Vol.2015-MPS-104 No.9
2015/7/27

近代書籍に特化した認識手法のための効率的な学習データ収集法

上坂和美[†1]　粟津妙華[†1]　石川由羽[†1]　高田雅美[†1]　城和貴[†1]

本稿では, 近代デジタルライブラリーの自動テキスト化に向けた学習データの収集を効率的に行うための Web アプリケーションの開発を行う．国立国会図書館では，多数の近代書籍を一般公開しているが，全文検索ができない．歴史的に貴重な文学で高度な検索を行うには画像データからのテキスト化が求められている．テキスト化のために PDC 特徴と SVM を用いた近代書籍に特化した多フォント活字認識手法が提案されている．この手法の性能を向上させるためには，大量の学習データが必要である．しかしながら手動で収集することは大変非効率である．そこで本研究では，データ収集支援を行う Web アプリケーションを提案する．

An Effective and Interactive Training Data Collection Method for Early-Modern Japanese Printed Character Recognition

KAZUMI KOSAKA[†1]　TAEKA AWAZU[†1]　YU ISHIKAWA[†1]
MASAMI TAKATA[†1]　KAZUKI JOE[†1]

In this paper, we present a web application that supports to collect training data efficiently for early-modern Japanese printed character recognition. The national diet library in Japan provides a lot of early-modern (AD1868-1945) Japanese printed books to the public, but full-text search is essentially impossible. In order to perform advanced search in historical literatures, it is required extracting texts from images. To solve this problem, we have already proposed a multi-font Kanji character recognition method using the PDC feature and an SVM. For growing in performance of this method, we need big amounts of training data. However, collecting training data by hand is extremely inefficient. Therefore, we propose a Web application that supports collecting training data.

1. はじめに

国立国会図書館[1]は，明治期から昭和初期にかけて刊行された近代書籍をおよそ 3500 万点所蔵している．平成 14 年から近代デジタルライブラリーとして，近代書籍の画像データを Web 上で提供している[2]．ただし，公開されている近代書籍は画像データとしてアーカイブ化されているため，全文検索することができない．よって，より近代書籍を利用しやすくするために本文のテキスト化が求められている．だが，画像データに既存の OCR を適用しても認識率が低く実用に耐え得るものではない．そこで我々は近代書籍に特化した多フォント活字認識手法といわれる手法を提案している [3][4][5]．この手法は手書き文字認識の手法を利用し，有効な識別手法の一種である SVM を特徴ベクトルの識別に用いる．近代書籍は，出版者や出版時期によって文字の形状が異なる．これらを網羅的に認識させるためには, SVM で必要となる学習データの数を大きくしなければならない．しかし，手動で様々な文字を学習データとして収集するのは人間にとって負担がかかり過ぎる．それゆえ効率的に収集を行うために，近代書籍用 OCR の学習データ収集支援 Web アプリケーションを提案する．本論文の構成は，以下の通りである．第 2 章では近代書籍に特化した多フォント活字認識手法について詳しく述べる．第 3 章では学習データの収集を支援する Web アプリケーションを提案する．第 4 章では，Web アプリケーションを用いた際の書籍の認識結果と手動での書籍の認識結果とを比較し，アプリケーションの有用性に関して考察する．

2. 多フォント活字認識手法

近代書籍には多フォント活字認識手法を使用している[3][4][5]．多フォント活字認識手法のために必要な作業工程は，前処理，文字切り出し，PDC 特徴(Peripheral Direction Contributivity,PDC)抽出[6]，SVM[7]を用いた特徴ベクトルの識別である．順に詳しく説明していく．

2.1 前処理

前処理としては画像の 2 値化・ノイズ除去・角度補正・ルビ除去を行う．

必要最小限の画像情報とするために, まず 2 値化を行う．

書籍画像には，近代書籍が活版印刷であるためノイズが含まれていることがほとんどである．画像上のノイズが文字線として誤認識されることを防ぐためノイズの除去を行う．また，ページのたわみのずれを解消するために，ページの歪んだ分だけ角度補正を行う．

