Vol.2014-MPS-97 No.10
2014/3/3

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

粟津 妙華[1,a)]　福尾 真実[1,b)]　高田 雅美[1,c)]　城 和貴[1,d)]

概要：近代書籍の漢字のフォントは現在のものとは大きく異なる特徴を持つため，既存 OCR を用いた認識率は低い．既に我々は近代書籍に適したオフライン多フォント漢字認識手法を提案している．本論文では，1000 種の漢字に多フォント漢字認識手法を適用し，漢字の特徴と必要教師データ数の関係性を分析する．我々の認識手法では PDC 特徴を用い SVM によって漢字を認識させるが，学習時において必要となる漢字と必要教師データ数の分析を十分に行っていなかった．そこで，PDC 特徴を用いて求めた漢字の特徴値と教師データ数による認識率の変化から，漢字の特徴値と必要教師データ数の関係性を分析する．

キーワード：多フォント漢字認識，機械学習，近代書籍，PDC 特徴

Analysis of the Relation between the Number of Necessary Training Data and Each Category for Multi-Fonts Kanji Character Recognition Method

Abstract: Since the fonts of the Kanji character in early-modern Japanese printed books dating from the Meiji era to the first half of Showa era (AD1868-1945) has the greatly different feature from the standard fonts determined by JIS. We have been proposed the multi-fonts Kanji character recognition method suitable for early-modern Japanese printed books. In this paper, the multi-fonts Kanji character recognition method is applied to 1000 kinds of Kanji characters, and the relationship the between the feature of Kanji characters and the numbers of necessary training data is analyzed. In our proposed method, when Kanji characters are recognized with the PDC feature and SVM, the training data is needed. However, it was not analyzing about the relation between Kanji characters and the number of necessary training data. Therefore, the reration between the feature of Kanji characters and the number of necessary training data is analyzed using the feature value of Kanji characters by PDC vector and the change of recognition rate for each category of training data.

Keywords: Multi-Fonts Kanji Character，Machine learning，Early-Modern Japanese Book，Peripheral Direction Contribution Vector

1. はじめに

国立国会図書館関西館では，明治期から昭和前期にかけての書籍約 35 万冊を公開している．これらの近代書籍は様々な分野にわたり，また現在は絶版になっている書籍も多く存在し，学術的に非常に貴重な資料となっている．そこで国立国会図書館 [1] では，図書館資料を文化財として永く後世に伝えるとともに広く利用を供にするという目的の下，所蔵資料のデジタルアーカイブ化を行っている．近代デジタルライブラリは国立国会図書館が提供している電子図書館サービスである．近代デジタルライブラリの WEB サイトでは，タイトル・著者名の他に出版者や出版年など詳細な項目を設定して近代書籍の検索を行うことが可能である．しかしながら，近代書籍の本文は画像として公開されているため，全文検索を行うことができない．全文検索は資料の学術的価値の向上や利用者の拡大には必要不可欠

[1] 奈良女子大学
Nara, Nara, 630–8506, Japan
[a)] awazu-taeka0802@ics.nara-wu.ac.jp
[b)] fukuo-manami0834@ics.nara-wu.ac.jp
[c)] takata@ics.nara-wu.ac.jp
[d)] joe@ics.nara-wu.ac.jp

である．全文検索を行うには，画像データである現在の近代デジタルライブラリのテキスト化が必要となるが，手動による数十万冊に及ぶ書籍のテキスト化は予算的に不可能である．

このような背景のもと，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の自動テキスト化に関する研究に着手している．近代書籍をテキスト化する際，画像データに既存 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ため，我々は手書き文字認識の手法を利用することで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 [2][3]．実際，近代書籍では出版者ごとに用いる活版が異なることは当然予測されることであるが，同じ出版者であっても時代によって活版が異なることも報告されている [4]．近代書籍の活字認識に手書き文字認識の手法 [5][6][7][8] を利用するのはこのような背景があるためである．

