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　 ８ 日現在

機関番号：３３９０８
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ０３０４９

研究課題名（和文）日本近代公文書の自動解読システム開発のための基盤構築の研究

研究課題名（英文）Research on Building Foundations for Developing a System for Automatic Decipherment of Modern Japanese Official Documents

研究代表者
山田　雅之（Yamada, Masashi）

中京大学・工学部・教授

研究者番号：９０２６２９４８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：近代の手書き文字による公文書を、コンピュータが自動的に解読するシステムの実現を目指し、その基盤となるデータセットと要素技術を開発した。題材として台湾総督府文書を利用し、約1700ページ、37万個の手書き文字に関する字形・字種の情報を有するデータセットを開発した。また、深層学習に基づく文字切り出し技術と個別文字認識技術を開発した。台湾総督府文書を対象とした実験では、それぞれの技術の精度は95%と89%に達した。

研究成果の学術的意義や社会的意義
近代公文書は近世古文書の流れを汲む近代古文書のため解読は容易ではなく、その自動解読システムを開発できれば、広く一般の国民や外国人研究者が近代公文書を利用できるようになる。本研究が開発したデータセットと要素技術は自動解読システム開発のための基盤となるものであり、また、文書認識技術の研究分野の発展に寄与するものである。

研究成果の概要（英文）：This research developed a dataset and elemental technologies which are foundations for developing an automatic decipherment system of handwritten official documents in modern era. The developed dataset includes information on shape and classes of about 370,000 handwritten characters sampled from 1,700 pages of the documents of Government-General of Taiwan. The developed techniques of character segmentation and isolated character recognition achieved the accuracy of 95% and 89% respectively in the experiment using the documents of Government-General of Taiwan as test samples.

研究分野：情報工学

キーワード：史料研究　近代公文書　データセット開発　手書き文書認識システム

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

現在、各行政機関等が保管している公文書のうち戦前期の文書の多くは、近世古文書の流れを汲む近代古文書のため一般行政職員が解読するのは容易ではない。そのうえ、近年の活字離れの進行により将来的には多くの公文記録が死蔵状態に陥る可能性すらある。一方、現在の我が国の外交に大きな影響を及ぼしている歴史認識問題の一つに、歴史的事実に対する錯誤があり、その原因として歴史史料を正確に理解出来ていない点が指摘されている。それゆえ如何に公文書史料を手軽に読めるようにするかが課題となっている。事実、その実情は、台湾総督府文書（図 1）の利用状況から見ることが出来る。中京大学社会科学研究所は三十数年にわたって文書を一般に利用できるようにするため、文書目録の編纂とデジタルデータ化、文書史料情報のメタデータの提供とを行ってきた。しかし、台湾人若手研究者の原文書の利用は決して多くはない。その原因の一つが、日本の古文書でもある原史料を容易に読むことが出来ないということにある。したがって、歴史的公文書ともなっている近代公文書を広く一般の国民が利用できるようにするとともに、多くの外国人研究者にも利用できるようにするため、近代公文書の自動解読システムの開発は喫緊の課題となっているといえよう。また、文書史料を文字情報レベルまでデジタル化し、整理可能な状態にするメリットは多岐にわたる。例えば、史料をデジタル化・構造化し、検索・分析を支援するデジタルアーカイブズの開発や、膨大な史料を統計的に分析する計量文献学的手法の適用が容易となる。このような史料研究の新たな展開のためにも自動解読システムは不可欠であろう。