ルビ文字はインクが滲んだ状態で親文字とつながって

†1 奈良女子大学
Nara Women's University

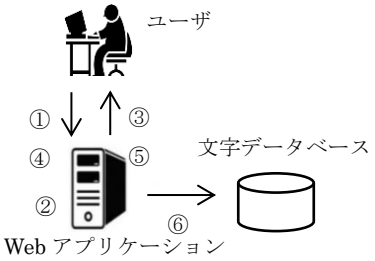

図 1　近代書籍用 OCR の学習データ収集支援 Web アプリケーションの概略図

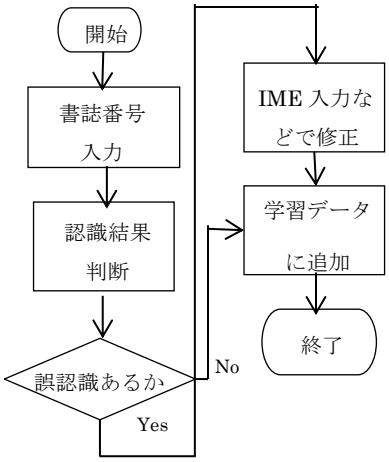

図 2　ユーザインターフェースのフローチャート

おり近代書籍の認識を妨げやすい．したがって次にルビ除去を行う．ルビ除去には漢字の横と縦の間の比率をとる手法[5]を用いる．行ごとにルビを含んだ漢字の横幅と縦幅の平均を計算する．その平均値に応じて遺伝的プログラミングによって生成された式でルビ除去を行う．

2.2　文字切り出し

2.1 節による前処理を行った書籍画像から，文字を 1 文字ずつ切り取る．この際，文字の外接矩形を用いて画像を縦・横・斜めの 8 方向に連結した黒画素部分にラベリング処理を行い，外接矩形を求める．

2.3　PDC 特徴

PDC 特徴[6]とは，丁寧に書かれた楷書体手書き漢字の認識に効果的な特徴の 1 つである．PDC 特徴は文字線の複雑さ，文字線の方向，文字線の接続関係，文字線の相対位置関係の 4 種類により文字線の構造情報を表している．

文字線内の黒点Pの方向寄与度d_Pをd_P= $(d_{1P}, d_{2P}, d_{3P}, d_{4P})$で表す．各要素$d_{mP}$(m=1,2,3,4) は，点 P から縦・横・斜めの 8 方向に触手を伸ばして求まる黒点連結長l_j (j= 1,2,…,8) を用いて，d_{mP}：

$$d_{mP} = \frac{l_m + l_{m+4}}{\sqrt{\sum_{j=1}^{4}(l_j + l_{j+4})^2}} \qquad (1)$$

で定義される．

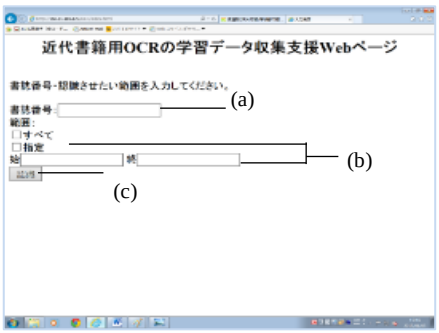

図 3　Web アプリケーション初期画面

2.4　SVM を用いた識別

最後に，Vapnik らによって考案された機械学習の一種である SVM[7]を PDC 特徴ベクトルに適用する．SVM は，計算量が比較的少なく単純な原理しか持たないにも関わらず，様々な分野で良好な結果を示している．

3.　近代書籍用 OCR の学習データ収集支援 Web アプリケーション

多フォント活字認識手法の認識精度を向上させるには，SVM で使用する学習データを効率良く収集することが必要である．そこで，学習データの収集を支援する Web アプリケーションを開発する．一般ユーザにとって操作が簡単な Web アプリケーションでなければならない．

