様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １７ 日現在

機関番号：２０１０３
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００２４１

研究課題名（和文）文書画像アーカイブに対するテキスト情報に依存しない内容解析

研究課題名（英文）Image-based contents analysis for untranscribed document image archives

研究代表者
寺沢　憲吾（Terasawa, Kengo）

公立はこだて未来大学・システム情報科学部・准教授

研究者番号：１０４３５９８５

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：本研究では、画像特徴に基づいて特定の文字列の出現頻度や出現パターンを解析することで、機械判読が困難である文書画像に対しても、頻出語の抽出や、抽出された頻出語の重要度の評価を可能にし、さらにはこれを用いて各文書の内容を要約したり、特定のトピックと関連の高い箇所を抽出したりすることで、文書画像デジタルアーカイブの有効活用を促進する手法を開発した。また、明治期の未翻刻の新聞画像を対象に実証実験を行い、開発した手法の性能と有効性を確認した。

研究成果の学術的意義や社会的意義
本研究の成果により、手書きであったり経年劣化を経ているなどの理由で機械判読が困難である文書画像に対しても、その内容の要約や、特定のトピックと関連の高い箇所を閲覧者に提示することが可能となる。これにより、各地で整備が進み蓄積されている文書画像デジタルアーカイブが、専門研究者のみならず、一般市民や地域史に興味を持つ人々などにとっても、使いやすく便利な文献資料として、その価値を高めていくことが期待される。

研究成果の概要（英文）：In this study, we achieved the extraction of frequently appearing words and evaluation of the importance of the extracted words from machine-unreadable untranscribed document images, using image-based analysis of the frequency and pattern of occurrence of certain text strings. We also achieved to summarize the content of each document and to extract the part that is highly related a specific topic. We conducted a experiment on untranscribed newspaper images published in Meiji Era, and confirmed the performance and effectiveness of the proposed method. Our achievement will promote effective use of digital archives of document images.

研究分野：画像、文章、音声等認識

キーワード：画像、文章、音声等認識　パターン認識　データベース　デジタルアーカイブ

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

文書をデジタル化して保存することはすでに一般的となっており、我が国でも、国立国会図書館、国立公文書館、国文学研究資料館などにおいて、総計 400 万点以上の文献資料がデジタルアーカイブとして蓄積されている。このように蓄積された文書画像の多くは予算等の制約からテキストデータ化されておらず画像データとして提供されているため、こうした文献画像を含むデジタルアーカイブの有効活用のために新たな情報検索の手法の開発が望まれている状況にあった。

一つの案は光学文字認識（OCR）の技術を用いて画像データからテキストデータを得ることであるが、従来型の手法では手書きを含む歴史的文書画像に対する認識精度は十分ではなく、従来手法とは異なる新しい文字認識の手法を開発することが必要である。もう一つの案はテキスト化を前提としない内容解析の手法を開発することである。画像特徴量のみに基づいた文書画像処理に関する研究としては、テキスト化せずに全文検索を行うワードスポッティングなどの研究が代表的である。

２．研究の目的

本研究は、手書き筆記における書体や字形のゆらぎ、あるいは経年劣化などの理由により機械判読が困難である文書画像を対象に、画像特徴に基づいて特定の文字列の出現頻度や出現パターン、あるいは複数の文字列の共起性などを解析することにより、文書画像に記載された文字列の構造を解き明かしていくことを目的とする。このことは、これまで質的に解読が困難であるとされてきた文書に対する有力な解読手段を提供するとともに、量的にテキストデータ化が困難とされてきた文書画像に対してテキストデータを作成する手段の提供にもつながる。すなわち、本研究の目的の第一はテキスト化を前提としない内容解析の手法を開発することであり、第二はそれを活用した従来手法とは異なる新しい文字認識の手法を開発することである。これらはいずれも近年および現在において増加しつつある文献資料のデジタルアーカイブを広く一般に活用してもらうためにも、専門家に深く活用してもらうためにも有益なものとなる。

この目的を見据えつつ、本研究課題の研究期間内においては、テキストに基づかない内容解析のうち、キーワード抽出や重要度の評価といった基盤となる部分の手法を確立することを目指す。

