Vol.2019-MPS-124 No.6
2019/7/29

帝国会議会議録における semantic segmentation を用いたレイアウト解析

飯田紗也香[†1]　竹本有紀[†1]　石川由羽[†2]　髙田雅美[†1]　城和貴[†1]

概要：本稿では，帝国議会会議録に対する，テキスト化のためのレイアウト解析の手法の提案する．文字認識の精度は文字切り出しの精度に依存する．文字切り出しの精度向上には，文書画像に対しレイアウト解析を行う必要がある．一般的に，レイアウト解析にはヒストグラムによる手法を用いる．しかし，ヒストグラムのパターンのみを用いて汎用的なレイアウト解析を行うことは難しく，正確な文書領域の抽出には，目視で文書の構成を考慮する必要がある．そこで，Semantic Segmentation を用いたレイアウト解析手法を提案する．提案手法を帝国議会会議録に適用し，ヒストグラムによるレイアウト解析の場合と切り出された文字数を比較し，提案した手法の有用性を確認する．

キーワード：レイアウト解析，帝国議会会議録，Semantic Segmentation

Layout analysis using semantic segmentation for Imperial meeting minutes

SAYAKA IIDA[†1]　YUKI TAKEMOTO[†1]
YU ISHIKAWA[†2]　MASAMI TAKATA[†1]
KAZUKI JOE[†1]

1. はじめに

国立国会図書館[1]は帝国議会会議録検索システム[2]という Web サービスを提供している．帝国議会会議録検索システムでは，明治から昭和初期の帝国議会全会期の本会議および速記録をデジタル画像で公開している．会議録はフォントが規格化される以前の活版印刷の文書であり，時代によりフォントおよび書式が異なる．会議録画像は目次や索引，発言者で検索が可能である．また，1945 年以降の会議録については，すでにテキスト化がされており本文内容での検索が可能である．しかし， 1944 年以前の会議録にはテキストデータが存在しないため，本文内容で検索不可能である．そこで，1994 年以前の会議録についてもテキスト化が求められている．

帝国議会会議録検索システムで公開されている会議録の枚数は約 20 万枚と膨大であり，人手によるテキスト化は難しい．フォントが規格化される以前の活版印刷の文書画像のテキスト化手法[3]が提案されているが，認識率向上のためには膨大な量の学習データが必要である．また，会議録のテキスト化を行うには，会議録画像に対するレイアウト解析が必要である．文字認識の精度は文字切り出しの精度に大きく依存する．文字切り出しの精度は文書画像からの文書領域抽出の精度に依存する．会議録画像中の文書は段組みになっており，文書以外に直線を含む．会議録画像から文書領域を抽出するにはレイアウト解析を行う必要がある．

一般的に，文書画像に対するレイアウト解析にはヒストグラムによる手法を用いる．ヒストグラムを用いた手法では，画素射影ヒストグラムの形状や変化量からパターンを検出し，文書画像から文書領域を抽出する．しかし，文書画像に含まれる文書以外の要素の違いにより，ヒストグラムのパターンのみを用いて汎用的なレイアウト解析を行うことは難しい．正確な文書領域の抽出には，目視で文書の構成を確認する必要がある．しかし，現在帝国議会会議録検索システムで公開されているテキスト化されていない会議録画像の枚数は膨大である．1 枚ずつ目視でレイアウトを確認し切り出しを行うことは非効率的である．そこで，本稿では会議録画像に対して Semantic Segmentation[4]を用いたレイアウト解析手法を提案する．提案手法を帝国議会会議録に適用し，ヒストグラムによるレイアウト解析の場合と文字切り出しの精度を比較し，提案した手法の有用性を示す．

以下，本稿の構成を示す．2 章で既存のレイアウト解析の手法であるヒストグラム法について述べる．3 章では，Semantic Segmentation を用いたレイアウト解析の手法の提案を行い，4 章では，実験方法を述べる．5 章でまとめについて述べる．

†1 奈良女子大学
Nara women's University
†2 滋賀大学
Shiga University

2. 既存のレイアウト解析手法

既存の文書画像のレイアウト解析はヒストグラムを用いた矩形の作成から成る. ヒストグラムを用いた手法では, まず文書を白黒の 2 値画像に変換し, 文書を構成する黒画素の射影ヒストグラムを求める. ヒストグラムの形状や変化量からパターンをみつけて線形に分割する. 切り出されたパターンを矩形で囲み, その矩形の面積や縦横比から判断して文書を切り出す. しかし, 文書画像に含まれる文書と文書以外の要素の違いを, ヒストグラムのパターンのみを用いて判別することは難しい. 正確な矩形の作成には, 目視で文書の構成を考慮する必要がある.

