Vol.2016-MPS-109 No.9
2016/7/25

近代書籍の文字認識のための特徴抽出手法の比較

上坂和美[†1]　藤本馨[†1]　石川由羽[†1]　高田雅美[†1]　城和貴[†1]

概要：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て，いくつか特徴量を適用し，認識率の比較を行う．国立国会図書館では，多数の近代書籍を Web 上で一般公開しているが，全文検索することができない．そのため，書籍の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 特徴を用いた近代書籍の認識に特化した多フォント活字認識手法が提案されているが，より精度の高い認識手法が求められる．そこで本研究では，特徴量として，PDC 特徴とともに，現在まで適用したことがなかった加重方向指数ヒストグラム特徴，セル特徴を適用し、これら 3 つの特徴量の認識の比較と分析を行う．

キーワード：近代書籍，多フォント活字認識手法，PDC 特徴，加重方向指数ヒストグラム特徴，セル特徴

Comparison of Feature Extraction Methods for Early-Modern Japanese Printed Character Recognition

Kazumi Kosaka[†1] Kaori Fujimoto[†1] Yu Ishikawa[†1] Masami Takata[†1] and Kazuki Joe[†1]

Abstract: In this paper, we compare feature extraction methods to use some feature 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. For growing in performance of this method, we compare feature extraction methods to use the weighted direction index histogram feature, the cellular feature and the PDC feature.

Keywords: early-modern Japanese printed books, multi-font Kanji character recognition method, PDC feature, weighted direction index histogram feature, the cellular feature

1. はじめに

国立国会図書館[1]では，明治期から昭和初期にかけて刊行された近代書籍とよばれる書籍を図書およそ 35 万点，雑誌およそ 8 千点を所蔵している．近代書籍は，哲学から産業，文学，自然科学，芸術等幅広い分野にわたるものである．これらは現在絶版になっているものや出版数の少ないものも多く，近代を研究する際において貴重な学術資料となっている．国立国会図書館では，平成 14 年から近代デジタルライブラリーといわれる Web サービスを開始した．このサービスは，近代書籍の画像データを Web 上で提供するものである[2]．近代書籍のデジタル化により，破損や紛失を防ぎ，原資料は良い状態のまま保存することができる．また，インターネット上に公開することで，いつでもどこでも無料で書籍データを閲覧することができる．近代デジタルライブラリーは平成28年5月末をもってサービスを終了し，国立国会図書館デジタルコレクションに統合される予定である[3][a]．統合後は引き続き，近代デジタルライブラリーで提供されていた資料は，同様に国立国会図書館デジタルコレクションの Web サイト上で閲覧可能である．近代デジタルライブラリーの Web サイトでは，タイトルや著者，出版者，出版年など詳細な項目を指定して書籍を検索することができる．ただし，公開されている近代書籍の本文はテキスト化されておらず，本文内容に関する検索をする形で利用することができない．特定の言葉による検索を行うことができないことは，利便性の低下につながる．したがって早急なテキスト化が求められているが，数十万冊に及ぶ膨大な書籍を手動でテキスト化することは予算的に不可能であると考えられる．近代書籍に特化した文字認識の研究は未だ行われていないため，我々は国立国会図書館関西館に協力を仰ぎ，近代デジタルライブラリーの自動テキスト化を目指した研究に着手している．近代書籍は活版印刷であり現在のように統一された規格が存在しないため，既存の光学文字認識 (Optical Character Recognition , OCR) ソフトウェアを適用した認識率は低い．そこで手書き文字認識の手法を用いたところ近代書籍から切り出された活字の認識が可能であることを報告している[4][5][6][7]．また，近代書籍は，出版者によって書体は異なるだけでなく同じ出版者であっても時代によっても異なることも報告されている[8]．これらの理由より，手書き文字認識の手法が近代