3.1　Web アプリケーションの構成

図 1 は，開発する Web アプリケーションの概略図である．図 1 の番号はアプリケーションの動作・操作手順を表す．図 2 は Web アプリケーションを操作するユーザインターフェースのフローチャートである．図 1 で示す概略図でのユーザの動き①④をフローチャートにして表している．図 1①では，ユーザは図 3(a)の箇所に全国書誌番号を入力し図 3(b)の箇所で認識させたい範囲を選択する．全国書誌番号は国立国会図書館が納本制度に基づいて割り振った番号で，全国書誌に掲載されている．ユーザは図 3(c)より認識させる．図 1②では，第 2 章で示された多フォント活字認識手法を用いて，認識を行う．図 1③では，原画像と認識結果を並べて表示出力する．ユーザは，認識結果を原画像と参照することで容易に誤認識の有無を確認することができる．図 1④では，誤認識があればユーザが正しい文字に修正を行う．IME パッドを用いて入力し訂正することもできる．図 1⑤では，認識された範囲の文字を SVM で用いる学習データの文字データベースに追加する．

3.2　外部設計・内部設計

Web アプリケーションの内部設計を図 4 に示す．サーブレット形式で開発するため，処理速度が早く，ユーザの負担を軽くすることができる．使用しているサーブレットコンテナは Apache Tomcat 7，Web サーバソフトウェアは，Apache HTTP Server 2.0 とする．この連携により，動的なコンテンツ処理は Apache Tomcat で処理し，静的コンテンツ

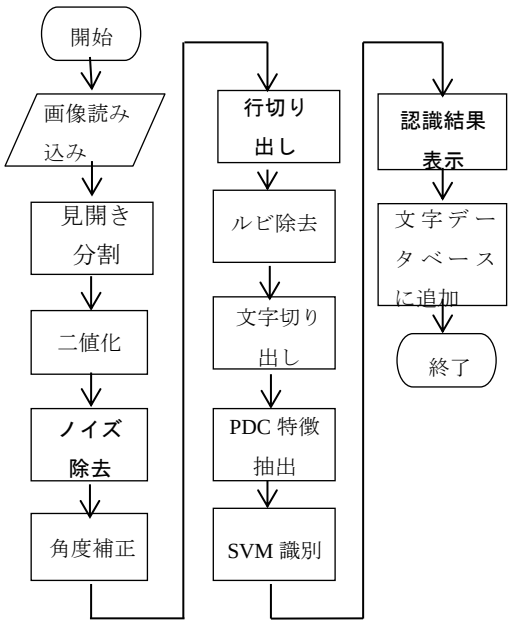

図 4　内部設計

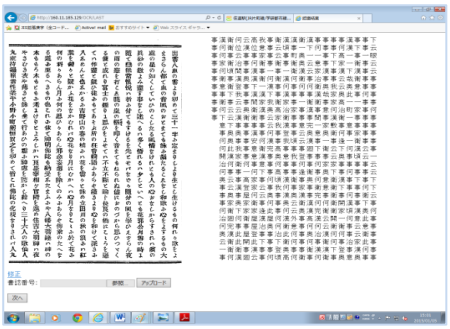

図 5　認識結果表示画面

は Apache で処理するといったように分担で処理を行うことできる．

まず Web アプリケーションでは見開きページごとに読み込み，画像を半ページずつ分割する．次に，半ページの画像を 1 行ずつに切り出し，ルビ除去の処理を行う．ルビ除去後，一文字ずつ切り出し，PDC 特徴抽出の処理に渡す．このアプリケーションでは SVM 識別に，SVM ライブラリの 1 つである LIBSVM2.81 を用いる．構築する文字データベースは，番号，文字，PDC 特徴の 3 項目で構成する．ここで用いる番号とは，第 1 水準から第 4 水準までの漢字・ひらがな・カタカナを 1 から順に割り当てた数字になっている．定義された以外の文字の場合，番号は 0 とする．文字データベースは， MySQL5.5.38 を用いる．SVM を適用した結果は，文字データベース構築の際に割り振った番号で表されている．したがって，各々番号に応じた文字に変換し，認識結果とする．図 5 は認識結果の表示画面である．