矩形の作成にはボトムアップな方法とトップダウンな方法の 2 つがある. トップダウン法では大まかなレイアウト構造を解析し, 徐々に細かい矩形を作成していく. たとえば文字列の境界, または段落の境界であると予想される空白に沿って画像を垂直方向もしくは水平方向に切り出し文書領域の矩形を作成する. トップダウン法では, 文書画像中に長方形ではない領域や縦横に混在する見出しなどを含む場合, それらを考慮できない. ボトムアップ法は画像から検出された矩形を徐々に統合する方法である. 文書画像中の小さなパーツを分類して, パーツ間の距離や形状, 面積を考慮し矩形を統合する.

本稿で対象とする帝国議会会議録を図 1 に示す. 会議録は 2 段から 5 段の複数の段に分かれており, 表題付きと表題無しのものがある. 文書の段組や表題は枠線で囲まれている. つまり, 会議録に対するレイアウト解析を行うには, ヒストグラムから段の境界と直線のパターンを検出する必要がある. しかし, 会議録画像に含まれる枠線は完全な直線ではなく, 会議録の原本のたわみや皺, 撮影時のずれにより歪んでいる. 枠線同様に文字の並びがずれている場合もある. また, 会議録にはインクの染みなどのノイズが含まれている. つまり, 会議録に対して水平および垂直方向の黒画素射影ヒストグラムのみを用いたレイアウト解析は難しく, ある程度目視で矩形の大きさや位置の調整を行わなくてはならない. 提案手法では, 領域抽出の際に目視による調整が必要ない Semantic Segmentation を用いる. Semantic Segmentation は同じオブジェクトに属する画素を自動でクラス分けする手法であり, 画像からの領域抽出に適している. 画像から文字領域や文書領域, 枠線領域をそれぞれ抽出し, ボトムアップな手法によるレイアウト解析を行う.

3. 提案手法

会議録画像に対するレイアウト解析の手法として, Semantic Segmentation を用いたレイアウト解析を提案する. Semantic Segmentation には SegNet Basic アーキテクチャを用いる. 提案手法では Semantic Segmentation の出力するセグメンテーションマップを利用し, レイアウト解析処理を

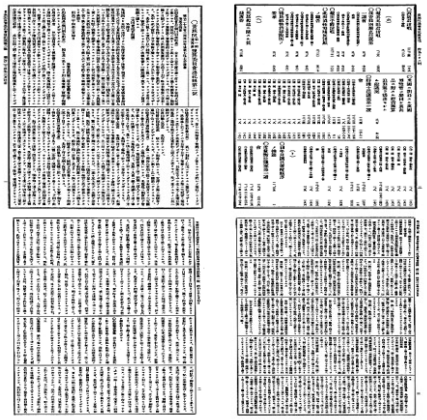

図 1　帝国議会会議録画像

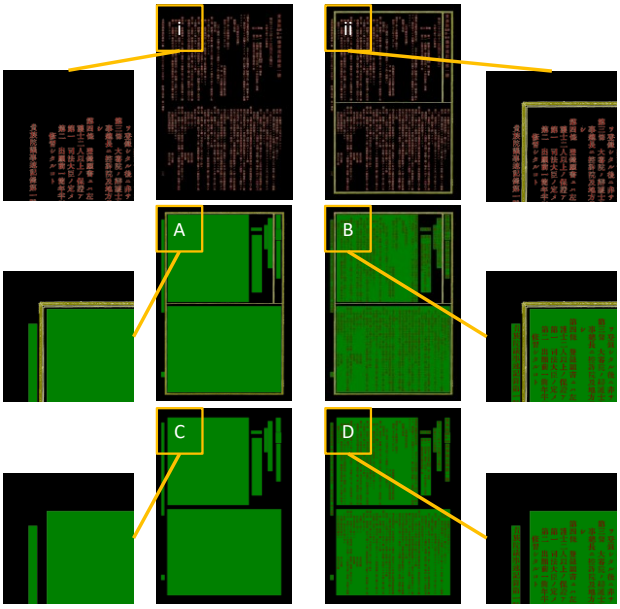

図 2　学習させるラベル 6 種

行う. また, 会議録画像の学習を行う際, 会議録画像に異なるフィルタサイズのガウシアンフィルタをかけ比較する.

Semantic Segmentation を行うため, モデルに会議録画像と対象となるラベルを対として学習させる. 本稿で使用したラベルは以下の 6 種類である.

