2009－MPS－73 (52)
2009／3／6

近代書籍に特化した多フォント活字認識法

芦田 尚美 * 高田 雅美 * 木目沢 司 † 城 和貴 *
ashida@ics.nara-wu.ac.jp
* 奈良女子大学大学院 人間文化研究科 情報科学専攻
† 国立国会図書館

概要

国立国会図書館は，所蔵する明治から大正期にかけての近代書籍を，近代デジタルライブラリーとしてWeb上で一般に公開している．公開されている書籍は，全て画像としてデータ化されており，全文検索を用いて書籍内容の検索を行うことができないため，早急なテキスト化が求められている．しかし，旧字体を多く含み活字の種類が特定できない近代書籍は，既存OCRによるテキスト化を適用が困難である．そこで，本研究では近代書籍に特化した活字認識の手法を提案する．

Multi-Fonts Character Recognition for Early-Modern Printed Books

Naomi Ashida* Masami Takata* Tsukasa Kimesawa† Kazuki Joe*
* Graduate School of Humanity and Science, Nara Women's University
† National Diet Library

Abstract

National Diet Library has a public web site as Early-Modern Digital Library for books in Meiji and Taisho periods. Since the archive is digitalized as images, text search for the books is not applicable and, therefore, required. Conventional OCR systems are not a good tool for such modern books because they have various font sets and most of them are very noizy. In this paper, we propose a character recognition method for early-modern printed books.

1 はじめに

国立国会図書館 [1] は明治から大正期にかけての書籍約160,000点を所蔵している．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] というプロジェクトが開始されている．このプロジェクトでは，書籍をページごとにマイクロフィルムに撮影した画像を，Web上で公開している．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，インターネットに繋ぐことのできる環境があれば，利用者はいつでも書籍閲覧をすることができる．

しかし，近代デジタルライブラリーで公開されている書籍は画像データであるので，全文検索などの通常のテキストデータを扱う際に利用する機能は存在しない．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，各出版社や出版された時代により異なる種類の活字が使用されており，さらにノイズの問題もあるため，既存のOCR適用が極めて困難である．

これらの問題を解決するために，本研究では近代書籍に特化した多フォント活字認識を行う手法を提案する．出版社や出版年代ごとに異なるフォントの違いを吸収するため，手書き漢字認識に用いられる特徴抽出法によって文字画像の特徴抽出を行う．既存のOCR技術を参考にし，近代書籍OCRが要求する機能を論じる．特徴ベクトル化した文字データを分類するため，パターン認識の一手法であるサポートベクターマシン (Support Vector Machine, SVM)[5] を用いて学習・分類を行う．

本稿の構成は次のようになる．2節において，本研究で使用する特徴ベクトル抽出の手法である外郭方向寄与度 (PDC) 特徴について説明する．PDC特徴は手

書き漢字認識のために考案された特徴抽出法であり，本研究で行うフォントが特定できない活字認識において有効もであると考える．3節において，本稿が提案する手法について述べる．既存のOCRが行う処理を参考に，本稿が実装する近代活字OCRの機能について論じる．続く4節で，提案する手法の有効性を検証するために実験を行う．近代書籍の文字画像を用いてフォントが特定されない活字認識を行い，実験結果から考察を行う．最後に5節において本稿についてまとめる．

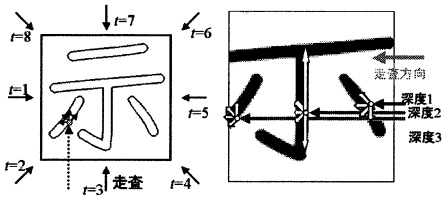

図 1: 文字の走査

2 PDC特徴

本稿で使用する文字画像の特徴抽出法である外郭方向寄与度 (Peripheral Direction Contributivity, PDC) 特徴 [3] について詳細を述べる．PDC特徴は，丁寧に書いた楷書体手書き漢字を正しく読み取る能力が高い特徴抽出法である．PDC特徴はパターン整合性に基づく特徴で，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種類の文字線構造情報を反映する．文字線の複雑さは線密度 (文字線の本数) を，文字線の方向と接続関係は方向寄与度を用いて抽出できる．文字線の相対位置関係は文字の n 番目の外郭形状を用いて表すことができる．