３．研究の方法

本研究では、歴史的文書画像の中でも歴史的に重要な情報が含まれる歴史的新聞画像を対象とし、新聞の特性を利用したキーワード抽出手法を提案する。新聞画像が単文字ごとにセグメント化されているところまで前処理で行われていることを前提に、各単文字画像から画像特徴量に基づく特徴ベクトルを抽出し、新聞画像をこの特徴ベクトルの系列データと考え、これに対してさまざまな解析を行う。解析においては、各特徴ベクトルをクラスタリングを用いて離散化し、割り当てられたクラスタ ID をその文字画像に対応する擬似コードとすることで、解析対象を離散データの系列として扱うことができるようになる。

実験評価にあたっては、1881 年（明治 14 年）に発行された「函館新聞」の画像データ（771 枚）のうち、全面広告等のページを除いた 474 枚を対象として用いる。対象となる画像に含まれている文字数はおよそ 100.7 万字である。この時期に発行された新聞は歴史資料としての価値がある一方で、分量が多いため全文テキスト化はなされていない。光学文字認識（OCR）を用いるとしても、この時期の新聞は活字で印刷されているものの字種や言葉の用法が現在と異なり、また印刷も鮮明ではなく経年劣化もあるため、現代の日本語を対象とした既存の光学文字認識手法をそのまま適用した場合は認識精度が著しく低い。こうした特性から、この画像データベースは本研究の実験素材として適切である。

４．研究成果

(1) 本研究において土台となるのは、単文字画像から画像特徴量に基づく特徴ベクトルを抽出する部分と、各特徴ベクトルを離散化して擬似コードとする部分である。本研究課題の研究期間内において、これらの手法は数度の改良を経て、離散化の精度の向上を得ている。

図１は本研究による画像特徴量抽出およびその離散化の結果を示したもので、図中左側の文書画像において赤丸で囲まれた「徴」という文字が、図中右側に示される擬似コードにおいて、すべて同じコード「245」が割り当てられている状態が示されている。その他の文字についても、同一の文字に対してはおおむね同一のコード

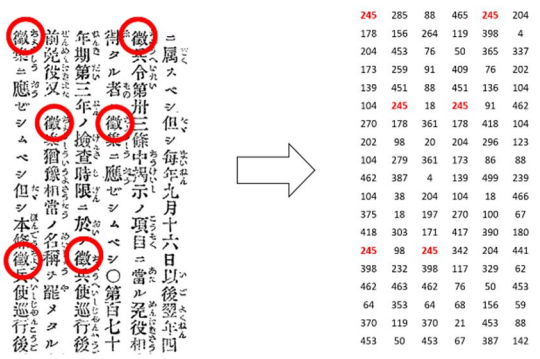


図１　文書画像の擬似コード系列化

が割り当てられており、良好な結果となっている。この結果を得るに至るまでの改良の要点は以下の通りである。

① 疑似コード化にあたっては、同一文字種に別コードが割り当てられるケースと、異なる文字種に同一のコードが割り当てられるケースの二つの不具合が起こりうるが、そのうち前者のケースを極力減らすようにクラスタリング手法を改善した。具体的には、k-means 法を単純に適用するのではなく、k の値を大から小に変化させて二段階で適用することで、前者の不具合を相当程度抑制できる手法を採用した。また、文書集合の状態から最適なクラスタ数を調整する手法も採用した。さらに、k-means 以外の GMM などのクラスタリング手法とも比較検討し、最適なものを選べるようにした。