- ラベル i　文字領域
- ラベル ii　文字領域と枠線領域
- ラベル A　文書領域と枠線領域
- ラベル B　文書領域と文字領域, 枠線領域
- ラベル C　文字領域
- ラベル D　文書領域と文字領域

モデルに学習させるラベルは図 2 の 6 種類である. ラベルについて, 赤色の領域が文字領域, 黄色の領域が枠線領域,

緑色の領域が文書領域である．ラベル i，ii は文書領域を含まないラベルであり，ラベル A から D は文書領域を含むラベルである．ラベル i, ii を用いたレイアウト解析処理では，ボトムアップ法と同様に細かい領域を統合し，矩形を作成する．ラベル A から D を用いたレイアウト解析処理ではトップダウン法と同様に，大まかに構造を抽出してから細部の矩形を作成する．レイアウト解析を行う目的は文書領域および文字領域を抽出することなので，枠線領域のみを含むラベルは用意していない．会議録画像は白黒二値画像である．白黒二値画像の色は0か255のみで表現されており，文書領域のようなラベルの境界が空白中に存在する画像のセグメンテーションは困難であると推測される．そこで，会議録の文字間の空白をガウシアンフィルタによるぼかしで埋め，文書領域の境界判別の補助とする．ガウシアンフィルタでは式(1)のガウス関数の畳み込みを用いて画像の平滑化を行う．本稿で用いるガウシアンフィルタのカーネルは正方形とする．

$$Gauss(ksize) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{2ksize^2}{2\sigma^2}) \quad (1)$$

カーネルの幅と高さは奇数である．σの値は式(2)を用いてカーネルサイズから決定する．

$$\sigma = 0.30(\frac{ksize- 2}{2}) + 0.80 \quad (2)$$

画像にガウシアンフィルタをかけた場合と，フィルタをかけない場合で比較する．文字間の距離を約 60px として，ガウシアンフィルタのカーネルサイズは 60 以下とする．同時に学習させるガウシアンフィルタのサイズ$K(n)$のパターンの条件は次の通りである．

- 条件 1：ガウシアンフィルタなし
- 条件 2：1 飛ばしの異なるカーネルサイズを用いる

$$K(n) = 2n - 1 \ \{n \in \mathbb{N} \mid 0 < n < 30\} \quad (3)$$

- 条件 3：5 飛ばしの異なるカーネルサイズを用いる

$$K(n) = 6n - 5 \ \{n \in \mathbb{N} \mid 0 < n < 10\} \quad (4)$$

- 条件 4：9 飛ばしの異なるカーネルサイズを用いる

$$K(n) = 10n - 9 \ \{n \in \mathbb{N} \mid 0 < n < 6\} \quad (5)$$

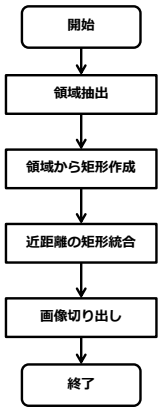

図 3 文字切り出し処理流れ

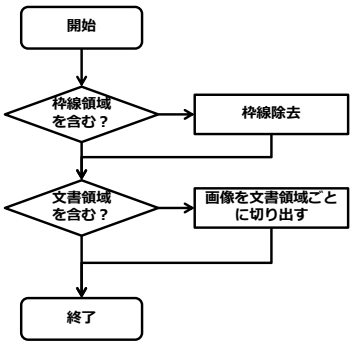

図 4 Semantic Segmentation の出力を用いた前処理

学習はラベル 6 種ごと，ガウシアンフィルタのカーネルサイズの条件 4 つごとの計 24 パターン行う．学習させる会議録画像の枚数は 5 枚である．計算量を減らすため，会議録画像は 1 枚を 1 辺 1024px の正方形に 20 分割して学習させる．

提案手法の流れについて説明する．初めに，Semantic Segmentation を行い領域抽出する．次に，Semantic Segmentation の出力を用いた前処理を行う．最後に，文字切り出し処理を行う．文字切り出し処理では文字領域を出力する場合，セグメンテーションマップ中の文字領域ラベルを用いる．

提案手法で用いる文字切り出し処理の手順について説明する．処理の流れは図 3 の通りである．文字切り出し処理ではボトムアップなレイアウト解析手法を用いて文書中の文字を 1 文字ずつ切り出す．まず，領域抽出処理では細かいパーツの領域を検出する．検出する領域は画像内の要素の最も外側の輪郭である．次に，検出した領域を囲む矩形を作成する．作成した矩形が重なる場合や矩形の距離が近い場合，矩形を統合する．最後に，作成した矩形で画像を切り出す．文字切り出し処理で用いる領域抽出では，検出する図形の最も外側の輪郭を検出する．そのため，文書画像上のインクの染みや枠線が領域抽出の妨げになる．また，文字を構成するパーツの距離が遠い文字はうまく切り出されない．