近代書籍の自動テキスト化を行うためには，認識対象の活字も自動で切り出さなければならないが，日本語に特有のルビが大きな障害となっている．一般にルビによる文字切り出しの失敗とルビの除去の不完全さが，その後の文字認識率を劣化させることが知られている．この問題を解決するため，我々はすでに遺伝的プログラミング [9] を用いてルビを除去する手法を提案している [10]．その提案手法は書籍を出版者・時代ごとに分類し，分類ごとに遺伝的プログラミングを用いてルビ除去フィルタを生成する手法である．

次に近代書籍の多様なフォントを吸収するため，外郭方向寄与度（Peripheral Direction Contribution，PDC）特徴 [11] を用いてサポートベクターマシン（Support Vecter Machine，SVM）によって文字を認識させるオフライン手書き文字認識技術を用いて，漢字の認識実験を行う．256 種の漢字を用いての認識実験では認識率は約 92% である [3]．本論文では 1000 種の漢字を用いる．日本工業規格（JIS）[12] では，使用頻度の高い漢字 2965 文字を JIS 第 1 水準としており，通常の文書であれば JIS 第 1 水準の文字だけで記述できる．この JIS 第 1 水準の中の 1000 種の漢字を実験に用いる．認識実験を行う際，教師データが必要となるが，漢字によって教師データ数が異なることは容易に想像できることである．そこで漢字の特徴値と必要教師データ数を分析する．

本論文の構成は，以下の通りである．2 章で漢字認識の前処理としての文字切り出しとオフライン手書き漢字認識とその実験について説明し，3 章において，漢字の特徴値の提案と必要教師データ数の関連性について述べ，最後にまとめを述べる．

2. オフライン手書き文字認識

既存の OCR ソフトである e.typist[15] と読取革命 [16] を用いて，文献 [10] のルビ除去実験で用いた 4500 行を認識させた場合，ルビのある漢字の認識率は e.typist で 48.7%，読取革命で 33.1% である．ルビを除去した漢字では e.typist で 83.2%，読取革命で 80.7% の認識率である．既存 OCR ソフトは，現在広く使用されているフォントを対象としている．しかし，近代書籍は現在のフォントとは大きく異なるフォントが多く存在しているため，認識率の低下を招く．また既存 OCR ソフトにはルビ除去の機能は付属しておらず，ルビのついた書籍が多い近代書籍を認識させることは困難である．そこで本論文では，以前提案した手法 [10] によりルビを除去した行に対し，文字切り出しを行う．さらに PDC 特徴を用いたオフライン手書き文字認識を行う．

2.1 文字切り出し

日本語の文字は漢字だけでなく，ひらがなやカタカナも複数のパーツから成り立っているものが多い．**図 1** は，複数のパーツからできているひらがなである．複数のパーツで構成されるため，単純に黒画素部分の外接矩形 [17][18] や黒画素射影ヒストグラム [3] で文字を切り出すことは難しい．黒画素射影ヒストグラムを用いて文字切り出しを行った場合，1 冊の書籍から切り出せる漢字数は 400 字程度であり，非常に効率が悪い．

そこで，まず黒画素の外接矩形を求め，矩形の一部が重なっているものを統合する．次に矩形を統合してできたパーツを以下の条件を満たした場合にさらに統合する．

- 上下のパーツ間の距離が文字と文字の間の平均距離の 0.3 倍以下
- 左右のパーツ間の距離が文字の横幅平均の 0.2 倍以下
- 統合後の文字の縦幅は文字の縦幅平均の 1.2 倍以下

上記 3 条件を満たしたときとする．文字切り出しの実験では文献 [10] で用いた 3 つの出版者・3 つの時代の計 9 グループに対して，各グループに 50 冊ずつ計 450 冊を用いる．実験に用いる行は総数 4500 行であり，1 冊の書籍から 10 行ずつ切り取ったものである．これらの行に対して提案する文字切り出しを行う．文字の総数は全部で 110526 個であり，このうち切り出しに失敗したのは，2182 個，成功率はおよそ 98.0% で良好な結果である．失敗の主な原因は，上下に分かれたパーツの統合の失敗である．特に漢数字の二，三が多い．漢数字の二では，上下のパーツ間の距離が上下の文字間の距離より大きい場合も多く，切り出しに失敗している．