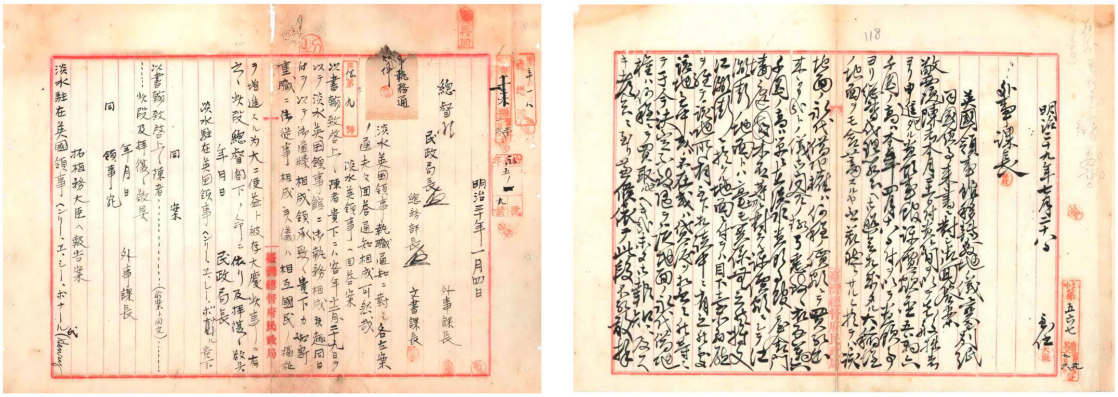

図１　台湾総督府文書の例

自動解読システムの核となるものは手書き文書認識技術である。これは字形情報と文脈情報を使って、注目している手書き文字の字種を推定する技術である。ここで、字形情報とは手書きの文字の 2 次元パターンであり、文脈情報とは、注目している文字の周辺パターン（前後の文字列など）である。手書き文書認識技術の進歩は近年めざましいが、近代公文書は次のような特徴を持つため自動解読は容易でない。

1) 文字の特徴：新・旧字体、略字、崩し字など様々な字体の手書き文字である(図 2) 。
2) 文字配置の特徴：文字のサイズや間隔が一定でなく、修正や挿入文も現れる(図 2)。
3) 言語的特徴：古語的語句・文体が使われる。

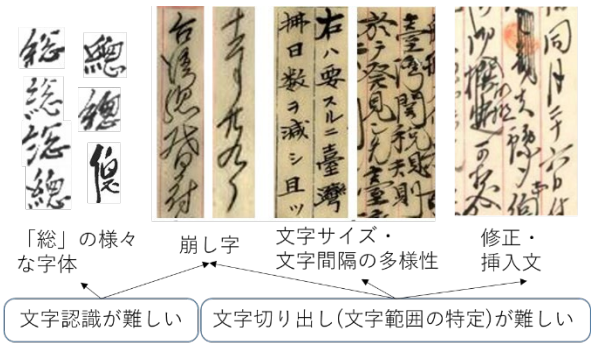

図 2 近代公文書は多様な字体・文字配置があり文書認識が難しい

現在、深層学習という機械学習手法を用いる認識技術が主流となっている。深層学習では予め用意したデータセット（十分な量の基礎データの集合）に基づき認識対象の特徴を学習する。本研究の場合、近代公文書の上記特徴を網羅するデータセットを用意する必要がある。また、本研究のみでなく、他の研究機関においても日本古文書を対象とした自動解読の試みは始まりつつあるが、既存技術のみでは十分な精度は得られておらず、より高い精度の達成には新規技術が必要となっている。

２．研究の目的

本研究は、近代公文書自動解読システム開発のための基盤を構築すること目的とする。ここで基盤とは、(a) 近代公文書の上記特徴を網羅する基礎データの集合（データセット）と、(b)システムを構成する要素技術のことである。前者のデータセットについては、近代公文書の画像データ、画像における個々の手書き文字の座標データと字種データ（図 3)、および、古語的語句・

慣用表現のリストにより構成する。後者の要素技術については、文書画像から個々の手書き文字を検出する文字切り出し技術と、切り出した文字を個別に文字認識する個別文字認識技術を開発する。

近代公文書自動解読システム開発のための上記の基盤を構築できれば、次のステップとして、広く利用可能な自動解読システムの構築に着手できる。また、近代公文書の自動解読という最終的な目標が達成できれば、そのシステムを構成する技術は、国公私立各機関、地方自治体、郷土資料館、図書館、企業、病院、国土地理院などが保存管理している古文書、地方文書、カルテ、日記など多様な文書の解読にも応用可能である。