[†1] 奈良女子大学 Nara Women's University

a 平成 28 年 6 月現在，近代デジタルライブラリーは終了し，国立国会図書館デジタルコレクションとして提供

Vol.2016-MPS-109 No.9
2016/7/25

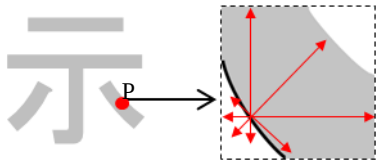

図 1:点 P における 8 方向の黒点連結長

書籍の活字認識に有用であると考えられる．現在，近代書籍の認識に特化した認識手法として多フォント活字認識手法が提案されている．この手法では，文字の特徴量として PDC 特徴を用いて表している．文献[5]では PDC 特徴を用いた認識率は，256 種でおよそ 92%となっている．発生する誤認識は，文字種の数や文字の種類によるため，今後は更なる認識率の向上が求められる．そこで現在まで，PDC 特徴以外の特徴量を近代書籍の画像認識に適用したことがないため，いくつか異なる特徴量を適用することとした．それぞれの特徴量の認識実験を行い，誤認識の傾向からより適した認識手法の提案を目指したいと考える．これまでは，低品質定型文字に対する高精度認識手法に用いる特徴量として，PDC 特徴以外に，手書き文字認識に実績がある加重方向指数ヒストグラム特徴とセル特徴を用いた手法が提案されている[9].

そこで，本論文では，PDC 特徴以外にセル特徴，加重方向指数ヒストグラム特徴を挙げ，3 つの特徴量の認識結果を比較することとした．認識結果から特徴量毎の誤認識を分析し，傾向を考察した．

本論文の構成は，以下の通りである．第 2 章では認識結果を比較するために用いる特徴量について述べる．次の第 3 章では第 2 章で述べた特徴量ごとの認識実験について述べ，結果を比較する．また，認識実験を行った際に起きた誤認識の文字種について分析し，誤認識の傾向を考察する．

2. 認識に用いる特徴量

現在用いている多フォント活字認識手法では，文字の特徴量として PDC 特徴を利用している．ここでは，PDC 特徴の他に，比較するために用いる加重方向指数ヒストグラム特徴，セル特徴に関して述べる．

2.1 外郭方向寄与度特徴
(Peripheral Direction Contributivity , PDC)

PDC 特徴[10]とは，文字線の方向性に注目し特徴量としたものである．PDC 特徴は，直線に強いという特性があるため，直線が多い漢字に適した特徴量とされている．

PDC 特徴は，文字線の方向，複雑さ，相対位置関係，接続位置関係の 4 つで表すことができる．文字線の複雑さは文字線の本数，相対位置関係は文字の外郭形状，接続位置関係は方向寄与度で表すとする．

方向寄与度は各黒画素に対して 4 次元ベクトルで示される．黒画素 P の方向寄与度をd_P= ($d_{1P}, d_{2P}, d_{3P}, d_{4P}$)で表す

と，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

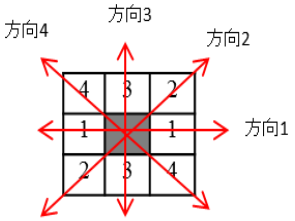

図 2 :注目画素と周辺 4 方向を表す数値

a)θ = 1 の時　　**b)**θ = 2 の時

図 3 :方向別の扇型領域

いて，d_{mP} は以下の式で定義することができる．

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

図 1 は，黒画素 P において 8 方向に伸ばした黒点連結の様子を表している．文字線の複雑さ，相対位置関係を表す外郭形状は，文字パターンを 45° ごとに 8 方向から走査し，横切る文字線の本線ごとに輪郭点をプロットしてできる形状のことである．n本目までの輪郭点をプロットしてできる形状を第n外郭形状といい，nを外郭深度と呼ぶ．文字を縦・横・斜めの 8 方向から走査し，各文字線を横切る輪郭点での方向寄与度を投影する．

投影する際の投影軸を 16 分割し，各区間内の平均値を PDC 特徴とする．PDC 特徴ベクトルP_nとすると，$P_n = 512 \times N$ で表される．$N = 3$とし，第 3 外郭形状で表すとすると，$P_n = 512 \times 3 = 1536$次元で表すことができる．

