4ZK-01

深層学習を用いた近代新聞画像からの広告挿絵の検出と日付ごとの整理

北清　敦也†　　寺沢　憲吾†
公立はこだて未来大学　システム情報科学部†

1. 背景と目的

近年では，数多くの歴史的文書画像がデジタルアーカイブとして保管されており，我々は容易に閲覧することができる．歴史的文書には主に文字や挿絵などが含まれておりどちらも当時の世情や環境を理解する上で重要な要素である．歴史の研究において，資料間の比較や分析を行いやすくすることには需要があり，安達・鈴木[1]は，歴史資料の閲覧・比較を簡便な操作で行えるシステムを開発した．また，研究に使用するデータを扱いやすく整理することにも需要があり，坂部[2]はデータを集計用に整備すると，さまざまな要素との関連分析が可能になると述べている．

以上の背景より本研究では，歴史的文書における挿絵，中でも明治期に発行された函館新聞に含まれる広告挿絵を自動的に検出し，検出した広告挿絵画像を日付ごとに整理する手法を提案する．広告挿絵検出には，畳み込みニューラルネットワーク（CNN）を使用する．物体検出アルゴリズムには YOLOv3 を用いる．日付ごとにどの種類の広告挿絵が含まれているかを整理するために，挿絵画像間の同一性の検証を行い，最終的に日付と広告挿絵の対応表として出力することで可視化させることを目標とする．

2. 関連研究と課題

青池らは，セマンティックセグメンテーションというアプローチを用いて，資料中における挿絵領域を自動抽出して，同様の挿絵を含む資料を検索する手法を提案した[3]．セマンティックセグメンテーションとは，ディープラーニングのアルゴリズムの一種で，画像内の全画素にラベル付けを行うものである．青池らの研究では，テストデータセットにおいて，画素単位で83%の正解率であった．しかし，ノイズが多く含まれた資料に対しては，ロバストな領域認識方法や，挿絵画像の特徴抽出方法の検討をする必要があると述べている．

Detection and Organization of Advertisement Illustration in Early Modern Newspaper Images Using Deep learning
†Atsuya Kitase †Kengo Terasawa
†School of Systems Information Science, Future University Hakodate

3. 提案手法

3.1 概要

本研究では，明治 11 年 1 月 12 日から明治 17 年 12 月 28 日にかけて発行された函館新聞の画像（3574 枚）から広告挿絵が含まれている新聞画像（454 枚）を抽出し，実験用データセットを作成した．広告挿絵は目視でカウントした結果 145 種類あることが分かった．実験に使用した函館新聞の画像と広告挿絵画像の例は図 1 に示す．明治期の函館新聞には同一の広告挿絵が複数回登場するという特徴がある．図 1 に示した旗と船の広告挿絵も別日に再掲載されている．

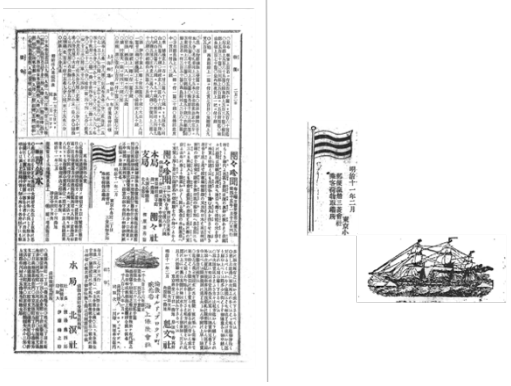

図 1.　函館新聞（左）新聞中の広告挿絵の例（右）

本手法では，はじめにデータセットの新聞画像中から広告挿絵画像を YOLOv3 を用いて自動検出する．次に検出した広告挿絵画像の同一性を検証するために AKAZE 特徴量を使用してクラスタリングを行う．最後に，クラスタ分けされた広告挿絵画像を日付ごとに整理する．

