Vol.2017-MPS-114 No.4
2017/7/17

アンサンブル学習を用いた近代書籍文字認識

藤本馨　石川由羽　高田雅美　城和貴

概要：近年，書籍のアーカイブ化が注目されている．アーカイブ化とは，書籍から抽出されるテキストデータがインターネット上のビッグデータとして使われるために，テキストデータを抽出することを意味している．現在，書籍はDTP で作成されることが多く，自身がテキストデータを含んでいるのでテキスト化する際に問題はない．しかし，半世紀以上前に出版された古い書籍は，現在の書籍と比べてフォントの違いがあるためテキスト化することは困難である．私たちは過去 10 年に渡り，近代書籍文字認識の研究をしている．本稿では，アンサンブル学習を行いた近代書籍文字手法の改良について述べる．

キーワード：近代書籍，PDC 特徴，加重方向指数ヒストグラム特徴，セル特徴，アンサンブル学習

1. はじめに

インターネット上での知識獲得の手法として、知的エージェントによる Web マイニングやテキストマイニングが利用されている。検索対象となるデータは、飛躍的に増えており、書籍の内容をテキストとして検索しやすい方向にニーズが進んでいる。従って、その方向性でさらなる技術革新が進むものと予想される。

現在書籍は DTP で作成されることが大部分であり、テキスト化は既になされた状態であるのに対し、過去に出版された書籍のテキスト化は人手によるものと OCR によるものに大別される。OCR は実用化され現在広く利用されているとはいえ、誤認識の可能性は排除できないため、重要な書籍のテキスト化には人手が必要となる。しかしながら知的エージェントによる知識獲得で重要なのは OCR を利用した広範囲な自動テキスト化である。また，書籍の自動テキスト化を行う際の制約で最も厳しいものは出版時期である。即ち現在の活字フォントが規格化される前のいわゆる活版印刷の時代の書籍である近代書籍に対する既存の OCR の利用は困難である[1]。近代書籍とは日本において明治から昭和初期にかけて刊行されたものである。国立国会図書館では 35 万点にも及ぶ近代書籍を国立国会図書館デジタルコレクション[2]という Web サイトで画像データとして公開している．

今世紀に入ってからストレッジの密度は飛躍的に向上し，書籍を画像としてデジタル化し保存することが可能となってきた。現在のところ各方面での近代書籍のアーカイブ化は貴重な資料の保管が第一の目的であるが、そこに現在の検索可能な知識とは異なる種類のものがあると考えた場合、近代書籍用 OCR（Optical Character Recognition , OCR）のニーズは格段に上がる。しかしながら，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かったため，我々は国立国会図書館関西館に協力を仰ぎ，近代書籍文字の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 OCR を適用した認識率は低い[1]．近代書籍は，出版者ごとに書体が異なるだけでなく同じ出版者であっても出版された時代が異なると，書体も異なることが報告されている[3]．これらの理由より，手書き文字の認識手法が近代書籍の活字認識に有用であると考えられる．そこで我々は近代書籍の認識に特化した認識手法として手書き文字認識手法をベースにした多フォント活字認識手法を提案している[4]．この手法では，文字の特徴量として PDC 特徴を、識別手法として教師あり機械学習の 1 つである SVM を用いる．さらに特徴量を，加重方向指数ヒストグラム特徴，セル特徴に変更し，認識実験を行っている．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 つの各特徴量における全ての実験で誤認識されるものはない．これより，複数の特徴量を用いることで認識率が向上すると考えられている[5]．

そこで，本稿では，アンサンブル学習を用いた近代書籍文字認識を行い認識率の向上を目指す．文字画像の特徴量として，PDC 特徴，加重方向ヒストグラム特徴，セル特徴の 3 つの各特徴量を適用する．また，識別器として既存手法に用いられる SVM に加え，OLVQ1 を用いる．3 つの特徴量と 2 つの識別器を組み合わせることにより 6 つの弱識別器を構成し，それらを用いてアンサンブル学習を行い，近代書籍文字を認識する．