2.2 加重方向指数ヒストグラム特徴

加重方向指数ヒストグラム特徴[11]とは，文字線の輪郭線に沿った方向に着目した特徴量である．

はじめに，入力された文字を平滑化し，輪郭線を抽出する．それを 7 × 7 領域に分割し，各領域において 4 方向指数ヒストグラムを求める．注目画素(S_i)の文字の輪郭部を左回りに輪郭追跡を行い，注目画素の一つ後の要素の位置する場所(S_{i+1})によって 4 方向指数を得る．図 2 は，注目画素(S_{ij})と 4 方向指数を表す数値を示したものである．これらの方向指数をヒストグラム化し，7 × 7 の小領域の方向指数とする．次に，2 次元ガウシアンフィルタを用いて 4 × 4 の領域に，先ほど求めた 7 × 7 の小領域の方向指数を集約する．結果，集約後の小領域 4 × 4 とその領域それぞれに 4 方向指数が得られるため，4 × 4 × 4 = 64 次元の特徴ベクトルとなる．これを加重方向指数ヒストグラム特徴とする．

2.3 セル特徴

セル特徴は，画像の輝度値が急激に変化する方向であるエッジ方向に着目した特徴量である[12].
注目セルとその 8 近傍を用いて，3×3 の局所領域のエッジ方向と大きさで表される微視的特徴量を求める．求めたエッジ方向は 8 方向に量子化する．

次に，周辺 4 画素の微視的特徴量を統合し，セル空間の特徴量を求める．これは方向別に各セルにおけるエッジの大きさの線形和で表すことができる．
局所的な特徴量を求めた後は，方向別に局所領域の広がりを求める．そのために、方向別に±45°の扇型の領域を定める．扇型のパターンは 2 つあり領域に含まれるセル数が異なる．図 3 は方向別の扇型領域を示し，内部に含まれるセルを緑色で示す．扇型の領域に対し積分することで，領域を相似形でかつ単調的に大きくし，原画像のストロークとぶつかるまで積分を続ける．
セル$\mathrm{D}(i_c,j_c)$の近傍セルをA,B,C,Eとし，内部情報を$A(\theta_c,t)$のように表す．θ_cは 1~8 の数字で表現された方向，tは積分回数を表す．特徴量は，注目セルと扇型領域内のセル特徴から求められ，積分が行われるたびに更新される．

3. 実験

本章の認識実験は，第 2 章で挙げた 3 つの特徴量，PDC 特徴，加重方向指数ヒストグラム特徴，セル特徴で文字画像の特徴量を表し，認識率の比較を行うものである．

3.1 認識実験

認識を行うために用いる識別処理の手法として，サポートベクターマシン(Support Vector Machines , SVM)を用いる．SVM は, Vapnik らによって考案された教師あり機械学習の 1 つであり，高い精度をもつ[13].
使用する文字種は，JIS 第一水準漢字，JIS 第二水準漢字[14]， ひらがなからなる 2678 種類とする．各文字種に対し 6 個の画像データを用意する[15]．識別器を生成するために，学習データとして 5 個の画像データを用いる．テストデータは，残り 1 個の画像データとする．テストデータを順次変更することによって 6 回の認識実験を行う．
実験環境は，CPU は Intel Xeon Processor (8core, 2.0GHz, L3 18MB,QPI6.4GT/sec) × 12 / 96 core , メモリは 768GB(8GB × 96)のクラウド計算機と，CPU は Intel(R) Core(TM) i7-4770K CPU @ 3.50GHz, メモリは 16GB の計算機を用いる．
認識実験の結果，PDC 特徴の認識率が最も高く，86.8%となっている. つづいて加重方向ヒストグラム特徴 85.5%, セル特徴 81.1%となっている．また，PDC 特徴は特徴量の次元数が最も多いため，最も時間がかかり，クラウド計算機でおよそ 486 時間，計算機でおよそ 428 時間かかる．セル特徴は，計算機でおよそ 199 時間，加重方向指数ヒストグラム特徴は，クラウド計算機でおよそ 431 時間，計算機でおよそ 183 時間かかった．

