Vol.2011-MPS-82 No.14
2011/3/7

# SVMに基づく多フォント漢字認識手法の評価

榎本　友理枝[†1]　高田　雅美[†1]
木目沢　司[†2]　城　和貴[†1]

国立国会図書館では，所蔵する明治から大正期にかけての近代書籍を画像データとしてアーカイブ化し，Web 上で一般に公開している．このデジタルアーカイブをより簡便に利用できるよう，近代書籍画像の早急なテキスト化が望まれている．本稿では，SVM に基づく近代書籍に特化した多フォント漢字認識手法の有効性を実証する．出版社が異なる書籍から切り出した様々なフォントの漢字 256 種を用いて識別実験を行った結果，常に 92% 以上の識別率を得ることができた．従って，文字画像に対して PDC 特徴を抽出し，SVM で学習・識別を行うという提案手法が近代書籍で使用されている多フォント漢字認識に対して有効な手法であるといえる．

**Evaluation of the SVM based Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books**

Yurie Enomoto,[†1] Masami Takata,[†1]
Tsukasa Kimesawa[†2] and Kazuki Joe[†1]

The national diet library in Japan provides a web based digital archive for early-modern printed books by image. To make better use of the digital archive, the book images should be converted to text data. In this paper, we evaluate the SVM based multi-fonts Kanji character recognition method for early-modern Japanese printed books. Using several sets of Kanji characters clipped from different publishers' books, we obtain the recognition rate of more than 92% for 256 kinds of Kanji characters. It proves our recognition method, which uses the PDC feature of given Kanji character images for learning and recognizing with a SVM, is effective for the recognition of multi-fonts Kanji character for early-modern Japanese printed books.

## 1. はじめに

国立国会図書館[1] では明治期から大正期にかけての書籍約 39 万冊を所蔵している．これらの近代書籍は哲学，歴史，自然科学，文学等の幅広い分野にわたり，また現在は絶版になっているものも多く，学術的に非常に貴重な資料である．通常，図書館での書籍の公開を考える場合，経年劣化や人の手による破損・紛失の危険を無視することができず，希少価値のある書籍を一般公開することは難しい．この問題を解決し一般向けに書籍を公開するために，近代デジタルライブラリー[2),3)] というプロジェクトが開始されている．これは，近代書籍をページごとにマイクロフィッシュ化し，それをフィルムにスキャンした書籍画像を Web 上で公開するものである．デジタルデータであるため，貴重な書籍の破損・紛失の恐れもなく，インターネット経由で利用者はいつでも書籍を閲覧することができる．現在，近代デジタルライブラリーでは約 17 万冊の書籍閲覧が可能である．

近代デジタルライブラリーの Web サイトでは，タイトル・著者名の他，出版年や日本十進分類（NDC）等詳細な項目を設定して書籍の検索を行うことができる．ただし，一部の書籍の目次部分はテキストデータ化されているが，本文は未だテキストデータ化されておらず，本文に含まれる文字列の検索など，テキストデータを扱う機能は存在しない．そのため，近代書籍データのより簡便な利用のために早急なテキスト化が望まれているが，近代デジタルライブラリーで公開されている書籍は前述の通り膨大であり，手作業によるテキスト化は効率的ではない．

一般的な文書画像であれば，光学文字認識（OCR）ソフトウェアによって文書画像からテキストデータへの変換を自動的に行うことができるが，近代書籍の多くは旧字体・異字体を多く含み，各出版社や出版年代により異なる種類の活字が使用されているため，市販の OCR ソフトウェアの適用が不可能である．

この問題を解決するために，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[4]．しかし，4) で行われた識別実験は出現頻度が比較的高い漢字 10 種類のみを使用したため，提案手法が有効であるとはいい難い．そこで本稿では，文字の種類を拡大した識別実験を行い，4) で提案された手法の有効性を実証する．文字種を拡大すると，識別実験に必

†1 奈良女子大学大学院人間文化研究科
Graduate School of Humanities and Sciences, Nara Women's University
†2 国立国会図書館
National Diet Library