本稿の構成は，以下の通りである．第 2 章では提案する認識手法について述べる．第 3 章では，アンサンブル学習を行うための予備実験について述べ，行った際に起きた各特徴量の誤認識について分析し，誤認識の傾向を考察する．第 4 章では、6 つの弱識別器を用いたアンサンブル学習による認識実験を行い，第 5 章ではまとめを述べる．

2. 認識手法

本稿では，アンサンブル学習を用いた近代書籍文字の認識手法を提案する．はじめに，近代書籍文字画像データから PDC 特徴，加重方向指数ヒストグラム特徴，セル特徴の 3 つの特徴量を抽出する．抽出した特徴量を，SVM と OLVQ1 の 2 つの各識別器により認識する．アンサンブル学習は弱学習器を増やすことにより認識能力が向上すること

が知られている．したがって，本稿では SVM と仕組みの異なる識別器 OLVQ1 を新たに加え, 3 つの特徴量と組み合わせることにより弱識別器を増やしアンサンブル学習を行うことにより，認識率の向上を目指す．3 つの特徴量と 2 つの識別器を組み合わせて生成される 6 つの弱学習器を用いて，アンサンブル学習による認識実験を行う．

実験に用いる重みは，Adaboost.M1 のアルゴリズムに従い，6 つの弱学習器が学習データを学習し，学習データを認識した結果から計算する．Adaboost.M1 は用いる弱識別器の誤認識に応じて弱点を補強するような重みを計算する．各弱学習器が学習データを学習しテストデータを認識した結果を用いて，Adaboost.M1 のアルゴリズムにより算出される各弱学習器の重みにより認識結果を得る．

3. 予備実験

本稿で行うアンサンブル学習を行うための予備実験として，各弱識別器による認識実験について 3.1 節で説明する．続いて 3.2 節で誤認識の分析を行い各弱識別器の誤認識の傾向について考察する．

3.1 各弱識別器による認識実験

認識対象となる文字は使用する文字種は，JIS 第一水準漢字，JIS 第二水準漢字， ひらがなからなる 2678 種類とする．これを 1 画像データセットと呼ぶ．このデータセットを 6 つ用意する[5]．データセットに含まれる文字画像データは複数の時代，出版者から集められている．データセットそれぞれに含まれる各文字種の画像には同じ出版者の書籍から取り出した文字は含まれていない．時代とは，大正時代，昭和時代のことであり，出版者とは，日吉堂，駸々堂，春陽堂，平民社，大倉書店，岩波書店，聚英閣，左久良書房，新潮社，実業之日本社である．

文字画像から抽出する 3 つの特徴量の次元数を説明する．PDC 特徴は，投影する際の投影軸を 16 分割し，文字を縦・横・斜めの 8 方向から走査する．各文字線を横切る輪郭点での走査で 3 本目までの 4 方向で表される方向寄与度を求める．したがって，PDC 特徴の次元数は 16×8×4×3=1536 次元である．加重方向指数ヒストグラムは，2 次元ガウシアンフィルタを用いて 18×18 の小領域に 4 方向指数ヒストグラムを集約する．したがって，18×18×4=1296 次元の特徴となる．セル特徴は，セル空間の大きさが 148×148 とするので 1152 次元で表す．

SVM を識別器として用いる場合, 特徴量データに対しグリッド検索を行い適切なパラメータを求める．RBF カーネルは次元の多い特徴量の認識を得意としているので，RBF カーネルを使用しパラメータを入力することで認識を行う．OLVQ1 を識別器として用いる場合，代表ベクトル数が 5 だと Adaboost.M1 で学習データを学習し認識する際に誤認識が起こらなくなり、弱識別器として使用できなくなる．