Semantic Segmentation を用いたレイアウト解析処理の流れを説明する．まず，学習済みのモデルを用いて会議録画像に Semantic Segmentation を用いたラベル分けを行う．次に出力されたセグメンテーションマップを用いて元画像の前処理を行う．処理の流れは図 4 のとおりである．出力されたセグメンテーションマップが枠線領域を含む場合，元画像から枠線と判別された領域中の黒画素を白画素へ置き換える．出力されたセグメンテーションマップが文書領域を含む場合，画像を文書領域ごとに切り出す．最後に，文字切り出し処理を行う．文字切り出し処理の流れは先に説明したとおりであるが，出力されたセグメンテーションマップが文字領域を含む場合，文字領域を用いて領域抽出を行う．

4. 実験方法

本稿では，帝国議会会議録画像を対象に文字切り出しの精度の比較を行う．対象とする会議録画像はモデルに学習された会議録画像とモデルに学習されていない会議録画像の 2 枚である．会議録画像に対して，ヒストグラムによる手法，提案手法を適用し文字切り出しを行った結果，正しく切り出された文字の割合を比較する．比較した結果から，Semantic Segmentation を用いたレイアウト解析の有用性を確認する．ヒストグラムによる手法では画素射影ヒストグラムを用いて会議録画像を段ごとに分割する．また，提案手法で出力するセグメンテーションマップ 6 種とガウシアンフィルタの条件 4 つの 24 パターンの結果について比較を行う．学習された画像の評価対象文字数は 1293 個，学習されていない画像の評価対象文字数は 837 個である．文字切り出し処理では文字を構成するパーツの距離が遠い場合，矩形の統合に失敗する．このような文字を評価対象外とした．

5. まとめ

本稿では，帝国議会会議録のテキスト化を目的として，Semantic Segmentation を用いたレイアウト解析手法を提案する．提案手法では，画像の各画素に意味を自動的に割り当てる Semantic Segmentation を用いて，テキスト領域，枠領域，文書領域を抽出する．Semantic Segmentation に用いる CNN アーキテクチャは SegNet Basic である．学習するラベル画像に含まれるオブジェクトの組み合わせは 6 種である．また，白黒二値画像に対する領域抽出の補助を目的に異なるカーネルサイズのガウシアンフィルタを画像に適用し同時に学習させる．

提案手法の有用性の検証のため，ヒストグラムを用いたレイアウト解析手法と提案手法によって文字切り出し処理を適用した結果，切り出された文字の割合で比較する．また，提案手法では，ラベルに含まれるオブジェクトの種類とガウシアンフィルタの条件を変更し，文字抽出処理によって抽出された文字の割合を比較する．ヒストグラムを用いたレイアウト解析手法では画素射影ヒストグラムを用いて会議録を段ごとに切り出した．提案手法では Semantic Segmentation の出力を用いたレイアウト解析を行い，文書領域を切り出した．ラベルに含まれるオブジェクトの種類と学習される画像に適用するガウシアンフィルタの条件を変えて，文字抽出処理によって抽出された文字の割合を比較する．ラベルは 1 つ目が文字領域，2 つ目が文字領域と枠線領域，3 つ目が文書領域と枠線領域，4 目が文書領域と文字領域，枠線領域，5 目が文字領域，6 目が文書領域と文字領域の 6 種類である．ガウシアンフィルタの条件は 4 種類である．

Semantic Segmentation を行う CNN に学習された画像，学習されていない画像に対して，ヒストグラムを用いた手法，提案手法を用いて文字切り出しを行い，それぞれの正確に切り出された文字数を比較する．また，ラベルおよびガウシアンフィルタの条件ごとの Semantic Segmentation の結果を比較する．

謝辞　本研究は文部科学省科学研究費補助金(17H01829)の助成を受けたものである.

参考文献

[1] 国立国会図書館 (参照　2019-06-19)
http://www.ndl.go.jp/
[2] 帝国議会会議録検索システム (参照　2019-06-19)
http://teikokugikai-i.ndl.go.jp/
[3] FUJIMOTO, Kaori, et al. Early-Modern Printed Character Recognition using Ensemble Learning. In: Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2017. p. 288-294.
[4] THOMA, Martin. A survey of semantic segmentation. arXiv preprint arXiv:1602.06541, 2016.
[5] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.