で大量に出土するため，どの削り屑がどの木簡から削りだされたものなのか全く分からない問題が挙げられる．再構成の作業は，歴史学の専門家でも手間と時間のかかる困難な作業であるため，ソフトウェアや人工知能技術を活用した省力化が望まれている．

これらの課題に対して，Truyen ら[1]は色特徴によるクラスタリングと，輪郭比較による木簡の再構成作業支援手法を提案した．このシステムは画像のグループ化と再構築の 2 つのモジュールから構成されている．このように 2 つのモジュールにすることでグループ化と再構成の工程を分けて行うことができる．グループ化モジュールでは，特徴データを用いて画像を自動的にグループ化する．抽出される特徴量は用手的に設計されたものであり，大量に存在する削り屑木簡画像に対して最適であるか疑問が残る．

また，Antoine ら[2]は断片で出土したエジプトの古文書であるパピルスに対して Siamese Network[3]を用いてクラスタリングを行う手法を提案した．この手法は深層学習に基づく手法ではあるが，あらかじめ的確なアノテーションが付与された訓練データセットが必要である．そのためほとんどアノテーションが付与されていない削り屑木簡画像に対しての適用は難しい．

本研究は，削り屑木簡画像のクラスタリングを目的とする．あらかじめ類似する削り屑画像をクラスタリングすることで，再構成プロセスの簡略化が期待できる．そのための手段として，

図 1　削り屑木簡の例

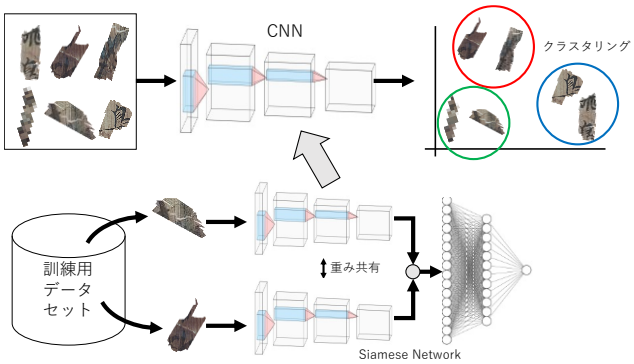

図 2　提案手法の概要

Siamese Network を教師なしデータである削り屑木簡で学習する手法を提案する．

提案手法

図 2 に提案手法の概要を示す．提案手法は，削り屑木簡のカラー画像から構成されるデータセットを入力とする．入力されたデータセットに含まれる各画像に対して畳み込みニューラルネットワーク（CNN）により特徴抽出を行う．抽出された特徴ベクトルデータセットに対して，クラスタリング手法を適用する．

CNN の学習は，Siamese Network の枠組みを利用して行う．通常，Siamese Network の学習には，入力された画像ペアが同一クラス由来か，別クラス由来かの情報を必要とする．すなわち学習にはクラス ID が付与された訓練用データセットが必要となる．また，提案手法の性能を定量的に評価するためにもクラス ID が付与された評価用データセットが必要となる．しかし，今日においては，クラス ID などアノテーションが付与された削り屑木簡画像のデータセットは存在し

Grouping and Reconstruction of Mokkan Fragments Using Deep Learning
† Saitama Institute of Technology
‡ Nara National Research Institute for Cultural Properties

本研究は JSPS 科研費 18H05221，20H00022 の助成を受けたものです．

ない．本研究では，仮想的にクラス ID が付与された実験用データセットを生成し，実験に用いる．

実験用データセットの生成

本研究で用いるデータセットは37,136個の削り屑木簡画像を含む．本研究では，切り出し元の木簡が識別できる ID が付与された，仮想的な削り屑木簡画像からなる実験用データセットを生成した．

図 3 に実験用データセット生成手順を示す．大小木簡画像を組み合わせ，大型木簡上に小型木簡をランダムに配置し，小型木簡の形状で大型木簡を切り抜く．切り抜かれた画像に大型木簡の ID をクラス ID として付与する．このようにして生成したクラス ID 付きデータを生成することで，木簡画像を用いた Siamese Network の学習が可能となる．大型木簡を 400 画像用いたので，生成された実験用データセットは 400 クラスの小型木簡画像データセットとなる．