表 1：弱識別器単体による平均認識率

弱識別器(特徴量・識別器)	平均認識率
PDC 特徴・SVM	87.47%
加重方向指数ヒストグラム・SVM	85.55%
セル特徴・SVM	81.01%
PDC 特徴・OLVQ1	75.94%
加重方向指数ヒストグラム・OLVQ1	78.42%
セル特徴・OLVQ1	79.38%

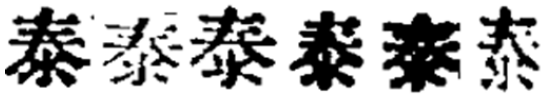

図 2：OLVQ1 が認識できる文字種の画像

したがって，本稿では、代表ベクトル数を 4，初期学習係数を 0.5 としている．

6 つの画像データセットのうち， 5 つの画像データセットを学習データとして用いる．テストデータは，残り 1 つの画像データセットとする．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．初めに 6 つの画像データセットから，3 つの特徴量をそれぞれ抽出する．その後，学習データにあたる 5 つの画像データセットの特徴量を SVM と OLVQ1 のそれぞれで学習し, テストデータにあたる 1 つの画像データセットの特徴量を認識する．各弱識別器につき 6 回の認識実験の平均認識率を表 1 に表す．

3.2 弱識別器ごとの誤認識の分析

特徴量が PDC 特徴，識別器が SVM である弱識別器は，かすれている文字の影響を受けやすく，文字線の太さの影響を受けにくいことが分かっている[5].

特徴量が加重方向指数ヒストグラム，識別器が SVM である弱識別器は，漢字の中心部分が似ているものや，画像がかすれているもじの影響を受けやすいことが分かっている[5].

特徴量がセル特徴，識別器が SVM である弱識別器は，画像の文字線の太さやかすれ，ひらがなの曲線のつながり等の相違の影響を受けやすいことが分かっている[5].

3 つの特徴量を用いて OLVQ1 を識別器とする場合，SVM を識別器とする場合に起きる誤認識の中で OLVQ1 が認識できる文字の割合は，およそ 71.80%となる．OLVQ1 は，各文字種につき用意されている 6 つの画像データのうち，文字線が太い画像と画像がかすれている画像が混じっているが，軸線を判別できる文字種を認識できる．図 2 は OLVQ1 が認識できる文字の文字画像の一例である．

特徴量が PDC 特徴, 識別器が OLVQ1 である弱識別器は，SVM が誤認識する文字のうちの約 38.79%を認識する．

図 3：文字の一部分がインクのにじみにより潰れている文字を含む文字種

3 つの特徴量を用いて OLVQ1 を識別器とする弱識別器の中で一番認識率が低い．また，この弱識別器は SVM の誤認識のうち，文字線の太さの影響を受けにくい．
特徴量が加重方向指数ヒストグラム特徴, 識別器が OLVQ1 である弱識別器は，SVM が誤認識する文字のうちの約 42.04%の文字を認識する．この弱識別器は SVM の誤認識のうち，各文字種につき用意されている 6 つの画像データのうち，一部がかすれているものや文字線の太さが異なるものが混じっているが，文字線の軸線が共通している特徴を持っている文字種を認識できる．

特徴量がセル特徴, 識別器が OLVQ1 である弱識別器は，SVM が誤認識する文字のうちの約 37.59%の文字を認識する．この弱識別器は SVM の誤認識のうち，文字の一部分がインクのにじみにより潰れている文字の影響を受けにくくなっている．図 3 は文字の一部分がインクのにじみにより潰れている文字を含む文字種である．

以上の考察より，OLVQ1 を識別器とした弱識別器は，SVM を識別器とした弱識別器の誤認識を補っていることが分かる．このことより，これらの弱識別器を用いることはアンサンブル学習に有効であることがわかる．