要な文字画像数が膨大となり，文字画像を全て手動で切り出すことが困難となる．したがって本稿では，書籍画像に対して自動文字切り出しを行う．そして，近代書籍に使用されているフォントの多様性を吸収するため，手書き文字認識で利用される外郭方向寄与度（PDC）特徴[5] を用いて文字の特徴ベクトル化を行う．この特徴ベクトルに対して，機械学習の一手法であるサポートベクターマシン（SVM）を用いて特徴ベクトルの学習・分類を行う．

本稿の構成は次のようになる．2 章において，近代書籍に特化した多フォント活字認識手法について，3 章において，自動文字切り出し手法について述べる．4 章において評価実験を行い，実験結果から考察を述べる．最後に本稿のまとめを行う．

## 2. 近代書籍に特化した多フォント活字認識手法

4) において提案された活字認識手法の流れを次に示す．

(1) 一文字ごとに分割された文字画像（入力パターン）の読み込み．
(2) 与えられた文字画像に対し，2 値化・サイズの正規化・平滑化・ノイズ除去の処理．
(3) 文字の PDC 特徴を抽出．まず手順（2）にて前処理を施した文字画像を周囲 8 方向から走査し，文字線の構造を取得．次に文字線の黒点連結長から PDC 特徴を算出．最後に文字種ごとに 1 つのクラスとし，ラベリング処理．
(4) SVM を用いた特徴ベクトルの学習．学習結果が未知データを入力した際に識別に使用する辞書．

手順（1）で文字画像を読み込んだ後，手順（2）で画像に対し前処理を行う．PDC 特徴を計算する際，画像を 2 値化する必要がある．また，サイズ正規化によって文字画像の余白を取り除き，異なる文献から切り取られた文字画像の大きさを統一する．さらに，近代書籍は印刷品質が劣悪である場合がほとんどなので，ノイズ除去を行う必要がある．ノイズ除去によって，画像上のノイズが文字線と誤認識され PDC 特徴の計算に影響を与えることを防ぐことができる．

さらに手順（3）で PDC 特徴を計算し，手順（4）で SVM の学習を行う．この際，クラスを表すラベルと手順（3）で求めた特徴ベクトルの各要素を並べたものの対が，SVM の学習・分類の対象であるパターンとなる．全ての文字画像をパターンに変換する．各クラスに属するパターンのうち半数を無作為に選び，これを教師データとし残りをテストデータとする．教師データのみを用いて SVM の学習を行う．学習済み SVM は未知のテストデータが入力されると，学習によって得た分離超平面からテストデータが属するクラスを予想することができる．

10 種類の漢字を使用して SVM とニューラルネットワークによる比較実験を行い，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% という結果を得ている[4]．したがって，ニューラルネットワークと比較し，SVM での学習が PDC 特徴ベクトルの学習・分類に適していると考えられる．しかし，日本の文字コードとして日本規格協会が選定している漢字の種類は第一水準と第二水準だけでも 6,349 種類存在する[6]．また漢字の構造は階層的になっており，部首など類似した構造も多く存在する．

そこで本稿では，漢字の種類を 256 種に拡大した識別実験を行う．4) では，識別実験に使用する文字画像を全て手動で切り出していた．しかし文字の種類を拡大すると，識別実験に必要な文字画像数が膨大となり，手動での文字切り出しが困難となるため，本稿では書籍画像に対して自動文字切り出しも行う．

## 3. 自動文字切り出し

文字切り出しの流れを次に示す．

（A） 書籍画像 1 ページ（入力パターン）の読み込み．右ページ・左ページに分割．
（B） 分割した半ページの書籍画像に対し，2 値化・ノイズ除去・アフィン変換の処理．
（C） レイアウト解析．
（D） 手順（C）にて行ったレイアウト解析の範囲内で行の切り出し．
（E） 切り出した行ごとにラベリングの処理．付けたラベルが文字領域であると判断し，文字の切り出し．