修正を行う際，ユーザは図 5(a)をクリックして認識結果のファイルをサーバから直接ダウンロードする．ダウンロードしたファイルを開き，自ら修正する．読み方などが不明な文字は IME パッドを用いる．IME パッド使用によりユーザ自ら不明な文字を検索する必要がないため，ユーザの負担は少なくなると考える．修正し終えたファイルはサーバへアップロードすることで修正完了とする．これらの動作はすべて Web 上のボタンで行う．認識したページ分をすべて表示し終えると最後に学習データへ追加するボタンのみの表示になる．Web アプリケーションはアップロードされたファイルより学習データが追加されデータベースは更新される．データベースの更新が終えると Web アプリケーションは操作完了の画面表示になる．

表 1　処理時間

処理 (時間)(時:h 分:m 秒:s)	Web アプリケーション	手動
前処理	5.22s	15h27m
PDC 特徴抽出	1.76m	7m
SVM 識別	8.42m	61.2m

表 2　Web アプリケーションの一致率

比較対象	個数	不一致数	一致率
(a)	1140	191	83.2%
(b)	934	69	92.8%
(c)	304	8	97.5%
(d)	630	61	90.4%
(e)	189	105	35.6%
(f)	17	17	0%

4. 実験

第 3 章で開発した Web アプリケーションに関して，手動で学習データを集めた場合と比較することによって，有用性を確認する．比較のために，明治 16 年に出版された書籍を用いる．1 ページの文字数は，1140 字である．サーバには CPU が Intel(R) Core(TM) i7-4770K CPU @ 3.50GHz，メモリは 16GB の計算機を用いる．

4.1 比較結果

Web アプリケーションの使用時と手動時での処理時間は表 1 に示す．Web アプリケーションでは，画像の前処理では約 9.4×10^{-5} 倍，PDC 特徴の抽出に関しては約 2.5×10^{-1} 倍，SVM 識別では約 1.4×10^{-1} 倍の処理時間で行えることが分かる．この結果により，手作業と比較すると Web アプリケーションを使うことで，同じ数だけ学習データの収集を行う作業が，肉体的にも負担が少なく，時間的にも大幅に短縮できるため，大変効率的であると考えられる．手動でかかったおよそ 16 時間 35 分と同じ時間 Web アプリケーションを用いると，およそ 96.9 倍の，約 110466 個の文字画像を学習データとして収集することができる．

次に，Web アプリケーションの使用時の認識結果と手動時での認識結果を比較し，一致率を調べる．一致率とは，それぞれの方法で得られた認識結果がどの程度一致しているかを表した割合であり，認識率とは異なる．比較結果を

表 2 で示す．比較対象は，以下のように分けて行う．

(a) 全ての文字を対象
(b) 文字切り出しの結果，欠けていないもののみを対象
(c) (b)での対象のうち，漢字のみを対象
(d) (b)での対象のうち，ひらがな・カタカナのみを対象
(e) 文字切り出しの結果欠けているものを対象
(f) 文字切り出しの結果一文字と判断さえず分解しているものを対象

(a)~(d)の結果より，(c)の欠けていない漢字のみを対象とした場合が最も一致率が高いことがわかる．ひらがな・カタカナは曲線や文字線の構造が単純である．一般的に，文字構造が単純であるほど抽出される PDC 特徴はより簡潔になる．ゆえに，漢字は認識率が最も高く結果が一致する可能性が高いと考えられる．したがって漢字よりもひらがなとカタカナの学習データの数を増やす必要がある．今後 Web アプリケーションを繰り返し用いることで，漢字だけでなくひらがな・カタカナも学習データとして随時追加されていくことが期待される．その結果，認識率も上昇し，手動の場合との一致率も向上するのではないかと考えられる．したがって，手動の場合と Web アプリケーションを用いた場合の処理にかかる時間および一致率を考慮すると，このアプリケーションは十分有用であるといえる．