方向寄与度は，文字内の各黒点について4次元ベクトルで表される．文字線内の黒点 P の方向寄与度 $\mathbf{d}_P$ を $\mathbf{d}_P = (d_{1P}, d_{2P}, d_{3P}, d_{4P})$ で表す．各要素 d_{mP} $(m = 1,2,3,4)$ は点 P から45°ずつ異なる8方向に触手を伸ばして求められる黒点連結長 l_i $(i = 1,2,\ldots,8)$ を用いて

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

で定義される．

文字パターンを45°おきに8方向から走査した場合に横切る文字線の1本目から n 本目までの輪郭点をプロットしてできる形状を第 n 外郭形状と呼ぶ．また，この n を外郭深度と呼ぶ．

第 n 外郭形状は漢字に対して次の性質を持つ [3].

1. 漢字は各走査方向で，外郭深度 $n = 3$ までとれば，文字線のほとんどすべての輪郭点 (96%以上) を含む．
2. $n = 1$ で文字の外郭形状を表すことができる．
3. 各文字線の相対位置関係は走査方向とその n の値で表される．
4. 文字線輪郭を含むことができる最小の n の値は文字線の複雑さを表す．

この第 n 外郭形状の性質を利用して，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を反映するPDC特徴を作成する．

PDC特徴の抽出方法を以下に述べる．文字を45°おきに8方向から走査し，横切る各文字線の最初に横切る輪郭点での方向寄与度を投影軸に投影する．外郭深度の最大値を N とした場合，各方向の走査で1本目から N 本目までの方向寄与度が成分ごとに投影される．図1右上は，$N = 3$ とした場合の方向寄与度を求めるための走査の様子である．走査方向を t $(t = 1,2,\ldots,8)$ とし，n 本目に横切った文字線の方向寄与度 m 成分 $(m = 1,2,3,4)$ の投影軸上の任意の点 z における関数を $H_{tmn}(z)$ とおく．投影軸を16区間に等分割し，各区間内の関数 $H_{tmn}(z)$ の値を平均して，PDC特徴が抽出される．この特徴の各要素を $P_{tmn}(k)$ $(k = 1,2,\ldots,16)$ とおくと，PDC特徴ベクトル P_N は次のような $512 \times N$ 次元ベクトルで表される．

$$\begin{aligned} P_N = \quad & (P_{111}(1), P_{111}(2), \ldots, P_{111}(16), P_{112}(1), \\ & \ldots, P_{11N}(16), P_{211}(1), \ldots, P_{tmn}(k), \\ & \ldots, P_{84N}(16)) \end{aligned}$$

これにより，n が大きい部分にも要素があれば，入力文字は複雑な文字線構造を持った字種であり，逆にその要素が全て0ならば，入力文字は簡単な文字線構造で構成されていることが分かる．

文字の走査の様子を図1に示す．8方向から走査を行い，到達した各外郭深度の点から黒点連結長を数え，走査方向に垂直な投影面に $m = 1,\cdots,4$ の黒点連結長を投影する．黒点連結長を縦横斜めの8方向の投影軸に投影し，各投影軸を16の区間に分割し，各区間の平均を求める．

[3] においてPDC特徴ベクトルによる文字の認識は，文字品質の悪い手書き漢字データセットに対し重み付き距離を用いた分類を行い95.4%の認識率を示している．このことから，フォントの特定できない活字認識を行う本研究においてもPDC特徴ベクトルを用いた分類が有効であることが予想できる．

3 提案手法

[4] において挙げられている一般的なOCRを図2に示す．一般的なOCRが持つ機能のうち，本稿では前処理，特徴抽出，識別部分からなる文字分類エンジンの開発を行う．文字の分類を行うためには，文字画像

