Single Shot MultiBox Detector と文書要素の階層構造を用いた文書画像の領域分割に関する研究

林侑生[1]　鶴田直之[1]　乙武北斗[1]

概要：住民の地方自治への参加を促進する目的で，地方議会の活動資料を自然言語処理してデータベース化する研究が進められている．多くの地方自治体は議会活動資料を PDF 文書で公開している．しかし，PDF 文書は，文書の内部データの自由度が高く，文字や語の順序が正しく保存されたテキストデータを抽出することが必ずしも容易ではない．そこで，本研究では，予め文書から画像処理によって文書の構成要素と配置を抽出し，テキストデータ抽出の手がかりとすることを検討している．本稿では，画像認識用深層学習 Single Shot MultiBox Detector (SSD)と文書の階層構造に関する知識を用いて文書の構成要素と配置を抽出方法を提案する．

キーワード：文書画像処理，深層学習，Single Shot MultiBox Detector，階層構造

A Study on Region Segmentation of Document Images Using Single Shot MultiBox Detector and Hierarchical Structure of Document Elements

YUKI HAYASHI[†1]　NAOYUKI TSURUTA[†1]　HOKUTO OTOTAKE[†1]

Abstract: To promote the participation of local residents in local government, studies are underway to create a database of local council activity reports using natural language processing. Many local governments publish council activity reports as PDF documents. However, PDF documents have a high degree of freedom in the internal data of the documents, and it is not always easy to extract text data in which the order of characters and words are preserved correctly. Therefore, in this study, we are investigating the use of image processing to extract document components and their arrangement from documents in advance as a clue for text data extraction. In this paper, we propose a method for extracting document components and their arrangement using deep learning Single Shot MultiBox Detector (SSD) for image recognition and knowledge about the hierarchical structure of documents.

Keywords: Document Image Processing，Deep Learning，Single Shot MultiBox Detector, Hierarchical Structure

1. はじめに

住民の地方自治への参加を促進する目的で，地方議会の活動資料を自然言語処理してデータベース化する研究が進められている[1],[2]．多くの地方自治体は議会活動資料を PDF 文書で公開している．しかし，PDF 文書は，文書の内部データの自由度が高く，文字や語の順序が正しく保存されたテキストデータを抽出することが必ずしも容易ではない．そこで，本研究では，予め文書から画像処理によって文書の構成要素と配置を抽出し，テキストデータ抽出の手がかりとすることを検討している．本稿では，画像認識用深層学習 Single Shot MultiBox Detector (SSD)と文書の階層構造に関する知識を用いて文書の構成要素と配置を抽出方法を提案する．

まず, 2 章では, 筆者らが行った先行研究[3]について述べ, 基本的な概念を定義し，示された課題をもとに文書の構成要素の階層構造に着目する理由を述べる．次に，3 章で提案手法を示し，4 章で実験結果に基づいて提案手法の効果を検証する．

2. 先行研究

2.1 地方自治体の議会活動資料と文書構造

図 1 に，地方自治体の広報誌の例を示す．一般的な傾向として，各ページに大きなタイトルがあり，小見出しや段落，図，表を文書の構成要素としている．文書の構成要素は四角いものが多く，構成要素同士が重なるケースは少ない．文章は，縦書きも横書きも存在し，1 ページの中で混在する場合もある．PDF ファイル内のテキストデータ抽出の手がかりになり得ることから，縦書きと横書きは区別して抽出する．また，図や表とそれらのキャプションのようにセットとなる構成要素についても，テキストデータ抽出の手がかりになり得ることから，図表全体を親，図表とキャプションを子とする階層構造として検出することを目的とする．先行研究で定義した文書構成要素のカテゴリは以下の通りである．カテゴリ名の_hor は横書きを，_ver は縦書きを意味する．

1 福岡大学工学部電子情報工学科
Department of Electronics Engineering and Computer Science, Fukuoka University

- bigtitle_hor（大見出し）
- bigtitle_var
- title_hor（見出し）
- title_var
- document_hor（本文）
- document_var
- table（表）
- fig-area（図や画像のエリア全体）
- figure（図や画像）
- caption（キャプション）