実験

作成した実験用データセットを使用して，削り屑木簡のクラスタリング実験を行った．まず，400クラス（490,002画像）のデータセットを訓練用（350 クラス），評価用（50 クラス）に分割し，訓練用サブセットを用いて Siamese Network により CNN の学習を行った．

学習済み CNN を用いて評価用サブセットの各木簡画像に対して特徴抽出した．抽出された特徴ベクトルに対して， クラス数が未知である場合を想定して，クラスタ数も推定できる Affinity Propagation（AP）法と，クラス数が既知である場合を想定した k-means 法,そして AP で求めた各クラスタの代表サンプル（セントロイドと呼ぶ）を階層型クラスタリングしたそれぞれの結果を比較した．

クラスタリング結果の性能評価には，相互情報量に補正を加えた Adjusted Mutual Information（AMI）を用いた．Siamese Network による CNN の学習の有効性を検証するために，同じ構成の CNN に全結合層を追加してクラス分類器として学習した場合との性能比較を行なった．

結果と考察

図 4 に，提案手法（図中 siamese）と，クラス分類器として学習した CNN（図中 CNN）のそれぞれで抽出した特徴ベクトルに対して，AP 法と k-means 法でクラスタリングした場合，AP 法での推定クラスタ数が大きくなったため，後処理としてセントロイドを階層型クラスタリング（図

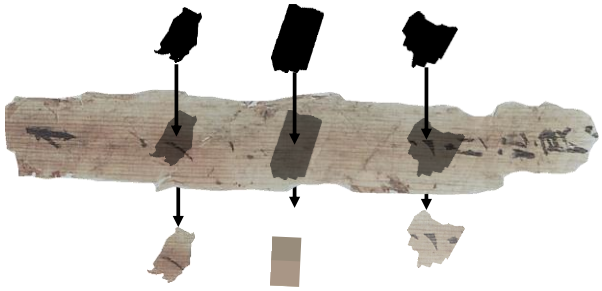

図 3　実験用データセットの生成方法

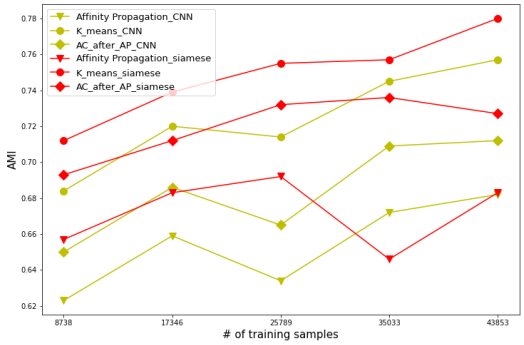

図 4　提案手法および CNN により抽出された特徴に対するクラスタリング性能の比較

中 AC_after_AP）にかけた場合の AMI の値を示した．特徴抽出器を学習するデータセットのサイズが，クラスタリング性能に与える影響を調べるため，学習データセットに含まれるクラス数に 70〜350 の 5 通りの条件を設定し，それぞれの条件で実験を行なった．いずれの条件においても，CNN よりも提案手法の AMI 値が大きくなった．このことは Siamese Network による CNN の学習が，クラスタリング性能を向上させることを示唆する．また，階層型クラスタリングを行った際には，AMI の向上が見られており，より精度の高い分類が実現されている．

まとめ

本研究では，再構成プロセスの簡略化のために，類似する削り屑画像をクラスタリングする手法を提案した．学習した CNN により抽出された特徴ベクトルに対して， Affinity Propagation 法と k-means 法を使ってクラスタリングを行った．いずれの手法においても高い精度のクラスタリングが実現できた．

参考文献

[1] Truyen Van Phan et al., A re-assembling scheme of fragmented Mokkan images, HIP '13
[2] Antoine Pirrone et al, Papy-S-Net: A Siamese Network to match papyrus fragments, HIP'19
[3] Jane Bromley et al. Signature Verification using a "Siamese" Time Delay Neural Network, NIPS'93