3.2 広告挿絵の自動検出

広告挿絵の検出に YOLOv3 を使用する．新聞画像から広告挿絵の箇所を矩形で囲いラベリングを行った．YOLO は物体検出だけではなく，クラス分類も同時に行うことができるが，未知画像の広告挿絵も検出できるようにするために「挿絵」という 1 クラスでラベリングを行う．145 種

類の広告挿絵画像が1回以上含まれるように学習用データセットを作成し検出を行った．検出結果として出力されたバウンディングボックスの情報を元に広告挿絵画像として切り抜き保存した．

3.3 広告挿絵画像のクラスタリング

切り抜いた広告挿絵画像の同一性を検証するために，AKAZE[4]を用いたクラスタリングを以下の手順で行う．

(1) 切り抜かれた全広告挿絵画像の特徴点をAKAZEを用いて検出する
(2) 広告挿絵画像間で総当たりマッチングを行い，特徴点の対応付けをする
(3) 対応のついた特徴点間の距離の平均を類似度として完全連結法でクラスタリング

3.4 日付ごとに整理

どの広告挿絵がいつの新聞に載せられていたかという点を確認できるように，表として自動出力する．

4. 実装結果と評価

広告挿絵検出については図 2 のような結果が得られた．新聞画像の 454 枚の中から 879 枚の広告挿絵が検出された．145 種類の挿絵画像のうち 142 種類の広告挿絵画像を検出することができた．

図.2 検出結果の例

未検出広告挿絵は，図 3 に示した 3 種類である．特徴として広告挿絵のサイズが他と比べ小さいものや，文字領域がノイズとなっているものがあった．

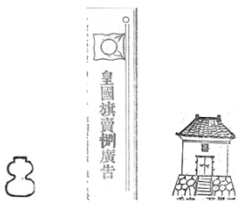

図.3 未検出画像

提案手法により広告挿絵画像をクラスタリングした結果は図 4 に示す．この図は正確にクラスタ分けされた例である．検出された広告挿絵は 142 種類であったため，最大クラスタ数を 142 個に設定してクラスタリングを行った．

図.4 クラスタ番号 78 に含まれた画像

クラスタリングについて評価を行った．本研究ではクラスタリング結果を成功とみなす条件を以下のように設定した．

① 1 クラスタに 1 種類の広告挿絵画像のみがクラスタ分けされていて他クラスタにその広告挿絵画像が分けられていない場合
② 1 クラスタに複数の種類の広告挿絵画像が含まれたが他のクラスタにそれらの広告挿絵画像が分けられていない場合

失敗となるのは，同種類の広告挿絵が複数のクラスタに分けられているという場合である．評価結果は，前節で検出できた 142 種類の広告挿絵画像のうち 127 種類は成功であり，15 種類が失敗であった．

5. まとめ

本研究では，明治時代の函館新聞を対象にして新聞内の広告挿絵を自動検出し日付ごとに整理することを目的とした．YOLOv3 を使用して広告挿絵の検出を行い，日付ごとに広告挿絵画像を保存するプログラムを実装した．また，広告挿絵画像間の同一性を検証するために AKAZE 特徴量を使用したクラスタリングも行った．今後は，広告挿絵と日付の対応表の作成と，広告挿絵検出やクラスタリングについて評価を行う．また，検出精度の向上を目指す．

参考文献

[1] 安達文夫，鈴木卓治：超精細画像による資料の比較閲覧機能の検討，情報処理学会研究報告人文科学とコンピュータ，Vol.2004，No.110（2004-CH-064），pp.9-16（2004）．

[2] 坂部裕美子：落語の寄席定席番組データの活用と課題，デジタルアーカイブ学会誌，Vol.4，No.S1，pp.1-4（2020）．

[3] 青池亨，里見航，川島隆徳：資料画像中の挿絵領域の自動抽出及び画像検索システムの実装，じんもんこん2018論文集，Vol.2018，pp.97-102（2018）．

[4] Alcantarilla, P.F., Nuevo, J., Bartoli, A.: Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces, BMVC, pp.13.1-13.11 (2013).