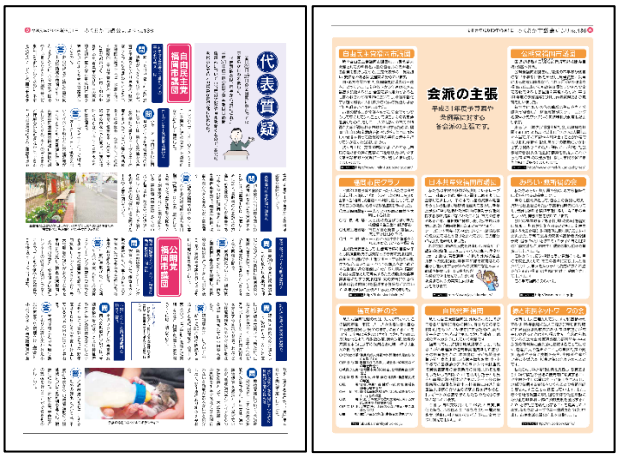

図 1. 自治体広報誌の例

2.2 SSD(Single Shot MultiBox Detector)

Single Shot MultiBox Detector (SSD)[4]は，一般物体認識用に提案された深層学習モデル一つである．一般物体認識では，物体の認識と検出（画像中の位置の特定）を同時に行う．対象物体の形状は多様なので大きさや形，位置の異なる矩形領域を画像から切り出し，それぞれに認識処理を行うため計算コストが増大する．これに対し， SSD では，切り出した矩形領域の認識処理の効率を高める工夫がなされており，1 度の演算（Single Shot）で物体の位置と認識の両方を行うため処理時間が短い．また，画像中に複数の物体が存在する場合の認識精度が従来の深層学習モデルよりも高いとされている．また，近年の深層学習モデルには pixel 単位で任意形状の物体領域を認識するモデル[7]も提案されているが，2.1 節で述べたように地方自治体の活動資料では，文書の構成要素は四角いものが多く，構成要素同士が重なるケースは多くはないことから SSD を選択した．

ここで使用した SSD の実装モデル[5]は， COCO 形式の学習データをサポートしているため，学習データの作成には，COCO 形式の出力が可能なアノテーションツール Fast Label[6]を用いた．

2.3 課題

先行研究では，261 枚の学習画像を用い，253 枚を訓練用に，8 枚をテスト用に用いて検証を行っている．SSD の推論では，検出候補領域におけるカテゴリ認識の信頼度を確率と呼び, 確率があるしきい値以上のものを検出とみなす．テストの際は，検出された個所と重なりを持つアノテーションのうち，重なりが最も大きかったアノテーションを正解候補とし，検出結果のカテゴリと正解候補のカテゴリが一致したら正解とみなす．図 2 には，高精度に検出・認識が行えた例を示す．この例では，文書の構成要素が位置およびカテゴリともに 100%の精度で認識が行えている．テスト全体では，F 値が 66.59%という性能が得られた．

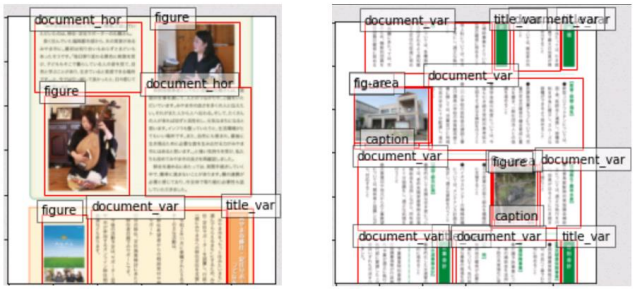

図 2. テストで精度が高かった例

一方，図 3 は誤認識を含むページの例である．カテゴリの正誤を無視すれば，ほぼすべての文書の構成要素が検出はされているものの，縦書き・横書きによらず，bigtitle（大見出し）と title（見出し）の誤認識や，図の中の caption と document（本文）との誤認識が多くみられることが分かった．特に，横書きの大見出し bigtitle_hor と横書き小見出し title_hor の性能が低く，それぞれ F 値が 52.85%，47.86%であった．