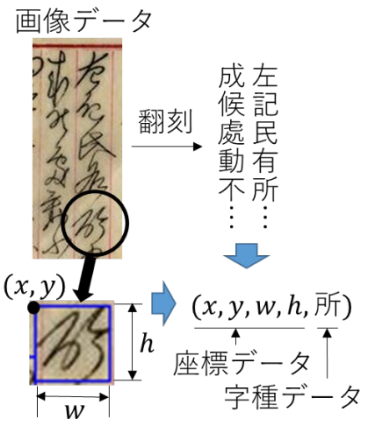

図 3 データセット：翻刻し、文字の座標・字種データを記録

３．研究の方法

本研究は、基礎データの採取および諸所の実験のため、具体的な題材として、台湾総督府文書を利用する。台湾総督府文書は明治 8 年(1895 年)～昭和 20 年(1945 年)までのあらゆる種類の公文書（上奏文、法令命令文、内閣文書、各省庁などの関連文書）が原型のままに残された雛形的存在であり、その量は 13,146 簿冊（１簿冊約 500 ページ）にのぼる。研究分担者らが 1982 年から開始した台湾総督府文書研究（現在は台湾総督府文書目録の編纂と同時に進めている台湾総督府文書史料検索データベースの構築）で蓄積してきた知識と古文書解読の知識を活用することで効率よく基礎データを採取できる。

（１）データセットの開発

台湾文献館から台湾総督府文書の画像データの提供を受け、それらを翻刻したのち、手書き文字の座標データ・字種データを作成する。翻刻は研究協力者である翻刻専門家および大学院生が行い、台湾総督府文書に精通する研究分担者が翻刻結果を検証・訂正する。個々の手書き文字の座標・字種データの作成作業の効率化を図るため、専用のソフトウェアを別途開発する。また、近代公文書における古語的語句･慣用表現を抽出･整理する。なお、このデータセットは下記（２）（３）において、深層学習の訓練データ、テストデータとして用いる。

（２）文字切り出し技術の開発

本研究の準備段階において経験則を用いた文字切り出し手法は精度 80%以下であることを確認した。本研究では精度を高めるため、深層学習による文字切り出し手法を検討する。

（３）個別文字認識技術の開発

深層学習を用いた個別文字認識手法を検討する。また、上記データセットでは字種ごとのデータ数の偏り出ることが予想されることから、データ増強法（人工的疑似サンプルを追加する方法）を用いてデータの偏りに対応することを検討する。

（４）自動解読システムの試作

上記（２）（３）の技術を用いた自動解読システムを試作し、解読精度の検証、新たな課題の抽出を行う。

（５）その他

研究期間中、半期ごとに研究組織会議を行い連携体制の維持を図る。研究成果は国内研究会および国際会議で公表する。

４．研究成果

（１）データセットの開発

台湾総督府文書 7 簿冊から画像 1,707 枚分の文書をサンプリングし、それらを翻刻した。さらに、画像中の個々の手書き文字の位置を求め、翻刻結果と対応づけることにより 363,519 個、約

況（いわんや）、抑（そもそも）、忝（かたじけなし）、偖（さて）、抔（など）、并ニ（ならびに）、迚モ（とても）、加之（しかのみならず）、都而（すべて）、決而（けっして）、乍去（さりながら）、乍恐（おそれながら）、有之（これあり）、依之（これにより）、斯様（かよう）、向後（きょうこう）、穴賢（あなかしこ）、幾許（いくばく）、今以（いまもって）事由（ことのよし）、事之外（ことのほか）、事候間（ことに・そうろう・あいだ）、悪敷者（あしきもの）、無覚束（おぼつかなし）、一件（いっけん）、如仰（おおせの・ごとし）、如件（くだんのごとし）、仍如件（よって・くだんのごとし）、如此（かくの・ごとし／ごとく）、如故（もとの・ごとし）、如何（いかん／いかに）、如何様（いかよう）、…