2.2 漢字 1000 種の認識実験

JIS 第 1 水準の漢字は 2965 種ある．通常の文書であれば，JIS 第 1 水準だけで十分書き表せると言われている．そこで，本実験では JIS 第 1 水準の漢字のうち 1000 種を

図 1　複数のパーツからなるひらがな
Fig. 1　A Hiragana with two components

図 2　複数のフォントを正しく認識した例
Fig. 2　An example of the correct recognition of different fonts

用いて認識実験を行う．実験には 2.1 節で切り出した漢字を利用する．そのため，文字種によって集まった画像の個数に大きな開きがある．文字切り出しを終えた段階では 3000 種以上の漢字があり，漢字種ごとの画像は，少ないものでは 1 枚，多いものは 50 枚以上である．文字種ごとにそれぞれ教師データとテストデータに分ける必要があるため，画像が 5 個以上のもの 1000 種総数 8448 枚を使用する．教師データ枚数を各漢字種の画像枚数の 1 割から 6 割まで変化させ，教師データの割合による認識率の違いを検証する．教師データの枚数を計算で求める際，1 以下であれば 1 とし，それ以外は四捨五入する． テストデータは画像数の 4 割とし，漢字種ごとにテストデータの枚数は変化する．1000 種の漢字のうち，各漢字種の画像数が 10 枚以下のものは 796 種（平均 8.98 個），10 枚〜50 枚のものは 181 種（平均 31.75 個），50 枚以上のものは 23 種（平均 88.18 個）である．認識率の結果を**表 1** に示す． 全体の認識率より，教師データの割合が増えると認識率も上昇する．全体の認識率に大きな影響を与えているのは，漢字種の画像数が 10 枚以下のものである．

画像数が 10 枚以下（平均 8.98 枚）の場合，教師データの割合が低いと教師データ数が非常に少なくなるため，複数のフォントの特徴の違いを吸収するのは難しく 75% と低い認識率となる．教師データを 5 割（平均 4.45 枚）以上にすると，94% の認識率となる．

画像枚数が 10 枚から 50 枚の漢字（平均 31.75 枚）では，1 割の教師データでも 93% の認識率であるが，3 割（平均 9.53 枚）以上あれば 99% の認識率となる．また，画像数が 50 枚以上（平均 88.18 枚）の場合，1 割の教師データ（平均 8.82 枚）で 99% 以上の認識率となる．実験結果から，教師データの枚数はおよそ 9 枚あれば 99% の認識率を得られることがわかる．**図 2** は，オフライン手書き文字認識手法によって複数の異なるフォントを正しく認識した例である．認識に失敗した画像は，文字がかすれたものやインクのにじみによりつぶれているものが多い．**図 3** は認識に失敗した例である．

表 1　漢字種ごとの画像数と教師データの割合における認識率 (%)
Table 1　The recognition rates with varying the rate of training data and the number of images for each Kanji character set (%)

教師データの割合	全体	漢字種ごとの画像数		
		10 枚以下	10〜50 枚	50 枚以上
1 割	89.65	75.06	93.28	99.33
2 割	94.66	87.89	96.69	99.89
3 割	96.19	91.37	99.43	99.82
4 割	96.50	92.38	99.08	99.89
5 割	97.14	94.07	99.54	99.89
6 割	97.18	94.09	99.77	99.96

図 3　認識に失敗した例
Fig. 3　A misrecognition example

3. 漢字の特徴値と必要教師データ数

2 章より，教師データの画像数はおよそ 9 枚で 99% の認識率を得られることがわかる．2 章の実験では教師データ数・テストデータ数ともに画像数の割合で決定しており，漢字ごとに変動している．そのため画像数の多い漢字では当然高い認識率を得ることになり，認識できた漢字の特徴と認識率の関連性が不明確である．また漢字によって必要教師データの枚数が異なることは容易に想像できる．当然 9 枚未満の教師データ数や既存のフォントで十分な認識率となる漢字も存在するはずである．近代書籍専用 OCR を開発するにあたって，少なくとも JIS 第 2 水準までの約 6000 種の漢字を対象としなければならず，漢字ごとに一律 9 枚の教師データを用意するのは非効率である．そこで，本章では必要教師データの枚数を算出するための漢字の特徴値を提案する．また提案した特徴値を用いた実験を行い，漢字の特徴値と必要教師データ数の関連性を検証する．

