様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 6 月 20 日現在

機関番号：14602
研究種目：基盤研究(C)
研究期間：2009～2011
課題番号：21500237
研究課題名（和文）近代デジタルライブラリの自動テキスト化

研究課題名（英文）Auto Recognition of Kanji Characters for Early-Modern Japanese Printed Books

研究代表者　城　和貴（Joe Kazuki）
奈良女子大学大学院・人間文化研究科・教授
研究者番号：　90283928

研究成果の概要（和文）：国立国会図書館関西館の近代デジタルライブラリでは明治大正昭和初期の近代書籍約 30 万点が画像ベースで利用できる。本研究ではこの貴重なデジタルアーカイブを有効利用するための自動テキスト化を、既存の手書き文字認識技術を利用して行った。また誤認識した漢字を認識システムに迅速に知らせるための利用者用ポータルサイトも開発した。さらに裏抜け除去やルビ除去等、これまでに知られていなかった問題点にも解決の道筋を示した。

研究成果の概要（英文）：The national diet library of Japan provides the public with 300 thousands kinds of early-modern printed books as an internet service. In this research we develop an early-modern printed Kanji recognition system to make a full use of the digital archive. The recognition method is an extension of a recognition method for off-line handwritten Kanji characters. A portal site is developed for extra-learning the system with the help of users. Furthermore we faced several problems such as bleed through and rubi removal to find some solutions.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,500,000	450,000	1,950,000
2010 年度	900,000	270,000	1,170,000
2011 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：情報科学
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：デジタルアーカイブ、近代書籍、文字認識、裏抜け除去、文字切り出し

１．研究開始当初の背景

近代デジタルライブラリとは、国立国会図書館関西館電子図書館課で開発・運営を推進している、同図書館所蔵の明治期・大正期刊行図書を収録した画像データベースである。平成 20 年 10 月現在、約 143,000 冊（101,414 タイトル）を収録しており、Web ブラウザを通して Jpeg もしくは Jpeg2000 形式で一般に公開されている。これらは画像による閲覧に限られているため、テキスト検索が不可能であり、貴重図書に含まれる情報の有効利用という観点からは問題がある。一方、同時期の図書のテキストによるデジタルアーカイブとして、青空文庫(www.aozora.gr.jp)が知られている。青空文庫とは、著作権が消滅しパブリックドメインに帰した文学作品を収集・公開しているインターネット上の電子図書館で、テキスト化はボランティアによって

手作業で行われており、平成 20 年 10 月現在約 7,000 タイトルが公開されている。青空文庫はそのアーカイブの拡張はボランティアによってのみなされるため、近代デジタルライブラリのような網羅的なアーカイブ化は極めて困難である。これに対して近代デジタルライブラリのアーカイブ化は、図書をマイクロフィルムに保存した後、Jpeg ならびに Jpeg2000 でデジタル化する定型業務であるため、将来にわたってさらなるアーカイブ化が期待されている。

現在に至るまで OCR は完全ではないものの、実用化され図書の自動テキスト化に多く利用されているのは周知の事実である。近代デジタルライブラリの画像データに対し、現在の OCR を利用して自動テキスト化を行う場合、次のような問題点がある。

・時代・出版社によりフォントが多様である
・旧字体に対応する OCR は稀である
・ノイズが極めて多い

一方、近年のセマンティック Web の隆盛に見られるように、近代デジタルライブラリを含むデジタルアーカイブは、リソースとして知識処理の対象となるべきである。そのためには近代書籍を対象とした OCR 技術の確立が急務の課題である。

近代書籍の活字認識に関する研究はこれまでにほとんど報告されておらず、近代以前の手書きを含む古書の文字認識に関する研究が各時代ごとに見受けられる程度である。研究代表者は 20 年程前にオフライン手書き文字認識に関する研究に携わっていた。この研究の概略は以下の通りである。学習用ならびにテスト用データとして、それぞれ ETL9 から平仮名 71 種類もしくは JIS 第一水準中 256 種類の 200 人分データを選び、外郭方向寄与度特徴を特徴ベクトルとして計算し、それを入力として平仮名もしくは 256 種類の漢字の種類を出力とする、階層的なフィードフォワード型ニューラルネット（学習方法は改良型バックプロパゲーション）を構築した。この研究における知見は、このような機械学習による手書き文字認識では、簡単な文字（平仮名）の方が複雑な文字（JIS 第一水準の漢字のうち画数の多いもの）よりも認識が難しいことであった。これは簡単な文字（＝よく使われる文字）程、各個人の固有の筆跡に変動が大きいことと考えられた。

