Vol.2014-MPS-98 No.20
Vol.2014-BIO-38 No.20
2014/6/27

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

粟津 妙華[1,a)]　高田 雅美[1,b)]　城 和貴[1,c)]

概要：国立国会図書館では，所蔵する明治から昭和前期の近代書籍を近代デジタルライブラリとして Web 上でページごとの画像データとして公開しているが, 文書内容での検索を行うことができない．そのため，自動でのテキストデータ化が望まれている．その際，問題となっているのがヒストグラムでは除去できないルビであり，我々はすでに近代書籍に特化したルビ除去手法を提案している．しかしながら，提案した手法は書籍に付加された外部情報を元にしており，実現可能性は低い．そこで本論文では，書籍画像から直接得られるデータを元に，進化計算によってルビ除去式を生成し，近代書籍から自動でルビを除去する手法を提案する．

キーワード：ルビ除去，近代書籍，遺伝的プログラミング，活字のアスペクト比

Ruby Removal Filters by Genetic Programming using the classification of printing type data for Early-Modern Japanese Printed Books

Abstract: In National Diet Library, books which are possessed in library as "the digital library from meiji era" are open to the public on Web. Since these are shown as image data and cannot search using document contents, an automatic text conversion is needed. There is a major obstacle to text conversion. It is ruby. Ruby can not be removed in the histogram method. Therefore, we have proposed a ruby removal method for early-modern Japanese printed books. However, since the proposed method is based on the external information added to the books, the feasibility is low. In this paper, we propose a method to remove the ruby automatically from early-modern Japanese printed books by generating ruby removal formula in Genetic Programming using the training data was based on the data of book image.

Keywords: Ruby Remove, Early-Modern Printed Books, Genetic Programming, Aspect ratio of the print

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 34 万冊を公開している．これらの近代書籍は，哲学・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書籍も多く，学術的に貴重な資料である．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い，近代デジタルライブラリとして電子図書館サービスを提供している．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．近代書籍は学術的に貴重なものを多く含むとはいえ，数十万冊に及ぶ書籍の手動による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

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

関する研究 [2] 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 [3][4]．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 [5]．近代書籍の活字認識に手書き文字認識の手法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，一般にルビによる文字切り出しの失敗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 [6]．特に近代書籍では，現在の書籍のように決まった規格はないため，既存のルビ除去技術を適用したのでは，肝心の文字認識率が大幅に低下してしまう．この問題を解決するため，我々は近代書籍に特化したルビ除去手法を開発している [7]．これは，遺伝的プログラミングを用い，出版者・時代ごとの専用ルビ除去式を生成する手法である．しかしながら，近代書籍の出版者は個人出版のものも多く含まれ，数は 1 万を超える．出版者・時代ごとに手動で教師データを収集することは非常に困難である．そこで，本論文では，分類方法を見直し，活字画像から得られるデータを用いて近代書籍を分類する．そして，その分類を元に遺伝的プログラミングを用いてルビ除去式を生成し，自動でルビを除去する手法を提案する．

本論文の構成は，以下の通りである．

2 章において本論文で用いた遺伝的プログラミングによるルビ除去式の生成について説明する．[7] で用いたアルゴリズムと同じであり，その概要を述べる．3 章において [7] で提案した分類によるルビ除去手法とその問題点について述べ，4 章において活字データを用いた近代書籍の分類とそれを元にしたルビ除去式の生成について説明する．5 章において，提案手法の有効性を調べる実験について述べる．活字データによる分類を用いた提案手法と，同様に自動でルビを除去する判別分析法を用いた黒画素射影ヒストグラム法の結果を比較し，考察を行う．

2. 遺伝的プログラミングを用いたルビ除去

本論文で提案するルビ除去手法のアルゴリズムは，[7] で用いたものと同じであり，その概要を示す．遺伝的プログラミングを用い，行における親文字とルビの境の近似式を自動生成する．始めに，教師データである各行から文字の位置情報などを推定し，それらの値を遺伝的プログラミングの終端要素として与え，ルビ除去式を生成する．除去式を適用後，残ったルビの一部を除去するために，孤立点除去を行う．**図 1** は，本論文で用いた手法のフローチャートである．概要は以下の通りである．