3.1 教師データのカテゴリごとの認識実験

全漢字種において教師データ枚数・テストデータ枚数ともに固定して実験を行う．実験では画像 12 枚の漢字を 1000 種用意する．さらに，現在広く使われており近代書籍のフォントに近い MS 明朝体の画像を各漢字に 1 枚用意する．テストデータは漢字ごとに近代書籍の漢字画像 3 枚とし，教師データを 1 枚から 10 枚まで変化させ認識率の推移を検証する．5 回の実験を行い，平均を求める．結果を**表 2** に示す．教師データのカテゴリごとに番号を振っておく．カテゴリ 1 の MS 明朝体 1 枚では 50% を切る認識率であるが，カテゴリ 2 の近代活字 1 枚では 67 %の認識を

表 2　教師データのカテゴリにおける認識率（%）
Table 2　The recognition rates in every category of training data (%)

教師データのカテゴリ	認識率
1: 明朝体　1 枚	49.6
2: 近代活字　1 枚	67.1
3: 明朝体 1 枚＋近代活字 1 枚	76.7
4: 明朝体 1 枚＋近代活字 2 枚	90.8
5: 明朝体 1 枚＋近代活字 3 枚	94.4
6: 明朝体 1 枚＋近代活字 4 枚	96.3
7: 明朝体 1 枚＋近代活字 5 枚	97.6
8: 明朝体 1 枚＋近代活字 6 枚	97.8
9: 明朝体 1 枚＋近代活字 7 枚	98.0
10:明朝体 1 枚＋近代活字 8 枚	98.4
11:明朝体 1 枚＋近代活字 9 枚	98.7

表 3　テストデータ 3 枚に対して認識率 100%となる漢字種数
Table 3　he number of Kanji characters recognized 100%

教師データのカテゴリ	認識率 100%の漢字種数 (1000 種中)
1: 明朝体　1 枚	306
2: 近代活字　1 枚	451
3: 明朝体 1 枚＋近代活字 1 枚	490
4: 明朝体 1 枚＋近代活字 2 枚	803
5: 明朝体 1 枚＋近代活字 3 枚	862
6: 明朝体 1 枚＋近代活字 4 枚	909
7: 明朝体 1 枚＋近代活字 5 枚	919
8: 明朝体 1 枚＋近代活字 6 枚	919
9: 明朝体 1 枚＋近代活字 7 枚	948
10:明朝体 1 枚＋近代活字 8 枚	948
11:明朝体 1 枚＋近代活字 9 枚	935

得られる．このことから，近代書籍に使用されている活字のフォントは現在のフォントとは違う特徴を持ち，それらの認識にはオフライン手書き文字認識手法が適していることがわかる．2.2 節において，およそ 9 枚の近代活字の教師データがあれば 99% の認識率が得られると推察したが，テストデータ 3 枚で固定した場合もカテゴリ 10 と 11 からおよそ 99% の認識率でほぼ同様の結果が得られた．

次に 1000 種のうちテストデータ 3 枚に対して 100% の認識率となった漢字種の数を教師データのカテゴリごとに**表 3** に示す．カテゴリ 1・2・3 では，1000 種のうち 100% の認識率となる漢字は 5 割を切る．しかし，カテゴリ 4 の MS 明朝体 1 枚と近代活字 2 枚の合計 3 枚の教師データにおいて，1000 種のうち 8 割の漢字種で認識率 100% となり，カテゴリ 6 では 9 割の漢字で認識率が 100% となる．カテゴリ 11 で認識できる漢字種が減ったが，これは増えた教師データがかすれなどにより粗雑な画像である場合，SVM における識別範囲がずれ，間違った認識結果となったものと考えられる．また，文献 [3] では 256 種の漢字を対象とし，教師データを 5 枚・テストデータを 4 枚で実験を行い認識率が 92% であるが，これはルビ除去と文字切り出しの部分で，きれいにルビを除去しきれていない画像や，文字切り出しで上下左右が多少切れているものも使用したためと考えられる．