近代デジタルライブラリーで公開されている書籍データは，書籍をページごとにマイクロフィッシュ化し，それをフィルムにスキャンした画像データである．ページをスキャンする際，ページの境目に向かって傾きが生じるため，左右のページに分割し，それぞれのページで傾き補正を行う必要がある．したがって手順（A）で，書籍画像を 1 ページずつ読み込んだ後左右のページに分割し，半ページごとに文字切り出しを行う．手順（B）において，文字領域抽出のための前処理を行う．具体的にはまず，文字領域を抽出する際に画素値の垂直射影分布を利用するので画像を 2 値化する．さらにメディアンフィルタを施し，ノイズ除去を行う．ノイズ除去によって，画像上のノイズが文字領域と誤認識されレイアウト解析に影響を与えることを防ぐことができる．手順（C）で，濃淡の垂直射影分布から閾値を算出しレイアウト解析を行う．手順（D）では，手順（C）で算出した垂直射影分布に対し Sondhi の自己相関関数[7] を用いて行間隔を算出する．最後に手順（E）において，手順（D）で算出した行間隔を基に行ごとにラベリング処理を行う．ラベリングされた黒画素連結部分を文

Vol.2011-MPS-82 No.14
2011/3/7

**図 1** 各書籍画像から切り出した文字画像例
Fig. 1 Clipped Kanji character images from different books

字領域とし，書籍画像から文字を切り出す．

本稿では自動文字切り出しで切り出すことができた文字画像を主に使用する．1 種類の文字に対して多様な字体のサンプルを得るために，出版社や出版年が異なる 9 冊の書籍画像から 1 つずつ文字画像を切り出す．つまり 1 種類の文字に対して 9 種類の PDC 特徴ベクトルを得ることができる．9 冊の書籍画像から切り出した 9 種類の文字画像の例を**図 1** に示す．図 1 に示す 9 枚の文字画像から PDC 特徴ベクトルを算出する．各次元の平均値を算出し，1,536 次元のベクトル値の標準偏差を算出すると，値は 11.53 となる．標準偏差値が小さいということは PDC 特徴ベクトルのずれが小さく，文字の揺らぎが小さいといえる．したがって少ないサンプル数での解析でも有効であると考え，本稿では 1 種類の文字に対して 9 個の文字画像を用意し，4 個をテスト用の未知データ，残りを教師用のデータとして使用する．

## 4. 評価実験

### 4.1 実験手法

本実験に用いたデータについて述べる．提案手法の有効性を実証するために，年代が様々な異なる字体で印刷された書籍から文字画像を得る．一般的に近代書籍が刊行された明治から大正期では，出版社ごとにフォントが異なる．さらに同じ出版社であっても出版年が異なれば，フォントが異なる場合が多い．したがって，1 種類の文字に対して多様な字体のサンプルを得るために，異なる書籍から 1 文字ずつ文字画像を切り出す必要がある．また，文字パターンの学習のためには，1 種類の文字について十分なサンプル数を得る必要がある．そこで，特定の文字がどの書籍に含まれているかを調べるため，本稿ではオンライン図書館の 1 種である青空文庫[8] を利用する．青空文庫は著作権が消滅した日本の文学作品を収集・公開している Web 上の電子図書館である．現在，青空文庫では約 7,200 冊の作品が提供されている．利用者は作品をテキストデータとしてダウンロードすることができる．青空文庫に登録されている作品の大部分は近代デジタルライブラリーに登録されている．したがって，青空文庫で提供されている作品のテキストデータを用いて，文字の出現頻度を容易に算出することができる．

本稿の実験で使用した書籍の一覧を**表 1** に示す．9 冊の書籍で共通して使用されている漢字を調べるために，9 冊の書籍に関して青空文庫で使用されている文字を全て調べる．その結果，表 1 の 9 冊の書籍で共通して使用されている漢字は 262 種類存在することが分かった．この 262 種類の漢字のうち無作為に選択し，近代デジタルライブラリーの書籍画像から切り出した文字画像を用いて実験を行う．本実験では 16 種，32 種，64 種，128 種，256 種と文字種を増やし，各種類数での認識率を調べる．