3.2 誤認識の分析

次に, 3.1 節で行った認識実験での誤認識の文字に注目し，誤認識の傾向を分析する．
文字の誤認識の種類の分類として，以下のような項目で分ける．
a)ひらがな
b)旁が同じで偏が異なる
c)偏が同じで旁が異なる
d)垂が同じで中が異なる
e)冠が同じで中が異なる
f)脚が同じで中が異なる
g)構が同じで中が異なる

表 1 :誤認識した特徴量の割合

誤認識の特徴量	割合(%)
A	26.8
B	9.5
C	14.0
D	17.8
E	6.1
F	7.1
E	18.7

h)画質(太さ)
i)画質(かすれ)
j)画質(中が分かりにくい)
k)旧字体
また，誤認識した特徴量に着目して分類するため，以下のように定める．
A)セル特徴
B)加重方向指数ヒストグラム特徴
C)PDC 特徴
D)加重方向指数ヒストグラム特徴とセル特徴
E)PDC 特徴とセル特徴
F)PDC 特徴と加重方向指数ヒストグラム特徴
G)PDC 特徴と加重方向指数ヒストグラム特徴とセル特徴
以上の A~G に分類し, a~k に基づいて誤認識の分析を行う．
A)セル特徴とは，セル特徴のみ誤認識したものである．
A~G で分類した際の各々の割合を表したものを表 1 として示す.
セル特徴は，セル特徴のみが誤認識である場合が多く誤認識全体の 26.8%である. 加重方向指数ヒストグラム特徴は，それのみの誤認識 9.5%よりもセル特徴とともに誤認識である場合の割合 17.8%の方が大きい．PDC 特徴の誤認識の場合は，E，F が少なく，PDC 特徴のみの場合で 14.0%か，3 つすべて誤認識の場合かに大きく分けられる．また，3 つの特徴量がすべて誤認識である場合の割合も誤認識全体の割合の 2 番目で 18.7%を占める．
つづいて，a)~k)に基づき，A)~G)における誤認識の分類の割合を述べる．
まず，A)では，a)21.3%c)27.1%h)18.6%i)16.1%の項目で誤認識される割合が多い．これは，セル特徴の場合エッジに着目しているので，画像の文字線の太さの差やかすれ，ひらがなであれば曲線のつながり等の相違の影響を受けやすいためであると考えられる．
B)は，g)42.9%が最も多く，続いて i)28.6%が多い．加重方向指数ヒストグラム特徴は，領域を大きく分割して 4 方向に集約するため，構の中身が似ているものや画像がかすれているものなどは誤認識が起こしやすいと考えられる．
次に，C)では，i)29.7%j)k)18.9%で割合が大きい．よって PDC 特徴は，かすれているものの影響は受けやすいが，太さの影響は受けにくいことが分かる．また，旧字体の誤認識が多いのは，旧字体が常用漢字と一部分だけ異なるのみで形が似ているためである考えられる．
つづいて，3 つの特徴量のうち 2 つの特徴量が誤認識したもの D)~F)を考える．
D)を分類すると, h) 19.2%i)30.8%で占める割合が大きい．それは，セル特徴と加重方向指数ヒストグラム特徴はエッジや文字線の方向に着目する特徴量のため，画像の状態の影響が大きかったからだと考えられる．

また，E)では i)66.7%，F)では c)33.3%のように文字構造が似ているものや画質のかすれでの誤認識が多い．
最後に，G)を分類すると，h)24.6%i)11.6%と画質の原因や，k)28.9%と構造が似ており一部分しか違わない文字が多い．以上より，画像の太さやかすれなど画質に問題がある場合は，どの特徴量でも誤認識の原因となる割合が大きい．特に，加重方向指数ヒストグラム特徴とセル特徴は画質の影響を受けやすいといえる．また，文字構造が似ており一部分だけが異なるものである場合で誤認識しやすい．