表 2・表 3 の結果から，ともにカテゴリ 3 から 4 で大幅に改善していることがわかる．この大幅な改善と何かしらの漢字の特徴値の関連がわかれば，必要教師データ数の決定に役立つと考えられる．

3.2　漢字の特徴値

漢字の特徴値によって必要教師データ数があらかじめ決まれば，専用 OCR を開発するための時間や労力の削減が期待できる．そこで，本論文では PDC 特徴 [11] の外郭深度を用いた文字構造の複雑さを表す特徴値を提案する．一

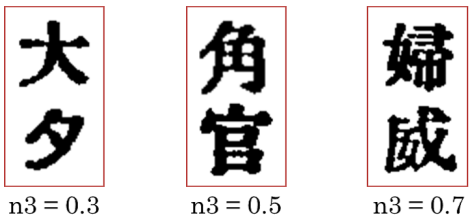

図 4　n3 値が 0.3，0.5，0.7 の近代活字
Fig. 4　The early-modern printing type that n3 value is 0.3, 0.5, 0.7

般に PDC 特徴ベクトルは，全走査方向（8 方向）∗ 深度（3 深度）∗ 寄与度成分（4 方向）∗ 区画（16 区画）= 1536 次元ベクトルで表される．このベクトルの要素のうち第 3 深度の要素に注目する．これは第 3 深度の要素が 0 でないものが多いとき，複雑な文字構造を持った漢字種であると考えられるからである．本論文では，第 3 深度の全要素のうち 0 ではない要素の割合を漢字の特徴値として提案し，認識率との関係性を検証する．この割合を n3 とすると，

$$n3 = \frac{\text{第 3 深度の要素のうち 0 でない要素数}}{\text{第 3 深度の全要素数}}$$

である．n3 値が 1 に近いほど複雑な文字構造となる．n3 値を求めるために使用した漢字は，3.1 節で使用した近代書籍の漢字画像 12 枚のうちランダムに選んだ 5 枚で，値の範囲は 0.06〜0.85 である． n3 値を求める際，外れ値は除き平均を求める．**図 4** は n3 値が 0.3，0.5，0.7 の近代活字である．

3.3　漢字の特徴値と認識実験

テストデータ 3 枚に対し 100% の認識率となった漢字種の平均 n3 値を求める．求めるのは MS 明朝体 1 枚の時・近代活字 1 枚の時と，それぞれの教師データのカテゴリで

表 4　教師データのカテゴリにおける近代活字の平均 n3 値
Table 4　Average n3 value in every category of training data

教師データのカテゴリ	平均 n3 値
1: 明朝体　1 枚	0.53
2: 近代活字　1 枚	0.57
3: 明朝体 1 枚＋近代活字 1 枚	0.59
4: 明朝体 1 枚＋近代活字 2 枚	0.69
5: 明朝体 1 枚＋近代活字 3 枚	0.69
6: 明朝体 1 枚＋近代活字 4 枚	0.67
7: 明朝体 1 枚＋近代活字 5 枚	0.68

新たに認識できるようになった漢字種の n3 値の平均である．結果を**表 4** に示す．ただし，MS 明朝体 1 枚と近代活字 7 枚を教師データにしたカテゴリ 9 以降はあまり差異がないため，表は省略する．

カテゴリ 3 から 4 で n3 値が大きく増加している．これは，3.1 節の表 2・表 3 の結果にも合致しており，n3 値が 0.7 前後の複雑な構造を持つ漢字を認識させるには，少なくともカテゴリ 4 の MS 明朝体 1 枚と近代活字 2 枚が必要であるということを示している．この結果から n3 値が必要教師データの画像数決定のおおよその指標になりうることがわかる．

