

機械学習による石器図面画像の自動抽出と分類

はじめに

考古学は蓄積型の学問であり、発掘調査報告書（以下、報告書）は、重要な基礎資料である。しかし、報告書は戦前含め推計125,000冊あり膨大にあるため、過去の蓄積に適切にアクセスしにくいという課題もある。そして考古学においては、遺構・遺物そのものが研究対象であるため、画像情報（図面・写真）は重要である。しかし、これまでに報告書に掲載された画像だけで写真400万件、図面500万件、表・グラフ50万件、合計950万件あると推計されている¹⁾。もはや人間の目によって1件ずつ確認して必要とする画像を探し出すことは困難な状況である。これまでにない情報検索方法が必要である。本稿では、報告書に掲載された膨大な画像にアクセスするための画像認識技術の取り組みを報告する。

PDFから画像の抽出

報告書の電子公開は主にPDFファイルによって実現されている。PDFファイルは、印刷物のレイアウトを継承したまま電子化できるなどメリットが多いが、データ自体は構造化されていないため、機械可読性は低い²⁾。まずは2020年度に作成した教師データをもとに機械学習のソフトウェアライブラリを使用して、PDFファイルから画像82万件を自動抽出した。写真、図面が混在し、遺構および遺物すべての画像データであるためまずは石器の図面画像のみを人間で選別した（図67）。



図67 大量に自動抽出した画像から石器図面を選別

石器種別ごとの教師データ作成

石器の図面画像をさらに54に種別設定し、種別ごとに教師データを作成した（図68）。作業には野口淳・国武貞克・森先一貴の協力を得た。種別は次の通り。

スタンプ形石器、ストーン・リタッチャー、ナイフ形石器、両面礫器、両面調整石器、円形搔器、削器、剥片、剥片石核、台形様石器、台形石器、台石、尖頭器、岩偶、彫器、御物石器、打製石斧、打製石斧（直刃斧）、挟入石器、搔器、敲石、斧形石器、有舌尖頭器、有茎石器、槌石、横長剥片、片面礫器、独鈷石、環石、石冠、石刃、石剣、石匙、石斧、石核、石棒、石皿、石篋、石鉢、石錐、石錘、石鏃、砥石、磨り石、磨石、磨製石斧（両刃）、磨製石斧（片刃）、細石刃、細石刃核、細部加工剥片、縦長剥片、舟形石器、鋸歯状石器、錐。

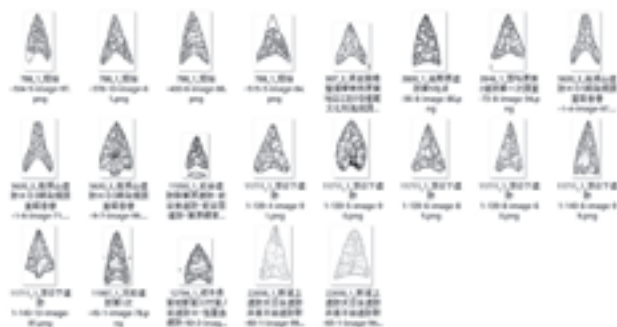


図68 石鏃の教師データとして選別した図面画像

機械学習にて種別の類似度を算出

機械学習にて個別の石器画像ごとに54の種別の類似度を算出した（図69）。数値が高い種別ほど類似していることを示す。この工程によって石器種別ごとに分類できたことになる。



ナイフ形石器(図面) 1
尖頭器(図面) 7
石鏃(図面) 13
類似画像

図69 機械学習にて各種別の類似度を算出

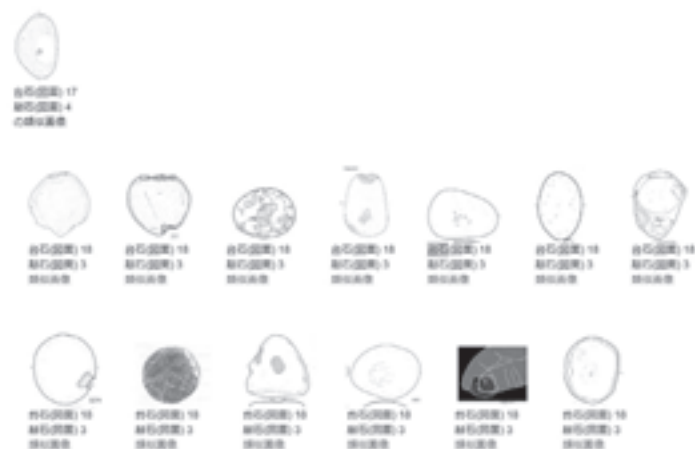


図70 台石の類似図面画像



図71 石匙の類似図面画像

類似画像の表示

当該画像に類似している画像を表示させる機能も開発した。おおむね類似している画像を表示させることに成功した。しかし、一部関係のない画像も混入しており、精度の向上の余地は残る（図70、71）。

今後の展開

今後は、土器・瓦・木製品など遺物ごとに教師データを作成し、機械学習による分類と類似度算出を進める予定である。また、画像を検索キーとする画像検索も実装を検討する。遺物名が不明であっても画像で検索できるため、用途不明遺物への対応に有効となる。遺物名がわからない市民にとっても便利となる。

おわりに

前述のとおり報告書掲載画像推計950万件を一人の人

間が確認することはほぼ不可能である。さらに画像が累積していく未来になっては加速度的に閲覧困難になると予測される。画像認識技術など大量データに打ち勝つための手段が必要である。さらに大量のデータならではの考古学研究が可能となれば、研究手法のブレイクスルーとなる可能性を秘めている。まずは、必要とする情報を得るために効率的・網羅的な検索をITによって実現し、収集した情報の精査や分析を人間で行うという役割分担が当面の大規模データへの対応モデルの一つの目標となる。

なお、本稿は、研究課題19K21643の成果の一部である。

（高田祐一・野口 淳）

註

- 1) 高田祐一「報告書のデータ量を推計する」『文化財の壺』7、4-5頁、2019。
- 2) 高田祐一「画像認識技術の文化財データへの適用実験」『紀要 2020』46-47頁、2020。