研究代表者は平成 19 年 12 月より国立国会図書館関西館に非常勤調査員として月 4 回勤務しているが、近代デジタルライブラリのテキスト化の可能性について同館電子図書館課のメンバーと議論を行った。同館所蔵の 143,000 冊に及ぶ蔵書を、青空文庫のように人手でテキスト化を行うのは予算的に不可能であり、既存の OCR 技術では全く役にたたないとの説明を受けたときに、研究代表者は過去の手書き文字認識研究から着想を得た。すなわち近代図書の多種多様なフォントや旧字体を含めると膨大な種類になる文字セットに対し、これを活字認識とは見なさずに、手書き文字と見なすという着想である。実際、近代書籍の多くの事例は、活字認識は困難であるが、手書き文字認識と見ると、それほど困難とは思えない。このような経緯と着想の結果、本研究課題を得ることとなった。

２．研究の目的

本研究課題には当初次の四つのサブテーマがあった。

1) 近代デジタルライブラリから出版社･刊行時期の異なる 256 種類の旧字体を含む文字を 200 セット手作業で切り出す
2) 1)で作成したデータに対して手書き文字認識手法を応用した近代図書活字認識技術を確立する
3) 近代図書画像データからの文字切り出し手法を確立する
4) 近代図書画像データと､それから抽出されたテキストの自動位置あわせ技術を確立する

1)と 2)は純粋に近代書籍の活字認識の手法を確立することが目的であり、3)は認識システムの自動構築を目的とした近代書籍の活字認識前処理であった。このような認識システムは一般に対象を完全に認識することは不可能であるので、4)で認識システムの学習補助 I/F を開発する。

以上の四つのサブテーマで研究に着手したが、実際に研究を行っていると新たな問題に直面したため、本研究課題では次の二つのサブテーマも研究の目的とした。

5) 裏抜けの除去手法の確立
6) 近代書籍用文字切り出し手法の確立

5)に関しては研究計画を立てた際には全く想定していなかった問題であった。近代書籍では、裏面の文字が表面でも観察されるノイズが発生する。この現象は裏面の文字が表面程鮮明ではなく輝度値が低い状態で発生する裏写りと呼ばれるものと、裏面の文字のインクが表面に完全に写りこむ裏抜けという現象に大別される。裏写りはヒストグラム変換等を利用することでほぼ完全にノイズ除去可能であり、現在のコピー機等では裏写り除去機能が実装されているものも多い。これに対して裏抜けはヒストグラム変換では理論的に除去不能であり、国内で裏抜けに関する研究がほとんど行われていないことが判明した。裏抜けに関する研究は欧州の中世の書籍、手紙、楽譜等に対するものが今世紀になってから多く報告されている。これは文化財のデジタルアーカイブを各方面で行うようになってから発覚した現象であると推測される。そして裏抜けの除去を完全に行って

いる研究はほとんどなく、現在でも裏抜け除去は基礎研究対象であるのが実情である。わが国では裏抜けに関する研究は、ノイズ除去という観点からは行われておらず、印刷技術や裏抜けしにくい印刷用紙の開発という方向で研究が行われてきた。

6)に関しては3)で既存の文字切り出し手法では、ルビの除去が困難であることが新たに発見された。特に大正期に入ってからの書籍のルビは、非常に複雑に配置されており、既存手法の射影を使ったものでは正しく文字を切り出せないことが判明した。

３．研究の方法

1)および 2) 近代デジタルライブラリから出版時期出版元の異なる２００タイトルを選出する。選出した各タイトルから、旧字体を含む 256 種類の活字を人手で切り取り、初期データとする。初期データに対し、ノイズ除去と正規化を行い、初期データと合わせてデータベース化を行う。同時にノイズ除去と正規化を当該データベースの初期データに対するクエリーとして実装する。データベースの各レコードに対して特長抽出を行うクエリーを実装する。この時の特徴抽出手法は、外郭方向寄与度特徴とする。特徴抽出後の出力データに対し、学習ならびにテスト・フェイズからなる認識器をクエリーとして実装する。認識手法には、サポートベクタマシン法（SVM）とする。以上のデータとクエリーを使って学習・テストを行った結果の性能を評価する統計ツールをクエリーとして実装する。