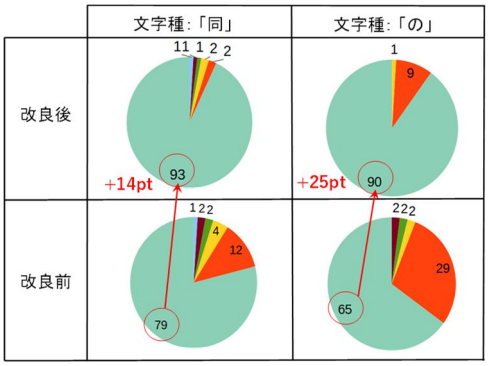


図２ 同一文字種に対する擬似コード割り当て分布

② 単文字画像からの特徴ベクトルの抽出においては、従来は HOG に代表される画素値の勾配分布に基づく特徴量を採用してきたが、機械学習のエンコーダ・デコーダモデルが盛んに研究され、優秀な結果を記録しているという近年の状況に鑑み、本研究においてもエンコーダ・デコーダモデルの一種であるオートエンコーダを採用した。前述の①の改良と合わせて、図２に示すように、同一文字種に別コードが割り当てられるケースを大幅に抑制できた。

(2) 擬似コードを解析し、頻出語抽出およびその重要度評価を行うことで、文書の内容を特徴付ける重要語を抽出する手法を開発した。重要度評価においては、一般的に広く用いられている TF-IDF では十分な精度が得られなかったため、本研究の目的に則した新たな手法を開発し、比較検討した。その中で最も有効であった方法は、文字列の出現間隔に着目する方法であった。具体的には、特定の文字列の出現間隔に着目し、それが重要度の低い（文書内容を特徴づける度合いが低い）語であればその出現間隔は偶然の確率分布（今回の場合幾何分布）に従い、逆に重要度の高い語であればその出現間隔は幾何分布とは異なる様態を示すと仮説を立て、ある特定の語に対し観測した出現間隔の分布の幾何分布への適合度合いを測ることで、語の重要度を評価する手法である。より具体的には、v を特定の文字列の特定の発行日における出現を表す二値ベクトルとし、0 を特定の文字列が含まれる発行日の集合、I をその間隔の集合として図３のように求める。I から求まる分布を F(t)、幾何分布を G(t)として、これらに対してカイ二乗検定を用いて、適合度合いを求める。

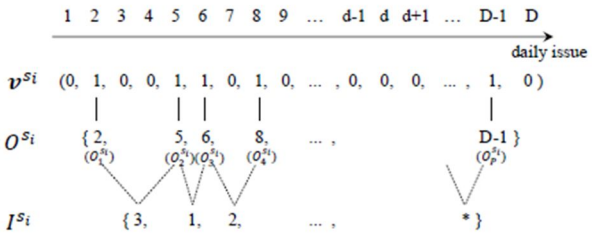


図３ 文字列の出現間隔の分布の計算方法

この手法は未翻刻の歴史的文書画像からの重要語抽出のために開発した手法であるが、手法自体は現代語にも適用可能であるため、予備実験として、現代語の新聞コーパスである「読売新聞コーパス 2017 年版」を用いた評価実験も行った。その結果、この手法はある程度の精度でその発行日の記事内容を特徴づける語を抽出することが可能であることがわかった。

予備実験の結果を受け、本実験として、1881 年（明治 14 年）に発行された「函館新聞」を対象として用いた実験を行った。この際、抽出すべき重要語としては、1974 年（昭和 49 年）に発

表１ 「函館新聞」（1881 年）からのキーワード抽出結果

| 特徴ベクトル抽出手法 | クラスタリング手法 | クラスタ数 | 官有物 | 関西貿易 | 黒田長官 | 玄武丸 | 小林重吉 | 山本忠礼 | 田中正右衛門 | 常野正義 | 抽出数 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HOG | k-means | 300 | ○ | ○ | ○ | ○ | | ○ | | | 5 |
| HOG | k-means | 400 | | ○ | | | | | | | 1 |
| HOG | k-means | 500 | ○ | ○ | ○ | ○ | | | | | 4 |
| AE | GMM | 300 | ○ | ○ | | ○ | ○ | | | | 4 |
| AE | GMM | 400 | ○ | ○ | ○ | | | | | | 3 |
| AE | GMM | 500 | ○ | ○ | ○ | ○ | | | | | 4 |

行された「函館市史 通説編 2」の中から 1881 年の函館新聞の記事を基に書かれている記事に着目し、これらの記事の中で出現する特徴的な名詞・固有名詞を目視で抽出してリスト化した後、文書頻度が 10 以上 19 以下であるもののみを選定した。実験の結果、本研究で開発した手法により、単一手法では選定した 8 つの重要語のうち 5 つ、複数手法を併用することにより 6 つの重要語を抽出することができるという成果が得られた（表 1 ）。