全ての画像は，大きさが不定の 2 値またはグレイスケール画像である．画像データに前処理を施す．まず PDC 特徴を計算する際，画像を 2 値化する必要がある．次に，マスクサイズ 3×3 のメディアンフィルタによるノイズ除去を行う．さらに PDC 特徴ベクトルの計算を正しく行うためには全ての文字画像に対し，同じ位置に文字が描画されている必要があるため，余白除去と位置補正を行う．これらの前処理によって，全ての画像は 128×128pixel の正方形の 2 値画像に変換される．

前処理を施した画像から PDC 特徴を計算し，SVM の学習を行う．なお本実験では，SVM のライブラリである LIB-SVM[9] を使用して実験を行う．

### 4.2 実験結果と考察

本稿では，漢字の種類を 16 種，32 種，64 種，128 種，256 種と増やし識別実験を行う．それぞれの種類数に関して 10 パターンずつ実験を行い，最終的な各種の認識率は 10 パターンで行った実験結果の平均値とする．

実験の結果を**表 2** に示す．表 2 に示すテストデータ数と誤答数は，各種 10 パターンの全実験の合計である．表 2 に示すように文字種を増やすと認識率は下がるが，常に 92% 以上の認識率を得ることができた．また，認識率の分散値は文字種が多くなるほど小さい．すなわち，文字種が多くなればなるほど，認識率は揺らぎが小さくなり安定していくと考えられる．この結果から，文字画像に対して PDC 特徴を抽出し，SVM で学習・識別を行うとい

Vol.2011-MPS-82 No.14
2011/3/7

**表 1** 実験に使用した近代書籍一覧
Table 1 The list of nine early-modern works

| 書籍番号 | タイトル | 著者 | 出版社 (者) | 出版年 |
|---|---|---|---|---|
| 1 | 堺事件 | 森 鴎外 | 鈴木 三重吉 | 大正 3 |
| 2 | かのように | 森 鴎外 | 籾山書店 | 大正 3 |
| 3 | 高瀬舟 | 森 鴎外 | 春陽堂 | 大正 7 |
| 4 | 吾輩ハ猫デアル | 夏目 漱石 | 大倉書店 | 明 38-40 |
| 5 | 倫敦塔 | 夏目 漱石 | 千章館 | 大正 4 |
| 6 | 煙草と悪魔 | 芥川 竜之介 | 新潮社 | 大正 11 |
| 7 | 奇怪な再会 | 芥川 竜之介 | 金星堂 | 大正 11 |
| 8 | 報恩記 | 芥川 竜之介 | 而立社 | 大正 13 |
| 9 | 或る日の大石内蔵之助 | 芥川 竜之介 | 文芸春秋社 | 大正 15 |

**表 2** SVM による実験結果
Table 2 The number of errors and the recognition rate

| 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 |

う提案手法がフォントの定まらない近代書籍の漢字認識に対して有効な手法であることが分かる．

誤認識例を**図 2**，**図 3**，**図 4** に示す．図 2 において部首が同じ漢字同士での誤認識の例を示す．また図 3 において類似した構造を持つ漢字同士での誤認識の例を示す．図 2 の Case1 から Case4 と，図 3 の Case11 から Case17 は，前処理の段階で文字線が欠け，誤認識が起こったことが分かる．つまり，文字線の欠如により類似した特徴がより強調されたと考えられる．Case1 から Case4 の場合類似した特徴は部首，Case11 から Case17 の場合類似した特徴は水平・垂直方向の文字線や文字線の傾きが考えられる．一方，図 2 の Case5 から Case8 と，図 3 の Case18 から Case20 は，前処理の段階で複雑な文字線の構造が取り除かれてしまい，部首など類似した特徴部分から誤認識が起こったと考えられる．特に Case6，8，19，20 の画像はノイズが激しい．また図 2 の Case9 は，前処理で取り除くことができなかったノイズの影響で誤認識が起こったと考えられる．さらに，図 2 の Case10 と，図 3 の Case21 から Case24 は，比較的明瞭な画像であるのでノイズの影響による誤認識である可能性は低い．しかし，それぞれ垂直方向の文字線や文字線の傾きなどの点で類似している