3)　近代書籍画像データから文字を切り出すツールを上記データベースにクエリーとして実装する。文字切り出しは近代書籍画像データから各ページを切り出し、アフィン変換を使って角度補正をおこない、垂直方向の射影を適用して各ページの累積輝度値ヒストグラムを作成し、文字列ごとの切り出しを行ってから単体の文字に分割する。

4)　近代書籍画像をテキスト化した文章データを閲覧者であるユーザが閲覧中に、誤認識文字を発見した場合、そのテキスト文章の位置に対応する元の画像データを表示し、テキスト修正画面を表示させ、誤りを訂正する機能を持つポータルサイトを開発する。対応する画像の表示にはメタデータを用いる。テキスト文章と画像データを照らし合わせたあとで、正しい文章を理解し、誤りを修正する機能もメタデータを利用して実装する。テキストデータには青空文庫のデータを用いる。

5)　本格的な裏抜けの研究は欧州を中心に最近数年の間に研究が進められている。主な手法として、シミュレーテッドアニーリングを使った方法、ウェーブレットを使った方法、マルコフランダムフィールドを使った方法、主成分分析＋クラスタリングを使った方法などが提案されている。これらは手紙や楽譜などが裏抜け除去対象であり、我々の対象である近代書籍とは条件が異なる。すなわち、裏抜けの発生する確率分布は出版社、出版年、書籍ごとに強く依存する。従って、ある書籍で裏抜けが観察されれば、その書籍では他の場所でも同様の裏抜けが発生している可能性が高く、その裏抜けの特徴も類似していると仮定する。この仮定のもと、本サブテーマでは遺伝的プログラミングを利用して、各書籍ごとに裏抜け除去を行うフィルターを自動生成し、そのフィルターを使って当該書籍の全ての裏抜けの除去を試みる。

6) 既存手法の 3)では特に大正期以降の書籍でルビの除去が失敗することが多いので、ルビと本文を単純に分離することは諦め、ルビを含むテキスト領域に異方性ガウシアンフィルタを適用し、二値化した後、領域抽出を施し、細長い連結成分を除去することでルビの除去を試みる。

４．研究成果

1)および 2)に関して、業績⑥では出現頻度が比較的高い漢字 10 種類の文字画像を使用して外郭方向寄与度特徴を計算し、得られた特徴ベクトルを SVM とニューラルネットワークによって学習させた後比較実験を行い，SVM による識別実験では 97:8%，ニューラルネットワークによる実験では精度 77:6% という結果を得た。また、業績④および③では、青空文庫でテキスト化されており、かつ近代デジタルライブラリで公開されているタイトルの中から、共通に使用されている漢字の種類ができるだけ多くなるように 9 作品を選び認識実験を行った。図 1 に 9 作品の「人」の違いを示す。共通に使用されている漢字は 260 種類まで確認できたので、16 種類から 256 種類まで 5 パターンで学習および認識実験を行ったところ、表 1 のように高認識率を達成した。

図 1 選ばれた 9 作品

表 1 SVM による実験結果

文字種	誤答数/テストデータ数	認識率 [%]	認識率分散
16 種類	13/640	97.969	2.95
32 種類	53/1280	95.859	3.79
64 種類	116/2560	95.469	2.63
128 種類	311/5120	93.926	0.82
256 種類	772/10240	92.461	0.02

3)に関して、業績④および③では、輝度値射影ヒストグラムによる文字切り出し手法を試したところ、図 2 のように平滑化のパラメータ設定が難しいことが判明した。そこで背景領域細線化による文字切り出し手法を使うことで図 3 のような特殊な場合以外は文字切り出しを行えるようになった。