(3) 前述の手法で得られた重要語抽出手法を用い、抽出された重要語の出現位置近辺を文書画像上でハイライト表示することにより、文書内容を読むことなく、その文書中で注目すべきトピックが記述されている箇所を読者に提示する手法を開発した。図 3 は 1881 年 9 月 7 日の函館新聞の画像で、左から順に 1 ページ目から 3 ページ目までを示したものであり、抽出されたキーワード近辺が赤色でハイライトされている。1 ページ目から抽出された単語は「聖上に」、「行在所」、「黒田長」、「田長官」、「右衛門」、「四年九」の 6 つであった。「聖上」は天皇（天子）につける敬称であり、「行在所」は天皇の行幸の際に旅先に設けた仮宮のことである。このように、天皇の動向に関する単語がキーワードとして抽出されることにより、この日の記事の中で特に注目すべき記事として提示することができる。また「黒田長官」は(2)で述べた抽出目標キーワードのリストに含まれるようにこの記事の時期に起きた重大な事件の一つである「開拓使官有物払下げ事件」に関わる人物の名前であり、この人物名をハイライト表示することで、「開拓使官有物払下げ事件」に関する記事を注目すべき記事として提示することができる。一方、2 ページ目以降にはハイライト表示がほとんどないことから、これらのページには連日掲載されるような人々の関心が高い記事が含まれないことがわかるので、関心が高い記事が掲載されているページとそうでないページをひと目で見分けることが可能となる。

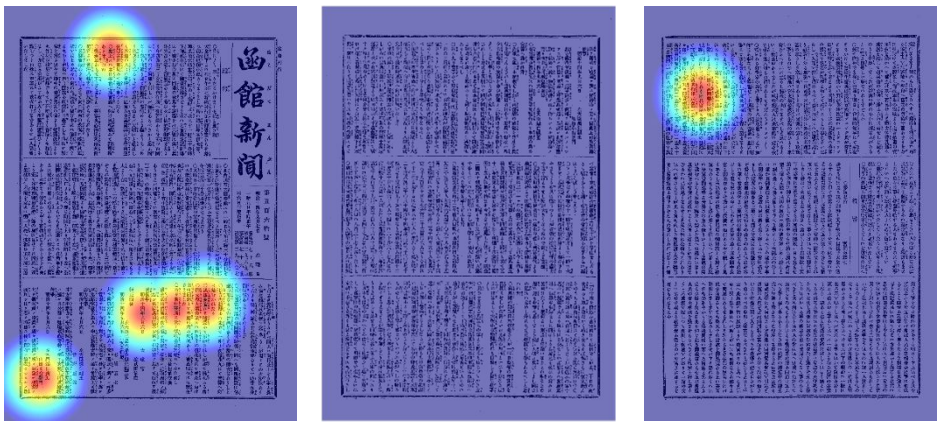


図 4 「函館新聞」（1881 年）において抽出された重要語近辺をハイライト表示した図

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　1件／うち国際学会　1件）

| |
|---|
| １．発表者名<br>Sora Ito and Kengo Terasawa |
| ２．発表標題<br>Extraction of Distinctive Keywords and Articles from Untranscribed Historical Newspaper Images |
| ３．学会等名<br>International Workshop on Advanced Image Technology, IWAIT2020（国際学会） |
| ４．発表年<br>2020年 |

| |
|---|
| １．発表者名<br>伊藤空, 寺沢憲吾 |
| ２．発表標題<br>文字認識が困難な文献史料画像の解析のための文字画像クラスタリング手法 |
| ３．学会等名<br>電子情報通信学会技術研究報告PRMU |
| ４．発表年<br>2018年 |

| |
|---|
| １．発表者名<br>寺沢憲吾 |
| ２．発表標題<br>歴史的文書画像に対する内容解析への取り組み |
| ３．学会等名<br>情報処理学会第116回人文科学とコンピュータ研究会発表会（招待講演） |
| ４．発表年<br>2018年 |

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|