2020年度 人工知能学会全国大会（第34回）

国立国会図書館の

デジタル化資料の検索・提供方法改善を目的とした技術開発とデータセット構築の取組について

青池 亨, 木下 貴文, 里見 航, 川島 隆徳(国立国会図書館)

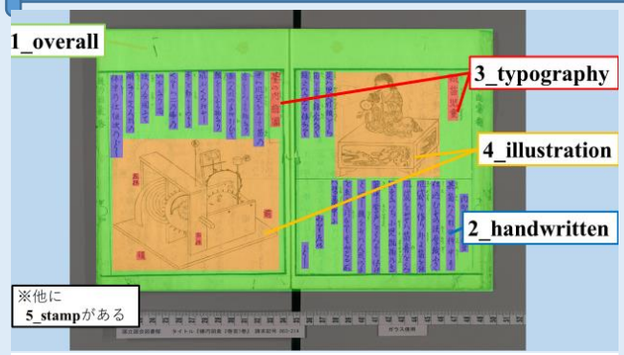
国立国会図書館(NDL) 電子情報部電子情報企画課 次世代システム開発研究室は、機械学習技術を応用することで所蔵資料の検索可能性と提供可能性を高めるべく調査研究活動を行っている。調査研究成果を組織内外問わず活用してもらうため、次の2点を意識して取り組んでいる。

【GitHubでのデータセット公開】NDLが保有する著作権保護期間満了資料の画像を利用して作成した機械学習用データセットを、自由に利用可能な条件で公開

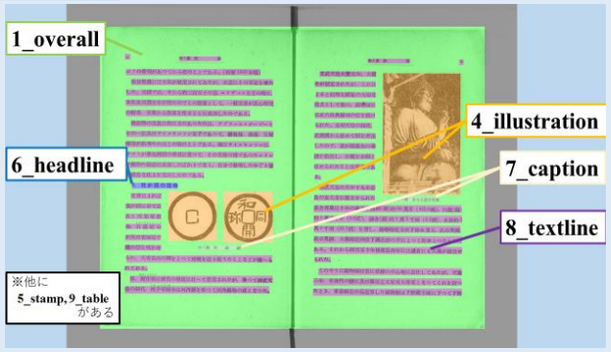
【成果のWebサービス化】研究成果をWebアプリケーションに実装し「**次世代デジタルライブラリー**(https://lab.ndl.go.jp/dl/)」と称した実験サービスとして一般に提供

以下、2019年11月に公開した資料レイアウトデータセット「**NDL-DocL**(https://github.com/ndl-lab/layout-dataset)」に関し、構築方法と当研究室内での利用について紹介する。

データセットの構築に関して

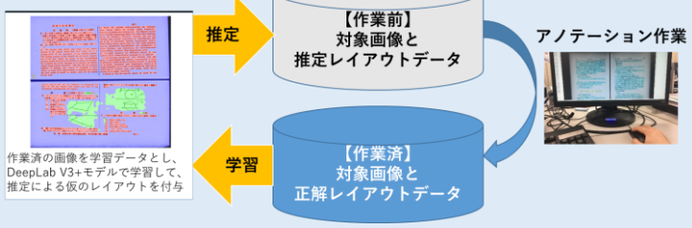


古典籍資料(**1,219画像**)

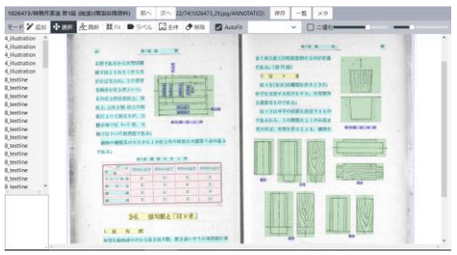


明治期以降刊行資料(**1,071画像**)