3.3 文字種に対する誤認識する弱識別器の数

Adaboost.M1 に用いる弱識別器の誤認識は分散している必要がある．2678 文字種の各文字に対して，誤認識を起こす弱識別器の数を，認識するデータセットごとに分析し，Adaboost.M1 に適用する弱識別器として適切であるかを検討する．
2678 文字の文字種ごとに 6 つの弱識別器のうち，いくつの弱識別器が誤認識しているかを調べる．6 つのテストデータをそれぞれ認識し，6 つの弱識別器全てにおいて誤認識を起こした文字種の割合の平均は 3.78%である．5 つの弱識別器においては 3.34%である．4 つの弱識別器においては 5.06%である．3 つの弱識別器においては 5.62%である．2 つの弱識別器においては 9.64%である．1 つの弱識別器においては 14.25%である．この結果より，1 つの弱識別器が誤認識する文字種の割合が高いことが分かる．

また，全テストデータの各文字種に対する誤認識する弱識別器の数の平均は 1.12 であり，各テストデータに対する標準偏差の平均は 1.44 である．このことから，6 つの弱識別器が誤認識する文字種は分散していることが分かる．したがって，この 6 つの弱識別器を用いて Adaboost.M1 を行い認識することは有効であると言える．

表 2 : 弱識別器による平均認識率，Adaboost.M1 による認識結果

認識データ	弱学習器による平均認識率	Adaboost.M1
テストデータ 1	77.81%	86.86%
テストデータ 2	83.51%	90.14%
テストデータ 3	84.25%	90.70%
テストデータ 4	84.01%	91.60%
テストデータ 5	82.47%	88.46%
テストデータ 6	79.44%	85.55%

表 3：誤認識を起こす文字種数と実験数

全ての実験で誤認識	0/1228
5 回の実験で誤認識	14/1228
4 回の実験で誤認識	15/1228
3 回の実験で誤認識	93/1228
2 回の実験で誤認識	271/1228
1 回の実験で誤認識	835/1228

4. 認識実験

アンサンブル学習を用いた認識実験について説明する．5.1 節では実験内容について述べ，5.2 節では実験結果について分析し，考察を行う．

4.1 実験内容

用いる弱識別器は 3 つの特徴量，PDC 特徴，加重方向指数ヒストグラム特徴，セル特徴と 2 つの識別器 SVM と OLVQ1 を組み合わせた 6 つの弱識別器を用いる．認識対象，特徴量の次元数，識別器の設定は予備実験と同様である．

予備実験と同様に 6 回の認識実験を行う． Adaboost.M1 のアルゴリズムに基づき，学習データを訓練し算出される重みを用いて各テストデータを認識する．6 つの弱識別器による 1 つのテストデータに対する平均認識率，Adaboost.M1 による認識結果を表 2 に表す．

認識結果より，Adaboost.M1 による認識結果は，弱学習器単体を用いる認識結果の平均よりも高い．認識率は向上しており，Adaboost.M1 による認識結果の平均認識率は 88.88%となる．

4.2 誤認識の分析

認識実験において誤認識される文字数は 1228 文字である．表 3 に誤認識を起こす文字種数と実験数の関係を表す．6 回の実験全てにおいて，誤認識される文字種はない．1・2 回の実験で誤認識される文字種は，学習データの文字画像が異なることが原因による誤認識であると考えられる．6 回の実験中，半数以上の実験で誤認識を起こす文字種について分析する．

3 回の実験で誤認識を起こす文字種は，6 つの画像データのうち，2 つ以上の画像がかすれにより一部欠損して

図 4：にじみがひどく文字の一部が潰れている文字種

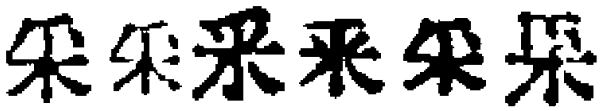

図 5：字体により文字が異なる文字種

図 6：中身分かりにくい文字が過半数を占める文字種

いるものや，にじみがひどく文字の一部が潰れているものを含む文字種である．図 4 はにじみがひどく文字の一部が潰れている文字種を表す．一部が読み取りにくい状態のものである場合，それらによって文字の一部が学習できないので誤認識が起こると考えられる．