3.3 誤認識となる文字種

3.2 節で分析したもののうち，G)3 つの特徴量がすべて誤認識したものの文字の種類をさらに分析する．

すると，各々の特徴量での正解数は，どの文字種に対してもばらつきがみられ，3 つの特徴量を用いて 6 回行った

認識実験の中で、どの特徴量を用いても 1 度も正しく認識されなかった文字種はなかった． つまり，3 つの特徴量を用いて認識結果として現れた文字の中で，最も出現回数が多かった文字を認識結果とした場合，認識率が上昇するのではないかと考えられる．

4. まとめ

本稿では，近代書籍の画像データをテキスト化するために用いる多フォント活字認識手法に使用する特徴量の比較について述べた．比較する特徴量は，PDC 特徴，加重方向指数ヒストグラム特徴，セル特徴とする．

実行時間は，最も次元数の多い PDC 特徴がかかった．認識率は， PDC 特徴が最も高く 86.8%であった．

次に，誤認識の分析を行った．誤認識したものをそれぞれの特徴量ごとに分類した割合は，セル特徴のみ誤認識した場合で 26.8%と最も高く，つづいて 3 つの特徴量すべてで誤認識した場合の 18.7%となる．

誤認識となった原因を分析した結果，加重方向指数ヒストグラム特徴とセル特徴は，文字画像の太さやかすれなど画像そのものの影響が大きいと考えられる．3 つの特徴量すべてで誤認識したものは，文字の一部分のみが異なるものや画像のかすれや太さが主な原因である．また，3 つの特徴量が 6 回すべてで誤認識したものはなかった．

そのため，今後としては，3 つの特徴量で 6 回認識させた結果から，最も出現回数が多い文字を認識結果とする多数決法を用いて認識させることにより，認識精度が向上するのではないかと考えられる．

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．

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] 国立国会図書館デジタルコレクション
http://dl.ndl.go.jp/
[4] 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 (P DPTA09), Vol.Ⅱ，pp. 728-734(2009).
[5] 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).
[6] 粟津妙華，高田雅美，城和貴．：遺伝的プログラミングを用いた近代書籍からのルビ除去，情報処理学会論文誌．数理モデル化と応用(TOM)，Vol. 6，No. 2，pp. 53-62(2013).
[7] 粟津妙華，高田雅美，城和貴．：活字データの分類を用いた進化計算による近代書籍からのルビ除去，情報処理学会論文誌．数理モデル化と応用(TOM)，Vol. 8(1)，pp. 72-79(2015).
[8] 福尾真実，高田雅美，城和貴．：同一出版者の近代書籍に対する漢字認識評価，情報処理学会研究報告．数理モデル化と問題解決(MPS)，2012-MPS-90(26)，1-6 (2012-09-12)
[9] 宮本一正，熊野信太郎，杉本喜一，玉川光明，英保茂．：複数特徴量を用いた低品質定型文字の一認識手法，電子情報通信学会論文誌．(D)，Vol.J82-D，No4，pp. 771-779(1999)
[10] 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).
[11] 鶴岡信治，栗田昌徳，原田智夫，木村文隆，三宅康二．：加重方向指数ヒストグラム法による手書き漢字・ひらがな認識，電子通信学会論文誌.(D)，Vol.J70-D，No.7，pp.1390‐1397 (1987).
[12] 岡隆一：：セル特徴を用いた手書き漢字の認識，電子通信学会論文誌.(D), Vol.J66-D, No.1, pp.17‐24 (1983).
[13] Cristianini，N. andShawe-Taylor，J.：サポートベクターマシン入門，共立出版(2005).
[14] 日本工業規格：https：//www.jisc.go.jp/
[15] 粟津妙華，福尾真実，高田雅美，城和貴．：多フォント漢字認識手法における各カテゴリと必要教師データ数の分析，情報処理学会研究報告．数理モデル化と問題解決(MPS)，2014-MPS-97(10)，1-6(2014-02-24)