**図 2** 部首が同じ漢字同士での誤認識例
Fig. 2 Miss-recognized characters with the same radical indices

部分があると判断され，誤認識が起こったと考えられる．次に図 4 において類似した構造を持たない漢字同士での誤認識の例を示す．Case25 から Case28 は，Case11 から Case17 の場合と同じ理由で誤認識が起こっている．また Case29 から Case34 は，Case5 から Case8 と Case18 から Case20 の場合と同じ理由で誤認識が起こったと考えられる．

このように文字画像中にノイズが含まれると，前処理として行う文字余白の除去が意図したように行われなかったり，またノイズの黒点を文字線と認識することで PDC 特徴ベクトルの計算が適切に行われなかったりすると考えられる．このことから，外郭方向寄与度法による PDC 特徴ベクトルの計算は，画像中のノイズに大きく影響を受けることが示される．

以上の考察より，誤認識された文字を次の 3 つのパターンに分類する．

（a） 部首が同じ漢字同士での誤認識
（b） 部首は違うが類似した構造を持つ漢字同士での誤認識
（c） 類似した構造を持たない漢字同士での誤認識

さらに**表 3** は書籍ごとの誤認識数を示す．表 3 より書籍番号 2，6，7，8，9 による誤認識が多いことが分かる．**図 5** に書籍番号 2，6，7，8，9 から切り出した画像例を示す．図 5 より書籍番号 2，6，7，8，9 は紙の劣化が激しくノイズを多く含んでいることが分かる．こ

Vol.2011-MPS-82 No.14
2011/3/7


**図 3** 部首は違うが類似した構造を持つ漢字同士での誤認識例
Fig. 3 Miss-recognized characters with similar structures

の結果からも，書籍画像に含まれるノイズが誤認識に大きな影響を与えることが示される．さらに**表 4** において文字数と誤認識パターン a，b，c ごとの誤認識数の関係を示す．表 4 より，3 つの誤認識パターンのうち，類似した構造を持つ漢字同士での誤認識であるパターン a，b の占める割合と類似した構造を持たない漢字同士での誤認識パターン c の占める割合はほぼ同じであることが分かる．

したがって本実験結果から誤認識の原因は次の 2 点が挙げられる．1 点目は書籍画像の紙質の劣化や汚れで生じたノイズによる誤認識である．そして 2 点目が漢字の特徴として挙げられる，部首などの類似構造や水平・垂直方向の文字線や文字線の傾きなどの類似形状を持つ文字同士での誤認識である．

## 5. ま と め

本稿では，近代書籍をテキストデータ化するために開発された多フォント漢字認識手法の有効性を実証した．提案手法を実証するために，文字種を 262 種類に拡大した評価実験を


**図 4** 類似した構造を持たない漢字同士での誤認識例
Fig. 4 Other miss-recognized characters

**表 3** 書籍ごとの誤認識数
Table 3 Miss-recognition by title

| 書籍番号 | タイトル | 16 種類 | 32 種類 | 64 種類 | 128 種類 | 256 種類 |
|---|---|---|---|---|---|---|
| 1 | 堺事件 | 0 | 0 | 0 | 4 | 10 |
| 2 | かのように | 2 | 14 | 18 | 49 | 126 |
| 3 | 高瀬舟 | 0 | 1 | 2 | 9 | 20 |
| 4 | 吾輩ハ猫デアル | 0 | 0 | 0 | 0 | 0 |
| 5 | 倫敦塔 | 1 | 2 | 17 | 35 | 109 |
| 6 | 煙草と悪魔 | 0 | 7 | 16 | 57 | 128 |
| 7 | 奇怪な再会 | 2 | 10 | 20 | 43 | 110 |
| 8 | 報恩記 | 6 | 8 | 29 | 64 | 159 |
| 9 | 或る日の大石内蔵之助 | 2 | 11 | 14 | 50 | 110 |