ここで，MS 明朝体と近代活字の n3 値を比較する．近代活字の n3 値は 3.2 節で求めた近代活字 5 枚の平均である．現在のフォントである MS 明朝体の n3 値（0〜0.88）を指標にできれば，さらなる効率化に繋がる．比較は近代活字の平均 n3 値から MS 明朝体の n3 値を引いた差異を求めることで行う．1000 種の漢字における差異の平均は＋ 0.009 である．しかし，MS 明朝体の n3 値を基準とする場合，n3 値が 0〜0.64 では差異は－の値（近代活字の n3 値の方が大きい），0.64〜0.88 までは＋の値（MS 明朝体の n3 値の方が大きい）が非常に多い．0〜0.64 までの差異と 0.64〜0.88 までの差異を以下に示す．

$$\text{MS 明朝体の n3 値} = \begin{cases} 0\text{〜}0.64：\text{-}0.04 \\ 0.64\text{〜}0.88：+0.03 \end{cases}$$

表 4 から，近代活字の n3 値が 0.7 前後の漢字種には少なくとも MS 明朝体 1 枚と近代活字 2 枚が必要であるとわかったが，これを MS 明朝体の n3 値に当てはめた場合，0.7〜0.75 の範囲となると考えられる．

次に認識できなかった漢字についてである．教師データの 11 カテゴリ全てで認識できない漢字種は 1000 種中 32 種である．これらの n3 値は 0.65〜0.75 の範囲にある．このことから，全ての漢字を n3 値のみで確実に必要教師データの枚数を決定できるわけではないことがわかる．1000 種中 968 種は PDC 特徴の深度 3 で十分な認識を得る特徴を抽出できることから，大半の漢字は n3 値で必要教師データ数を算出できる．しかし，少数ではあるが深度 3 では十分な特徴を抽出できず，認識できない漢字がある．表 2・表 3 のカテゴリ 4 以降，数値の上昇が鈍化しているのは，本実験で用いた PDC 特徴の外郭深度 3 では全ての漢字の複雑さを吸収しきれていないためであると考えられる．そのため，必要教師データ数算出のためのより正確な指標を求め，さらに多くの漢字を認識させるためには，PDC 特徴の外郭深度を深める必要があると考えられる．

今回の実験結果では，n3 値はおおよその必要教師データを算出するための指標になりうることがわかった．また MS 明朝体による n3 値が 0.7 より大きければ，教師データは少なくとも MS 明朝体 1 枚と近代活字 2 枚が必要であり，それらを用意できれば 80 %以上の漢字でほぼ 100 %の認識率を得られることがわかる．このことから，MS 明朝体の n3 値を用いることで，より効率の良い近代書籍専用 OCR の開発が期待できる．

4. まとめ

本論文では，その後 PDC 特徴を用いた SVM による漢字 1000 種の認識実験を行った．さらに漢字の特徴値と必要教師データの画像数の関係性を検証した．JIS 第 1 水準 2965 文字中 1000 種の漢字で，およそ 9 枚の教師データがあれば 99% の認識率となると推測することができる．さらに，漢字の特徴値と必要教師データの画像数の関連を検証するため，PDC 特徴ベクトルを用いた新たな特徴値を提案し認識実験を行った．その結果より，提案した特徴値と必要教師データ数に間に関連性があることがわかる．

漢字の認識実験では，1000 種での実験を行った．結果は漢字種ごとの画像数が 10 枚以下では 5 割を教師データとすることで 94% の認識率となり，10 枚〜50 枚では 3 割，50 枚以上では 1 割を教師データ数とすることで 99% の認識率を得られ，良好な結果と言える．ここから，教師データはおよそ 9 枚あれば，多くの漢字種で非常に高い認識率となると推察することができる．しかし，実験では教師データ・テストデータとも変動であり，画像数が多い漢字種では高い認識率となる．また漢字種によって必要教師データ数が違うことが予想されたため，教師データ・テストデータとも固定し実験を行い，漢字の特徴値と必要教師データ数の関連性を検証した．関連性の検証のため，PDC 特徴を用いた新たな漢字の特徴値を提案し，教師データをカテゴリごとに分け認識実験を行った結果，提案した特徴値と必要教師データ数との間に関連性が見いだすことができる．この結果から，提案した漢字の特徴値を用いることで，おおよその必要教師データ数を決定することができると考えられる．