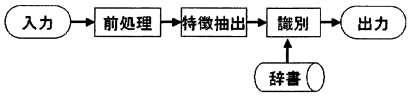

図 2: システムの概要

を何らかの特徴ベクトルとして表現するための画像の特徴抽出の方法と，抽出した特徴ベクトルを識別する分類器が必要である．

本稿が提案する手法の流れを次に示す．

1. 入力パターンとして一文字ごとに分割された文字画像を読み込む．
2. 与えられた文字画像に対し，2 値化・サイズの正規化・ノイズ除去の処理を行う．
3. 文字の PDC 特徴を抽出する．2) にて前処理を施された文字画像を周囲 8 方向から走査し，文字線の構造を得る．文字線の黒点連結長から PDC 特徴を計算する．文字種ごとに一つのクラスとし，ラベリングを行う．
4. 作成した特徴ベクトルに対し機械学習を行う．学習済み分類器は，未知データが入力されたとき照合すべき辞書となる．

手順 1) で文字画像を読み込んだ後，2) で画像に対し前処理を行う．PDC 特徴を計算する際，画像上の任意の点の画素値は黒か白のどちらかでなければならないため，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ことができる．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であるので，ノイズの除去を行う必要がある．ノイズの除去によって，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．

3) で PDC 特徴を計算し，4) で機械学習を行う．この際，クラスを表すラベルと 3) で求めた特徴ベクトルの各要素を並べたものの対が，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 50 個を無作為に選び，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学習を行う．学習済み分類器は未知のテストデータが入力されると，学習結果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．

4 実験

4.1 実験方法

提案手法の有効性を実証するため，実験を行う．実験に用いる文字画像は，近代デジタルライブラリーの書籍画像から切り出したものを用いる．実験には，「行」，「三」，「人」，「生」，「十」，「来」，「小」，「中」，「年」，「彼」の 10 種類の文字画像を使用する．それぞれの文字を 1 つのクラスとし，1 クラスにつき 100 ～ 209 個の文字画像を用いる．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の画像である．

画像データに前処理を施す．最初に文字画像を 2 値化する．次に，マスクサイズ 3×3 のメディアンフィルタを施し，ノイズの除去を行う．続けて文字画像の上下左右それぞれの辺から最も近い黒点位置から，画像の余白を計算し除去する．その後画像の縦横の長い方の辺の長さが 128 ピクセルとなるよう，線形なサイズ変更を行う．最後に，128×128 ピクセルの画像の中央に文字が位置するよう，位置の補正を行う．これらの処理を全ての文字画像に対して行う．

前処理を行った文字画像から PDC 特徴ベクトルを計算し，特徴ベクトルの学習を行う．各クラスにつき 50 個のデータを無作為に選び出し，教師データとして使用する．残りは未知データとして評価に使用する．機械学習の一手法である Support Vector Machine(SVM) [5] を用いて学習を行う．

一般的に機械学習を行うとき，特徴ベクトルの要素の数よりも十分大きな数の訓練データが必要である．本研究で使用する PDC 特徴ベクトルは 1536 次元の高次元特徴ベクトルであり，訓練データとして文字種ごとに 1536 個以上の異なる文字画像を用意することは困難である．しかし，SVM は訓練データ数よりも大きな次元の特徴ベクトルを用いる問題に対しても良好な分類性能を示す報告がなされている [6]．以上の理由から，本研究では SVM を用いて学習を行う．

SVM のパラメータを交差検定によって決定し，教師データの学習を行う．実験には，SVM ライブラリである LIB-SVM [7] を使用して実験を行った．SVM のカーネル関数は，RBF カーネルを使用した．

4.2 実験結果

学習済み SVM に未知のテストデータを入力し，学習結果の評価を行った．教師データに対する認識率が 100% となるように学習を行った．実験から，テストデータ全体では 97.8%という認識率を得た．個々のクラスでの認識率と誤答数を表 1 に示す．