データセット構築のサイクル



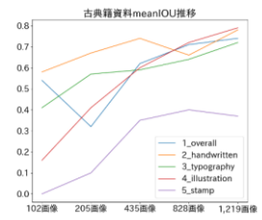
資料用アノテーションツールの開発

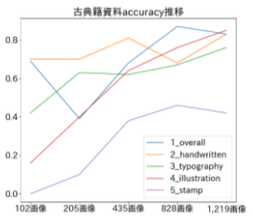


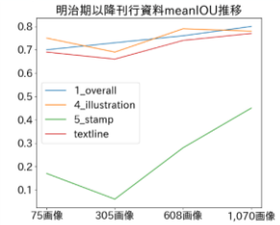
二値化による文字列を囲む矩形の自動フィット
XY軸方向の矩形の分割等を可能とした。
https://github.com/ndl-lab/layout-dataset/tree/master/tugidigi-annotation

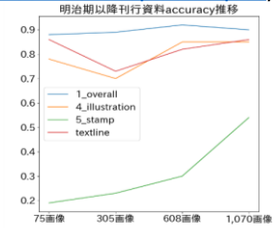
データセット作成枚数と仮レイアウトの推定精度の推移

- テスト用のデータセットを古典籍資料から211画像、明治期以降刊行資料から200画像分作成し、データセット作成途中における仮レイアウトを付与する推定モデルの性能を定量評価した。
- 仮レイアウト段階では6_headline,7_caption, 8_textlineの区別はつけていないため、まとめてtextlineとした。
- モデルはDeepLab V3+のkeras実装を改変して用いた(https://github.com/ndl-lab/tensorflow-deeplab-v3-plus)。









構築したデータセットを利用した調査研究に関して

高速なレイアウト認識方法の検討と応用

作成したデータセットを活用して、実際にNDLが保存・提供しているデジタル化資料に対して実用上高速なレイアウト認識手法を検討した。

適用した手法

ESPNETv2(https://github.com/sacmehta/ESPNetv2)

対象

日本十進分類法で「5類(技術・工学)」「7類(芸術・美術)」に分類される著作権保護期間満了のデジタル化資料全件、30,693点(合計2,734,508画像)
※いずれも高精細JPEG2000形式

利用したレイアウト

NDL-DocLのラベルをOverall, Textline, Illustrationの3クラスに分けて学習した。

性能評価

実資料への適用に先立って、認識性能の評価を行った。
テスト用データセットは左記と同様である。
資料の多くの領域を占めるTextlineは性能が低いが、IllustrationとOverallについては高性能であった。
処理速度は8.76画像/秒で、DeepLab V3+に対して10倍程度高速であった。

ラベル	ESPNETv2	DeepLab V3+(比較)
Overall	**0.86/0.94**	0.83/0.92
Textline	0.74/0.83	**0.79/0.85**
Illustration	**0.82/0.87**	0.69/0.83

※数値はラベル領域毎のmeanIOU/Accracyを表す

適用結果

- GTX1080Ti 1基を利用して、約2週間(337時間)で処理が完了した。
- 実験で自動認識・座標抽出したIllustration(1,734,911点)は、全て次世代デジタルライブラリー(https://lab.ndl.go.jp/dl/)に登録済であり、任意の対象資料について認識結果を確認することが可能である。

今後について

- 今回の手法について、保有する著作権保護期間満了のデジタル化資料全画像に拡大して適用する処理を実施中である。2020年度前半内に完了する見通しである。
- 速度面・精度面でのモデル検討が引き続き重要となる。文字認識と組み合わせた古い文献の資料画像に対応した全文テキスト化についても調査研究を進めている。

参考文献

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, Hartwig Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, ECCV 2018, pp. 801-818, 2018
Sachin Mehta, Mohammad Rastegari, Linda Shapiro, Hannaneh Hajishirzi, ESPNetv2: A Light-weight, Power Efficient, and General Purpose Convolutional Neural Network, CVPR 2019, pp. 9190-9200, 2019