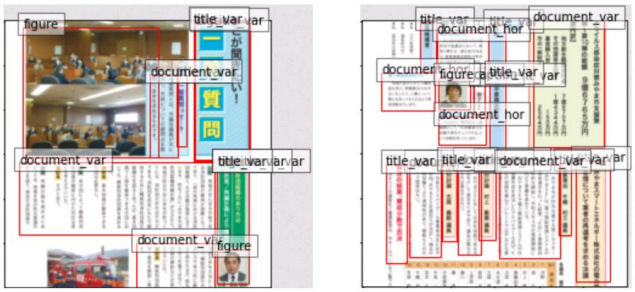

図 3. テストで精度が低かった例

また，2.1 節で示したカテゴリのうち，fig-area(画像領域)と figure（画像）， caption（画像のキャプション）は文書中で入れ子構造になるものであった．しかし，アノテーションにおいて figure や caption は必ずしも fig-area の中に存在するようにアノテーション作業が行われておらず，更には学習データにおけるカテゴリ数においても fig-area の割合は以下の表 1 のようになっており，学習データがあまりに少なく，階層構造として検出することが困難であった．

表 1. 先行研究のカテゴリごとの学習データの割合

カテゴリ	アノテーション数	割合
bigtitle_hor	187	3.72%
bigtitle_var	72	1.43%
caption	416	8.27%
document_hor	778	15.47%
document_var	1170	23.27%
fig-area	357	7.10%
figure	877	17.44%
table	90	1.79%
title_hor	433	8.61%
title_var	648	12.89%
合計	5028	100.00%

3. 提案手法

3.1 文書構造の階層性と定義

本研究においては，文書の構成要素間の階層構造に着目してカテゴリの見直しを行い，精度向上の方法を提案する．以下に，再定義したカテゴリと，カテゴリ間の階層関係（図 4）を示す．

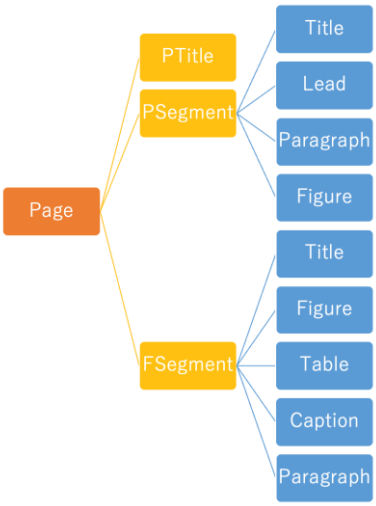

図 4　文書構成要素間の階層関係

ここで，カテゴリ名内の H は横書き，V は縦書きを表す．全てのページには，1 つ以内の大きな見出しがあり，その他の構成要素は大きくは文章（Paragraph）主体もしくは図表（Figure）主体の Segment に分類される．それぞれのセグメント内は，更に以下に示す構成要素からなる．

- PTitleV（ページ内に一つだけ存在する大見出し）
- PTitleH
- PSegment（文章主体の領域）
- TitleV（領域内に一つだけ存在するタイトル）
- TitleH
- LeadV（PSegment 内の 1〜2 行の文章，リード文）
- LeadH
- ParagraphV（領域内における段落）
- ParagraphH
- FSegment（画像主体の領域）
- Figure（図や画像，挿絵，表）
- Table（FSegment 内の表全体）
- CaptionV（図や表に一つ存在するキャプション）
- CaptionH

3.2 学習データの準備

学習データは 215 枚の画像を用いた．アノテーションは，具体例を使ったマニュアルを作成し，アノテーションを事業として行っている業者に依頼した．マニュアルでは判断がつかない事柄に関しては，その都度，質疑応答を行いながら進めた．