4 回の実験で誤認識を起こす文字種は，6 つの画像データのうち，図 5 のように字体により文字の一部が異なるものや，3 つ以上の画像がインクのにじみにより文字が目視で読み取ることができないものを含む文字種である．字体により文字の一部が異なる文字種の場合，軸線の異なる字の特徴量を学習することになり，テストデータを認識することが難しくなると考えられる．また，インクのにじみにより目視で読み取ることが出来ない文字種の場合，文字の中身がデータにより異なるので認識が困難であると考えられる．

5 回の実験で誤認識を起こす文字種は，一部のみが異なる文字が存在する文字種(賴・瀨，諸・諸，輿・輿)と 6 つの画像データのうち，インクのにじみ，かすれにより文字が変形しており，図 6 のように中身が分かりにくい文字が過半数を占める文字種である．一部のみが異なる文字が存在する文字種の場合，文字線の太さにより，違いが文字画像データ上では，はっきりと表れないので誤認識が起こると考えられる．また，中身が分かりにくい文字が過半数を占める文字種の場合，文字種に共通する特徴が抽出されないことにより誤認識が起こると考えられる．

Adaboost.M1 により誤認識される文字種は，文字画像データの文字線の太さや，インクのにじみ，かすれなどの画質によるものが多いことが分かる．PDC，加重方向ヒストグラム，セル特徴をそれぞれ SVM で認識する際に起こる共通する誤認識と比べると誤認識数は減っているが，誤認識範囲は被っている[5]．これより，OLVQ1 を加えて Adaboost.M1 により認識実験を行うことにより認識率が向上するが，OLVQ1 を識別器として用いる場合，SVM を識別器として用いる場合に起きる誤認識をカバーすることができない範囲があることが分かる．したがって，認識率を向上させるためには，新たな弱識別器を用意し Adaboost.M1 によるアンサンブル学習を行うことにより認識率が向上することが考えられる．

5. まとめ

本稿では，近代書籍文字の更なる認識率向上のため，アンサンブル学習を用いた近代書籍文字認識を行っている．文字画像の特徴量として，PDC 特徴，加重方向ヒストグラム特徴，セル特徴の 3 つの各特徴量を適用し，SVM と OLVQ1 の 2 つの識別器をそれぞれ用いる．3 つの特徴量と 2 つの識別器を組み合わせることにより構成される 6 つの弱学習器を用いて Adaboost.M1 アルゴリズムに従いアンサンブル学習を行い，近代書籍文字を認識する．アンサンブル学習による認識実験の結果，Adaboost.M1 による認識結果の平均認識率は 88.88%となっている．弱学習器単体で認識するよりも認識率が向上している．誤認識の分析により，文字画像データの画質によるものが多いことが分かる．OLVQ1 を識別器として用いる場合，SVM を識別器として用いる場合に起きる誤認識をカバーすることができない範囲があるので，今後は，画質が悪い文字画像を認識できるような弱識別器を用意し Adaboost.M1 によるアンサンブル学習を行うことにより認識率が向上することが考えられる．

謝辞

本研究は科研費・基盤研究 B(No17H01829).の助成を受けたものである．

参考文献

[1]国立国会図書館全文テキスト化実証実験報告書(online)：http://www.ndl.go.jp/jp/aboutus/digitization/fulltextreport.html (参照 2017-03-20)
[2]国立国会図書館デジタルコレクション http://dl.ndl.go.jp(参照 2017-03-20)
[3]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)
[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]Kosaka,K, Fujimoto,K, Ishikawa,Y, Takata,M, and Joe,K.：Comparison of Feature Extraction Methods for Early-Modern Japanese Printed Character Recognition(PDPTA’ 16)to appear