図 4 古語的語句・慣用表現リスト（一部のみ）

3500 字種の手書き文字の座標・字種データを有するデータセットを作成した。また、計 342 の語句・慣用表現を収集・整理した（図 4）。

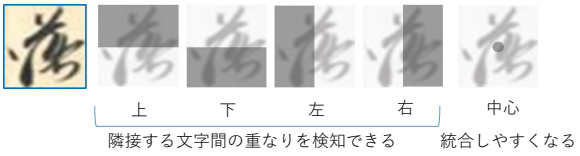

図 5 各文字の上下左右・中心を検出

（2）文字切り出し技術の開発

近代公文書の手書き文字は文字間隔や文字サイズが不均一であり、隣接する文字が接触したり、重なっている場合も多い。このような手書き文字を個別に検出するために、図 5 に示すように、個々の文字の上下左右の領域および中心を検出する手法を開発した。検出のための深層学習ネットワークには FCN(Fully Convolutional Networks)を用いた。この手法により適合率 97%、再現率 98%、精度 95%で個々の手書き文字を検出可能であることを実験により確認した。この成果は文書認識分野のトップカンファレンス ICDAR2019（採択率 53%）に採択された。

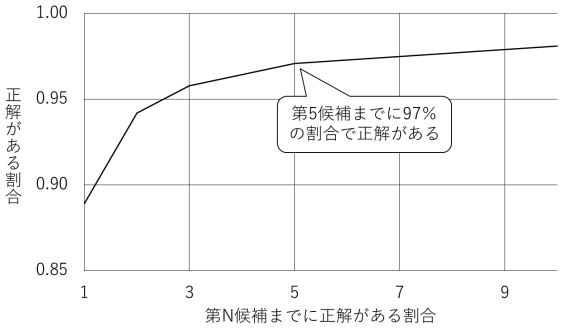

図 6 第 5 候補までに正解のある割合は 97%

（3）個別文字認識技術の開発

近代公文書に出現する文字は、字種ごとに出現頻度が異なるため、本研究で作成したデータセットにおいても、字種ごとのデータ数が不均一である。そこで本研究では、字種ごとのデータ数が均一となるようにデータ増強する手法を開発した。文字認識のための深層学習ネットワークには Google Net を用いて実験を行い、文字認識精度が 89%に達すること、および、第 5 候補までに正解がある割合は 97%であることを確認した（図 6）。この成果により画像認識分野の国際会議 IWAIT2019 で最優秀論文賞を受けた。

（4）自動解読システムの試作

上記の文字切り出し技術と個別文字認識技術を用いて、文書画像からその解読結果を出力する自動解読システムを試作した。自動解読の精度は 85%=95%×89%（第 5 候補まで含めると 92%）となる。図 7 には、台湾総督府文書の半ページ分に対する自動解読結果を示す。図中①、②は認識に失敗している例である。①は「年」を「南」と認識している。②は正解「都」が第 2 候補となり、「部」が第 1 候補になってしまっている。さらに精度を上げるためには、(a)多様な字形への対応と(b)字種は異なるが類似する字形への対応が必要である。(a)については、図 7①のように、同種の文字でも様々な崩し字や異体字があり、現時点のデータセットはそれらを網羅できていない。このためデータセット開発を継続し、サンプル数を増やす必要がある。(b)については、図 7②のように、字形が類似する手書き文字の判別を失敗する場合があり、これを防ぐために、文脈情報を利用して文字認識する仕組みが必要である。