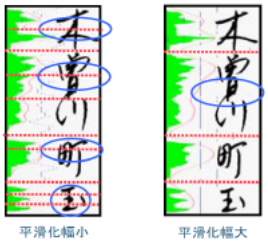

図 2 ヒストグラムによる文字切り出し

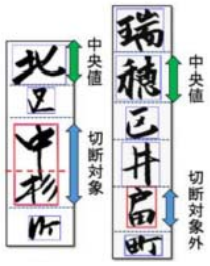

図 3 領域細線化による文字切り出し

また文字切り出しを行う前処理では、推定線幅に基づくノイズ除去を提案し、その結果として文字切り出しと文字認識の両方の性能を上げることが可能となった。線幅推定手段では，まず連結成分ごとにラベリング処理を行う．ラベリング処理をした各連結黒画素成分の面積を算出し，最大の連結黒点群を A とする．次に連結黒点群 A 内の点 $p_n(p_n \in A)$ を通る最短の黒点連結直線の長さ lpn を求める．文字線幅の走査の様子を図 4 に示す．各点から走査を行い，l_{pn} の中央値を文字の推定線幅とする．この文字の推定線幅を用いると，連結黒画素成分の面積＜(推定線幅£ 推定線幅)/2 となる連結成分をノイズとして削除する。推定線幅に基づくノイズ除去の例を図 5 に示す。

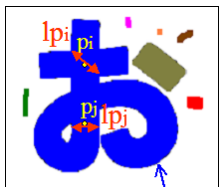

図 4 文字線幅の走査

図 5 推定線幅に基づくノイズ除去の例

4)に関しては業績⑤で近代デジタルライブラリのテキスト化支援ポータルサイトを開発した。図 6 は，ポータルサイトのシステムの全体図である．ポータルサイトは，GUI とデータベースの 2 つの部分にわかれており，GUI での入力に対して．3 つのデータベースのうち必要なデータベースにアクセスすることによって処理を行う．GUI の開発環境は，サーバに無料で利用できる Apache Tomcat6.0 を用いる．次にデータベースについて述べる．ポータルサイトに含まれているデータベースは大きくわけて 3 つある．画像データを格納した画像データベース，テキスト文章を格納したテキストデータベース，管理情報を格納した管理データベースである．全員が利用可能な書籍検索は管理データのみを検索の際に使用する．また，テキスト表示では，テキストデータと管理データの 2 つを利用する．管理者と id 所有ユーザが行えるログインには管理データのみを使用する．画像表示では，テキストデータ，管理データ，画像データ，3 つすべてを利用する．同様に，テキスト修正でも，テキストデータ，管理データ，画像データ，3 つすべてを利用する．

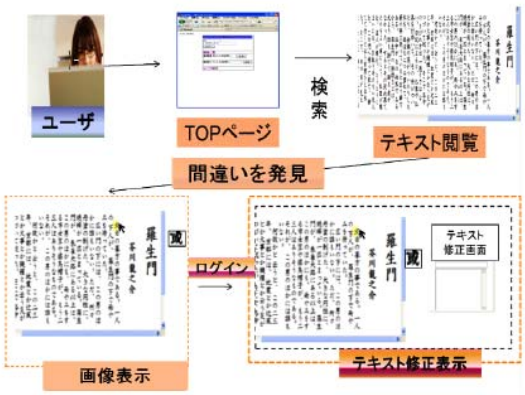

図 6 ポータルサイトの概要

5)に関しては業績②で遺伝的プログラミングを用いた裏抜けの検討を行った。四則演算や基本関数等のプリミティブを与えた状態で、個体数 1000、交差率 0.8、突然変異率 0.1 で図 7 のような学習セットを 20 種類与えて学習させたところ、fabs(omoteimg[i][j]+((sqrt(abs((omoteimg[i][j+1]*(Min(uraimg[i][j])*sqrt(abs(sqrt(omoteimg[i][j])+((omoteimg[i][j]*Min(uraimg[i][j]))*fabs(omoteimg[i][j]+omoteimg[i][j])))))))+(omoteimg[i][j]*sqrt(abs(sqrt(omoteimg[i][j])+((omoteimg[i][j+1]*omoteimg[i][j+1])*fabs(fabs(omoteimg[i][j+1]+1)+(fabs(omoteimg[i][j+1]+Min(uraimg[i][j]))*Avrg(omoteimg[i][j]))))))))*Min(omoteimg[i][j]))*Min(ura img[i][j]))) というフィルターを得た。このフィルタを用いた裏抜け除去例を図 8 に示す。