今後は複雑な文字構造の漢字のより正確な指標とより正確な認識のために， PDC 特徴の深度を深める必要がある．さらに文字種を増やすことも課題である．そのために，文字切り出しの精度を上げなければならない．また前後の

ページのインクがヒストグラムの差異では取れないほど強く染みこみ，表面の文字と重なっているものが存在する．これを裏抜けといい，これは文字切り出しにも文字認識にも障害となる．この問題を解決するために，近代書籍に特化した裏抜け除去手法の開発が必要であると考える．

参考文献

[1] 国立国会図書館 (online): *http://www.ndl.go.jp/* (accessed 2014-2-3).
[2] 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 The 2009 International Conference on Parallel and Distributed Processing Technologies and Applications (PDPTA 2009)*, Vol. II, pp. 728–734 (2009).
[3] Fukuo, M., Enomoto, Y., Yoshii, N., Takata, M., Kimesawa, T., and Joe, K.: Evaluation 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 (PDPTA 2011)*, Vol. II, pp. 727–732 (2011).
[4] 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, Vol. 2012-MPS-90, No. 26 (2012).
[5] 篠沢佳久, 大駒誠一: テンプレートマッチングによるオフライン手書き文字認識ニューラルネットワークの作成, 情報処理学会論文誌, Vol. 42, No. 1, pp. 16-25 (2001).
[6] 鶴岡 信治: 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, 電子情報通信学会論文誌 D 情報・システム, Vol. 7, No. 7, pp. 1390-1397 (1987).
[7] 浜本義彦, 政水克典, 内村俊二, 富田眞吾: 手書き漢字認識のための Gabor 特徴, 電子情報通信学会論文誌. D-II, 情報・システム, II-情報処理, Vol. J79-D-2, No. 2, pp. 202-209 (1996).
[8] 後藤 英昭, 平山理継, 阿曽 弘具: 局所多値しきい値処理による濃淡文書画像からの文字パターンの抽出, 電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理, Vol. J82-D-2, No. 11, pp. 2188-2192 (1999).
[9] 伊庭斎志: 遺伝的プログラミング入門, 東京大学出版会 (2001).
[10] 粟津妙華, 福尾真実, 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, Vol. 6, No. 2, pp. 53-62 (2013).
[11] 萩田紀博, 内藤誠一郎, 増田 功: 外郭方向寄与度特徴による手書き漢字の識別, 電子通信学会論文誌 D, Vol. 66, No. 10, pp. 1185-1192 (1983).
[12] 日本工業規格 (online): *http://www.jisc.go.jp/* (accessed 2014-2-3).
[13] Cortes, Corinna. and Vapnik, Vladimir N.: Support-Vector Networks, Machine Learning, Vol. 20, pp. 273-297 (1995).
[14] T. Kohonen, G. Barna, and R. Chrisley: Statistical pattern recognition with neural networks : benchmarking studies, IEEE International Conference on Neural Networks, Vol. 1, pp.61-68 (1988)
[15] e.Typist v14.0 (online): *http://mediadrive.jp/products/et/index.html* (accessed 2014-2-3).
[16] 読取革命 (online) *http://kindai.ndl.go.jp/* (accessed 2014-2-3).
[17] 馬場口登, 塚本正敏, 相原恒博: 手書き日本文字列からの文字切り出しの基礎的考察, 電子通信学会論文誌.(D), Vol. J68-D, No. 12, pp. 2123–2131 (1985).
[18] 長谷博行, 辻正博, 園田浩一郎, 米田正明, 酒井充: 汎用を目指した自動文書画像認識システム : 要素抽出技術の問題点と検討, 電子情報通信学会技術研究報告. PRU, パターン認識・理解, Vol. 94, No. 242, pp. 49–56 (1994).