（5）その他

上記の研究成果は学術論文 1 篇，国際会議発表 5 件、国内学会発表 9 件により公表した。

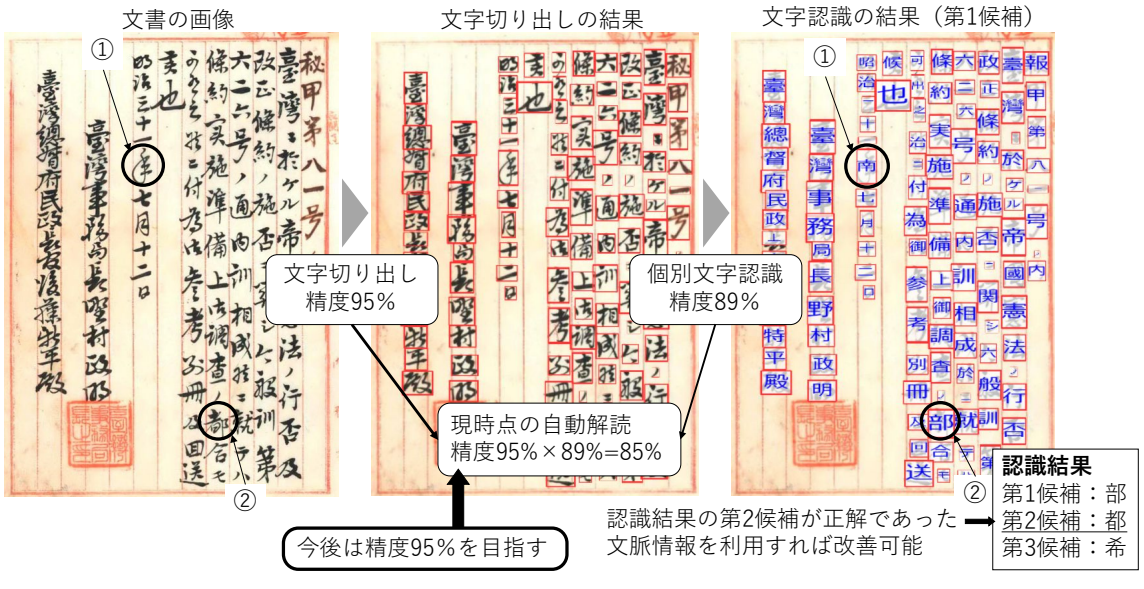

図 7 自動解読結果の例

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 山田雅之,目加田慶人,長谷川純一,鈴木哲造,東山京子,檜山幸夫,寺沢憲吾,川嶋稔夫	４．巻 38
２．論文標題 ディジ タル・ヒューマニティーズプロジェクト ―近代公文書自動解読のための基盤的研究―	５．発行年 2018年
３．雑誌名 社会科学研究	６．最初と最後の頁 1,23
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計14件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 Kei Watanabe, Shinji Takahashi, Yuki Kamaya, Masashi Yamada, Yoshito Mekada, Junichi Hasegawa, Shinya Miyazaki
２．発表標題 Japanese Character Segmentation for Historical Handwritten Official Documents Using Fully Convolutional Networks
３．学会等名 15th Int. Conference on Document Analysis and Recognition (ICDAR)（国際学会）
４．発表年 2019年

１．発表者名 釜谷勇輝，渡辺佳，高橋真治，山田雅之，目加田慶人，長谷川純一，中貴俊，宮崎慎也
２．発表標題 近代公文書自動解読システムのためのFCNによる手書き文字切り出し
３．学会等名 第10回社会情報学会中部支部/第5回芸術科学会中部支部合同研究会，SSICJ2019-1
４．発表年 2019年

１．発表者名 伊藤里華，山田雅之，目加田慶人，長谷川純一，中貴俊，宮崎慎也，鈴木哲造，東山京子，檜山幸夫
２．発表標題 近代公文書の文字認識に関する実験と考察
３．学会等名 第17回情報学ワークショップ
４．発表年 2019年

１．発表者名
高橋真治，山田雅之，目加田慶人，長谷川純一，中貴俊，宮崎慎也，鈴木哲造，東山京子，檜山幸夫

２．発表標題
文字領域検出と文字認識を用いた近代公文書翻刻支援システムの開発

３．学会等名
第17回情報学ワークショップ

４．発表年
2019年

１．発表者名
Kei Watanabe, Shinji Takahashi, Yuhei Takagi, Masashi Yamada, Yoshito Mekada, Junichi Hasegawa, Takatoshi Naka, Shinya Miyazaki

２．発表標題
Detection of Characters and their Boundary from Images of Modern Japanese Official Documents using Fully CNN-based Filter

３．学会等名
NICOGRAPH International 2018（国際学会）

４．発表年
2018年

１．発表者名
Shinji Takahashi, Kei Watanabe, Rika Itoh, Masashi Yamada, Junichi Hasegawa, Takatoshi Naka, Shinya Miyazaki, Kyoko Higashiyama, Yukio Hiyama