4.2 Web アプリケーション使用時と手動時との比較

Web アプリケーションの使用を重ねた際の認識率と文字データベースの文字数・文字種数の推移を確認する．使用当初時，文字データベースに含まれる文字種は 1,000 種，総文字数 4,345 とする．図 6 は，使用する毎の文字データベースに含まれる文字種の数と認識率の結果の推移を表している．図 6 より，文字種の数はおよそ 1.3 倍，認識率はおよそ 3 倍に上昇している．次に Web アプリケーションを用いた場合と，手動で行ったとした場合の作業時間を比較する．Web アプリケーションを用いて図 6 で表す学習データ数を収集する総作業時間は，およそ 10.5 時間となる．同様に同じ学習データ数の収集を手作業で行った場合，およそ 42.9 時間になると考えられる．したがって，学習データの収集作業は，Web プリケーションを用いることで手作業でのおよそ $\frac{1}{4}$ の時間で行えることがわかる．

5. まとめ

本稿では，近代書籍に特化した多フォント活字認識手法で用いられる SVM で必要な学習データを効率よく収集するアプリケーションの開発について述べた．この Web アプリケーションを用いた場合と，手動で行った際の認識結果・処理時間を比較した．全体の処理時間は，Web アプリケーションを用いた場合およそ 10.3 分，手動ではおよそ 16 時間 35 分となった．また，Web アプリケーションを用いた認識結果と手動での認識結果の漢字のみの場合の一致率は約 97.5%となった．そこで Web アプリケーションの使用を重ねた際の認識率と文字データベースの文字数・文字種数の推移を確認した．認識率は使用当初と比較しておよそ 3 倍となる．また，文字データベースの文字種数はおよそ 1.3 倍になる．手作業で行ったとする時間と比べて Web アプリケーションは処理時間・一致率，また使用した際の認識率・文字データベースの文字種と文字数の推移より十分有用であるといえる．

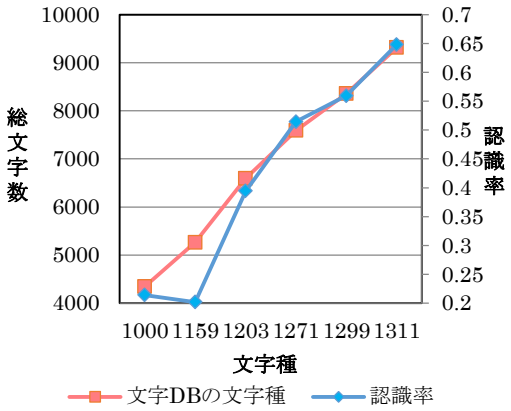

図 6 文字データベースの文字種数と認識率

謝辞 本研究は科研費・新学術領域研究(No26280119).の助成を受けたものである．また，プログラム記述に関してご協力いただいた奈良女子大学生である岡村沙季さん，高北奈津子さん，柿原玲奈さん，兼松明未さん，冨澤皐月さんに感謝致します．ありがとうございました．

参考文献

1) 国立国会図書館
http://www.ndl.go.jp/
2) 近代デジタルライブラリー
http://kindai.ndl.go.jp/
3) Ishikawa,C.，Ashida,N.，Enomoto,Y.，Takata,M.，Kimesawa,T.，and Joe,K.：Recognition of Multi-Fonts Character in Early-Modern Printed Books，Proceedings of International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA’09)，Vol. Ⅱ，pp．728-734(2009).
4) Fukuo,M.，Enomoto,Y.，Yoshii,N.，Takata,M.，Kimesawa,T. and Joe,K.：Evalua-Tion 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 Technologies and Applications (PDPTA2011)，Vol. Ⅱ，pp．727-732(2011).
5) Awazu,T.，Fukuo,M，Takata,M and Joe,K.：A Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books with Ruby Characters, International Conference on Pattern Recognition Applications and Methods (ICPRAM 2014), 637-645 (2014.3)
6) 萩田博紀，内藤誠一郎，増田功.：外郭方向寄与度特長による手書き漢字の認識，電子通信学会論文誌．(D)，Vol.J66-D，No.10，pp．1185-1192(1983).
7) Cristianini，N．andShawe-Taylor，J.：Support vector machine Introduction，Kyoritsu Publisher(2005).