行った．文字種を拡大すると，識別実験に必要な文字画像数が膨大となるため，本稿では，書籍画像に対して自動文字切り出しを行った．識別実験では，自動文字切り出しで切り出すことができた文字画像を主に使用した．近代デジタルライブラリーにて公開されている近代書籍 9 冊に共通して使用されている漢字 262 種類を切り出し，PDC 特徴ベクトルを抽出した．本稿では 262 種類の漢字から無作為に 16 種，32 種，64 種，128 種，256 種と文字種

Vol.2011-MPS-82 No.14
2011/3/7

**表 4** 誤認識パターンごとの誤認識数
Table 4 Miss-recognition by pattern

| 文字種 | パターン a | パターン b | パターン c |
|---|---|---|---|
| 16 種類 | 1 | 5 | 7 |
| 32 種類 | 5 | 19 | 29 |
| 64 種類 | 11 | 42 | 63 |
| 128 種類 | 54 | 104 | 153 |
| 256 種類 | 137 | 291 | 344 |

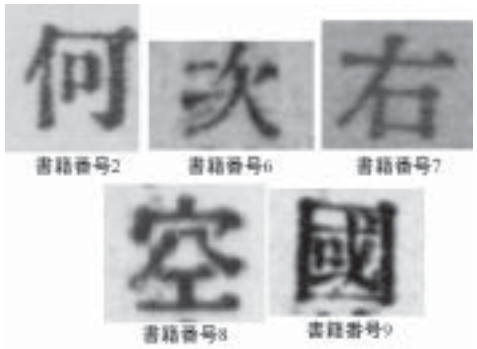


**図 5** 誤認識した文字切り出し画像例
Fig. 5 Examples of miss-recognized clipped kanji characters

を増やし，抽出した特徴ベクトルを用いて SVM で学習を行い，未知データに対する認識率を調べた．5 セット（16 種，32 種，64 種，128 種，256 種）の各認識率は各セット 10 回実験を行った結果の平均値とした．実験結果から，16 種類の場合 97.969%，32 種類の場合 95.859%，64 種類の場合 95.469%，128 種類の場合 93.926%，256 種類の場合 92.461% の認識率を得た．

この結果から，文字画像に対して PDC 特徴を抽出し，SVM で学習・識別を行うという提案手法がフォントの特定されない近代書籍の漢字認識に対して有効な手法であることが実証できた．本稿では誤認識の原因を 2 点挙げた．1 点目が書籍画像の紙質の劣化や汚れで生じたノイズによる誤認識である．そして 2 点目が漢字の特徴として挙げられる，部首などの共通構造を持つ文字同士での誤認識である．識別精度をさらに向上させるために，紙質の劣化に対応できるノイズ除去の改善やクラスタリングが課題として挙げられる．

## 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) 近代デジタルライブラリーパンフレット: http://kindai.ndl.go.jp/information/kindai(jpn).pdf
4) Chisato Ishikawa, Naomi Ashida, Yurie Enomoto, Masami Takata, Tsukasa Kimesawa, and Kazuki Joe.: Recognition of Multi-Fonts Character in Early-Modern Printed Books, In Proceedings of International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'09), pp.728-734 (2009/7).
5) 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，Vol.J66-D No.10，pp.1185-1192，(1983).
6) 日本規格協会: www.jsa.or.jp/
7) Man Mohan Sondhi: "New Methods of Pitch Extraction", IEEE Transactions on Audio and Electroacoustics, Vol.AU-16, No.2, June 1968, pp.262-266.
8) 青空文庫 http://www.aozora.gr.jp/
9) Chih-Chung Chang and Chih-Jen Lin.: LIBSVM: a library for support vector machines: http://www.csie.ntu.edu.tw/˜cjlin/libsvm, (2001).