図 7 裏抜け表画像（左）裏抜け裏画像（中）教師画像（右）

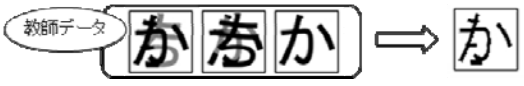

図 8 裏抜け除去例

6)に関しては業績①でルビを含むテキスト領域に異方性ガウシアンフィルタを適用し、二値化した後、領域抽出を施し、細長い連結成分を除去することでルビの除去を行った。その効果を検証するために、近代デジタルライブラリの中から画像数 200 枚の「田舎教師」に対して、3)の手法との比較を行ったところ、ルビ除去率は二倍に向上した。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計６件）
①福尾真実，高田雅美，城和貴、近代デジタルライブラリーの文字切り出しにおける実際的手法、情報処理学会数理モデル化と問題解決研究会研究報告、査読なし、MPS-87-31、2012、http://www.bookpark.ne.jp/cm/ipsj/search.asp?flag=6&keyword=IPSJ-MPS12087034&mode=PDF
②栗津妙華，高田雅美，城和貴、進化計算による裏抜け除去、進化計算学会進化計算シンポジウム 2011 予稿集、査読あり、PDF を現地 CD 配布、2011
③M. Fukuo, Y. Enomoto, N. Yoshii, M. Takata, T. Kimesawa, K. Joe, Evaluation of the SVM Based Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books, Proceedings of the 2011 International Conference on Parallel and Distributed Processing Techniques and Applications, 査読あり、Vol.2, 2011, p.727-732
④榎本友理枝，高田雅美，木目沢司，城和貴、SVMに基づく多フォント漢字認識手法の評価、情報処理学会数理モデル化と問題解決研究会研究報告、査読なし、MPS-82-14、2011、http://www.bookpark.ne.jp/cm/ipsj/search.asp?flag=6&keyword=IPSJ-MPS11082014&mode=PDF
⑤黒田佳世，榎本友里枝、高田雅美、城和貴、近代デジタルライブラリーテキスト化支援のためのポータルサイトの設計、情報処理学会数理モデル化と問題解決研究会研究報告、査読なし、MPS-81-35、2010、http://www.bookpark.ne.jp/cm/ipsj/search.asp?flag=6&keyword=IPSJ-MPS10081035&mode=PDF
⑥C.Ishikawa, N.Ashida, Y.Enomoto, M.Takata, T.Kimesawa, K.Joe、Recognition of Multi-Fonts Character in Early-Modern Printed Books、Proceedings of the 2009 International Conference on Parallel and Distributed Processing Techniques and Applications、査読あり、Vol.2、2010、p. 728-734

〔学会発表〕（計６件）
①福尾真実，高田雅美，城和貴、近代デジタルライブラリーの文字切り出しにおける実際的手法、情報処理学会第 87 回数理モデル化と問題解決研究会、2012 年 3 月 2 日、鹿児島県指宿市
②栗津妙華，高田雅美，城和貴、進化計算による裏抜け除去、進化計算学会第 4 回進化計算シンポジウム、2011 年 12 月 18 日、宮城県岩沼市
③M. Fukuo, Y. Enomoto, N. Yoshii, M. Takata, T. Kimesawa, K. Joe, Evaluation of the SVM Based Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books, The 2011 International Conference on Parallel and Distributed Processing Techniques and Applications, 2011 年 7 月 18 日、米国ラスベガス市
④榎本友理枝，高田雅美，木目沢司，城和貴、SVM に基づく多フォント漢字認識手法の評価、情報処理学会第 82 回数理モデル化と問題解決研究会、2011 年 3 月 7 日、宮崎県宮崎市
⑤黒田佳世，榎本友里枝、高田雅美、城和貴、近代デジタルライブラリーテキスト化支援のためのポータルサイトの設計、情報処理学会第 81 回数理モデル化と問題解決研究会、2010 年 12 月 17 日、福岡県福岡市
⑥C.Ishikawa, N.Ashida, Y.Enomoto, M.Takata, T.Kimesawa, K.Joe、Recognition of Multi-Fonts Character in Early-Modern Printed Books、The 2009 International Conference on Parallel and Distributed Processing Techniques and Applications, 2009 年 7 月 13 日、米国ラスベガス市

６．研究組織
(1)研究代表者
城　和貴（Joe Kazuki）
奈良女子大学大学院・人間文化研究科・教授
研究者番号：90283928

(2)研究分担者
高田　雅美（Takata Masami）
奈良女子大学大学院・人間文化研究科・助教
研究者番号：20397574