２．発表標題
Development of Handwritten Japanese Character Dataset for Auto-Transcription of Modern Japanese Official Documents

３．学会等名
NICOGRAPH International 2018（国際学会）

４．発表年
2018年

１．発表者名
目加田慶人

２．発表標題
人工知能と台湾総督府文書

３．学会等名
日台学術シンポジウム（国際学会）

４．発表年
2018年

１．発表者名
Zongjhe Yang, Keisuke Doman, Masashi Yamada, Yoshito Mekada

２．発表標題
Character recognition of modern japanese official documents using CNN for imblanced learning data

３．学会等名
2019 Int. Workshop on Advanced Image Technology（国際学会）

４．発表年
2019年

１．発表者名
小島雅大，道満恵介，山田雅之，目加田慶人

２．発表標題
手書き文字の傾きを考慮した切り出しによる古文書文字認識に関する検討

３．学会等名
第16回情報学ワークショップ

４．発表年
2018年

１．発表者名
伊藤里華，高橋真治，渡辺佳，山田雅之，目加田慶人，長谷川純一，中貴俊，宮崎慎也，鈴木哲造，東山京子，檜山幸夫

２．発表標題
近代公文書の手書き字形データセットの開発と個別文字領域検出手法の検討

３．学会等名
第16回情報学ワークショップ

４．発表年
2018年

１．発表者名
釜谷勇輝，山田雅之，目加田慶人，長谷川純一，檜山幸夫，東山京子，中貴俊，宮崎慎也，寺沢憲吾，川嶋稔夫

２．発表標題
近代公文書自動解読のための手書き字形データセット構築

３．学会等名
平成29年度電気・電子・情報関係学会東海支部連合大会

４．発表年
2017年

１．発表者名
楊宗哲，道満恵介，山田雅之，目加田慶人

２．発表標題
畳込みニューラルネットワークを用いた日本近代公文書字認識

３．学会等名
平成29年度電気・電子・情報関係学会東海支部連合大会

４．発表年
2017年

１．発表者名
高木裕平，山田雅之，目加田慶人，長谷川純一，中貴俊，宮崎慎也

２．発表標題
Fully-CNNを用いた近代公文書画像からの文字検出

３．学会等名
情報処理学会第80回全国大会

４．発表年
2018年

１．発表者名
羽田 竜馬，道満 恵介，山田 雅之，目加田 慶人

２．発表標題
畳み込みニューラルネットワークによる手書き漢字の部首認識

３．学会等名
2018年電子情報通信学会総合大会講演論文集

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
深層学習による文字・画像認識 日本近代公文書の解読支援で戦前期の行政を理解する
https://www.chukyo-u.ac.jp/research_2/news/2017/06/011775.html
第13回先端研究交流会　戦略的研究3テーマを報告
https://www.chukyo-u.ac.jp/research_2/news/2019/02/013523.html

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	目加田　慶人 (Mekada Yoshito) (00282377)	中京大学・工学部・教授 (33908)	
研究分担者	寺沢　憲吾 (Terasawa Kengo) (10435985)	公立はこだて未来大学・システム情報科学部・准教授 (20103)	
研究分担者	川嶋　稔夫 (Kawashima Toshio) (20152952)	公立はこだて未来大学・システム情報科学部・教授 (20103)	
研究分担者	長谷川　純一 (Hasegawa Junichi) (30126891)	中京大学・工学部・教授 (33908)	
研究分担者	檜山　幸夫 (Hiyama Yukio) (40148242)	中京大学・社会科学研究所・特任研究員 (33908)	
研究分担者	東山　京子 (Higashiyama Kyoko) (80570077)	中京大学・社会科学研究所・研究員 (33908)	
連携研究者	鈴木　哲造 (Suzuki Tetsuzo) (10771123)	中京大学・社会科学研究所・研究員 (33908)	
連携研究者	岩壁　義光 (Iwakabe Yoshimitsu) (30124506)	学習院大学・学習院大学資料館・大学非常勤講師 (32606)	