3.3 階層構造を用いた検出

カテゴリ間に階層構造を持たせた場合，最適な検出結果を得る問題は，親グループの検出候補と子グループの 2 部マッチング問題と捉えることができる．この考え方に基づく認識・検出精度を向上させる単純な方法として，次の二つの方法が考えられる．

一つは，親カテゴリが高い確率で検出されているにもかかわらず，そこに含まれている子カテゴリが検出されていない場合で，検出時の確率のしきい値を下げ，未検出になっていた子カテゴリ候補を検出とみなすトップダウン方式である．もう一つは，逆に，複数の子カテゴリが高い確率で検出されているにもかかわらず，それらを包含する親カテゴリが検出されていない場合で，検出時の確率のしきい値を下げ，未検出になっていた親カテゴリ候補を検出とみなすボトムアップ方式である．

4. 実験による評価

まず，階層構造の知識を使う前の SSD のみによる性能を示す．実験では，確率が 80%以上のものを検出とみなした．結果は，テスト全体では，平均適合率が 44.25%，平均再現率が 46.64%，平均 F 値が 45.41%であった．先行研究と比べると性能が低下しており，カテゴリ数が増えて難易度が上がったものと考えられる．階層構造も含めて正しく検出することは，2 章で述べた通り元々の目的であるので，この性能低下はやむを得ないと考える．

カテゴリ毎の検出精度を見ると，図領域 FSegment を親とする構成要素が全く検出されていなかった．一方，親の FSegment は，適合率で 85.38%，再現率で 57.22%，F 値では 68.52%と比較的精度が高かった．そこで，先述の階層構造を用いたトップダウン方式を適用した．その結果，未検出だった図 Figure と表 Table の一部が検出とみなされるようになり，性能が向上した（表 2）．

表 2．階層構造を用いた効果（再現率）

	適用前	適用後
PTitleV	85.71%	85.71%
PTitleH	24.21%	24.21%
PSegment	55.47%	55.47%
TitleV	78.46%	78.46%
TitleH	48.45%	48.45%
LeadV	81.14%	81.14%
LeadH	68.83%	68.83%
ParagraphV	83.21%	83.21%
ParagraphH	70.28%	70.28%
FSegment	57.22%	57.22%
Figure	0.00%	**15.11%**
Table	0.00%	**45.65%**
CaptionV	0.00%	0.00%
CaptionH	0.00%	0.00%
平均	46.64%	50.98%

5. おわりに

本研究では先行研究における課題であった階層構造の推定を解決するため，カテゴリの見直しや階層構造を用いた精度向上手法の適用を行った．その結果，図セグメントを親とする図領域と表領域の検出・認識精度を向上することができた．混同行列をみると異なった親の子カテゴリとの誤認識も確認できるため，親カテゴリをまたがった 2 部マッチングアルゴリズムの採用により，より性能が向上するものと考えられる．

謝辞

学習データの作成にご協力いただきました(株)正興電機製作所に感謝申し上げます．なお，本研究は JSPS 科研費 JP22K12740 の助成を受けたものです．

参考文献

[1]Yasutomo Kimura, Keiichi Takamaru, Takuma Tanaka, Akio Kobayashi, Hiroki Sakaji, Yuzu, Uchida, Hokuto Ototake and Shigeru Masuyama: Creating Japanese Political Corpus from Local Assembly Minutes of 47 Prefectures, ALR12, The COLING 2016 Organizing Committee, pp.78-85, 2016.

[2]乙武北斗，内田ゆず，高丸圭一，木村泰知: 構造化データ作成を目的とした PDF 地方議会資料のテキスト抽出に関する分析，第 37 回ファジィシステムシンポジウム講演論文集，pp.431-436，2021.

[3]田中季樹，「Single Shot MultiBox Detector を用いた PDF 文書の領域分割に関する研究」，福岡大学工学部卒業論文, 2023.3

[4]Liu, W. et al. (2016). SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2

[5]https://github.com/amdegroot/ssd.pytorch (参照 2023-12-15).

[6]https://fastlabel.ai/ (参照 2023-12-15).

[7]Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick; Mask R-CNN, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969