表 1 から，クラス 8 の誤答率が他のクラスより高いもののほとんどのクラスについて誤答数が 0 または 1 という結果を得ることができた．全 736 個のテストデータのうち，16 個の誤認識が見られた．これらの一部を図 3 に示す．16 個のうちの約半数は図 3(a)，(b)，(c)，

表 1: 実験結果

クラス	正答数/データ数	誤答数	認識率 [%]
全体	720/736	16	97.8
クラス 1(行)	52/52	0	100
クラス 2(三)	52/53	1	98.1
クラス 3(人)	83/84	1	98.8
クラス 4(生)	50/50	0	100
クラス 5(十)	49/50	1	98.0
クラス 6(来)	84/85	1	98.8
クラス 7(小)	50/50	0	100
クラス 8(中)	147/159	12	92.5
クラス 9(年)	103/103	0	100
クラス 10(彼)	50/50	0	100

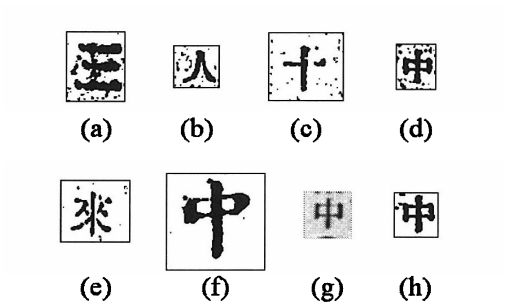

図 3: 誤認識した文字

(d) が示すように一見してノイズが多く含まれており，それ以外の文字は (e)，(f)，(g)，(h) のような目立ったノイズは見られない明瞭な画像であった．

4.3 考察

実験の結果から次の点が考察される．誤認識を起こした 16 個の画像のうち，ノイズが原因であると推察されるものは半数の 8 個であった．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と考えられる．このことから，外郭方向寄与度法による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．ノイズ以外の理由による誤認識のうち，「小」という文字として誤認識された例が 3 例，「十」の文字として誤認識された例が 4 例見られた．「年」，「彼」など今回使用した文字種の中でも比較的画数が多い文字は誤認識が見られなかった．このことから，極端に文字線が単純である場合には類似する文字構造を持った文字同士の誤認識が起こりやすいと考えられる．

クラス 8「中」の誤認識が他のクラスと比較し著しく多く見られた．クラス 8 の誤認識 12 例のうち，ノイズによるものと判断される誤認識は 5 例であった．残り 7 例は類似する文字との誤認識であると予想される．クラス 8 のみ誤認識が多い理由は，テストデータとして使用した文字画像が 159 個であり，他のクラスの 2 ～ 3 倍以上の個数のテストを行ったため，誤りが多くなったと考えられる．

5 まとめ

本研究において，フォントの特定されない近代活字の認識を行った．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から文字を切り出し，PDC 特徴ベクトルを抽出した．10 種類の漢字から抽出した特徴ベクトルを用いて SVM の学習を行い，未知データに対する認識率を調べた．実験結果から，97.8% の認識率を得た．個々のクラスでは，そのほとんどが 98% 以上の認識率を示すという結果を得た．この結果から，ノイズの少ない文字画像に対して PDC 特徴を抽出し，SVM で分類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活字認識に対して有効な手法であることが実証された．

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] 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).

[4] 入江 文平.: 道しるべ : 文字認識技術—文字認識は終わっているか?—，情報処理学会論文誌 Vol.42 No.6，(2001).

[5] Nello Cristianini and John Shawe-Taylor.: サポートベクターマシン入門, 共立出版, (2005).

[6] 賀沢 秀人.: なぜ SVM を使うのか?～ユーザーから見た利点～，2004 年電子情報通信学会総合大会，(2004).

[7] Chih-Chung Chang and Chih-Jen Lin.: LIBSVM: a library for support vector machines: http://www.csie.ntu.edu.tw/~cjlin/libsvm, (2001).