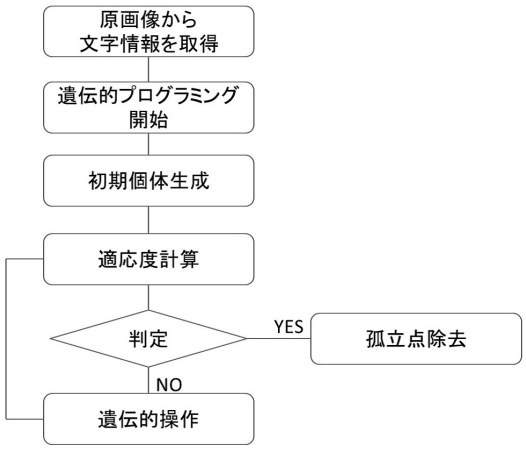

図 1 提案手法のフロー
Fig. 1 Flow of the proposed method

(1) 教師データの原画像である各行からルビ付き文字列の座標位置と文字の横幅の推定
(2) 手順（1）の値を与え，遺伝的プログラミングを用い除去式を生成
  (a) 初期個体群の生成
  (b) 手順（1）で求めた位置情報と横幅を終端要素として与え，適応度を計算
  (c) 終了条件の確認
  (d) ルーレット選択で，個体群の半数を交叉
  (e) ランダム選択で選んだ個体を突然変異
  (f) 適応度の計算
  (g) 適応度の低い個体を削除，新たに個体を生成
  (h) 手順（2c）に戻る
(3) 生成式で除去後，メディアンフィルタを適用し，残ったルビの一部に対し孤立点除去を行う

手順（1）では，教師データを読み込み，原画像であるルビのある行から，ルビ付き文字列の位置と文字の横幅の推定を行う．

手順（2）では，手順（1）で求めた値を与え，遺伝的プログラミングを用い除去式を生成する．この際，非終端要素には，四則演算子と絶対値，三角関数 sin・cos を用いる．終端要素には 1〜9 の定数と π，手順（1）で求めた文字の横幅・それぞれのルビ付き文字列の縦方向の座標位置が入る．

手順（2a）では，初期個体を生成する．個体は終端要素・非終端要素を用い，木構造で表現された式である．これを指定された個体数生成する．

手順（2b）では，適応度の計算を行う．適応度は，生成式で表された曲線の右側の黒画素部分を原画像から削除した画像と目標画像の輝度値の一致率とする．

手順（2c）で用いる終了条件は，適応度が 1 になるか，指定世代数だけ実行することである．

手順（2d）では，ルーレット選択で交叉させる親個体を

Vol.2014-MPS-98 No.20
Vol.2014-BIO-38 No.20
2014/6/27

選び交叉させる．

手順（2e）では，ランダム選択で選んだ個体を突然変異させる．

手順（2f）では，遺伝的操作で作成された次世代の適応度を手順（2b）と同じ方法で計算する．

手順（2g）では，適応度の低い個体を半数削除する．

以上の操作を，終了条件が満たされるまで繰り返す．

手順（3）では，除去できずにわずかに残ったルビに対し孤立点除去を行う．

3. 出版者・時代ごとの分類とその問題点

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はない．現在の規格に沿ったルビは，ルビとルビの付いている文字（親文字）の間に一定の間隔が空いている．しかし，近代書籍によく見られるルビは，親文字とルビの近接度が高く，連結しているものも多い．連結している場合，親文字とルビのそれぞれの特徴値を求めることが困難であるため，サポートベクターマシンなどの機械学習では，ルビだけを分離することはできない．また，黒画素射影ヒストグラムを用いた場合も，近接度の高さが原因でルビの除去率は高くない．この問題を解決するために，我々はすでに近代書籍に特化したルビ除去手法を提案している [7]．

近代書籍は活版印刷であるため，活版の数だけフォントが存在し，それらは出版者や時代によって特徴が異なることは容易に想像できる．そこで，この手法では，出版者・時代ごとの専用ルビ除去式を生成する．教師データとして，出版者・時代ごとに分類した近代書籍の行を用い，遺伝的プログラミングによって出版者・時代ごとの専用ルビ除去式を生成する．

出版者・時代という分類を用いたルビ除去の実験では，およそ 98%前後の除去率となっている．例えば「春陽堂・明治中期」で生成された式ではルビ除去率は 99.0%である．また，判別分析法を用いた黒画素射影ヒストグラム法では，およそ 83%の除去率であり，提案した手法の有効性は示されている．

しかしながら，近代書籍の出版者の数は膨大である．近代デジタルライブラリで公開されている書籍の出版者だけでも 1 万を超える．出版者・時代という書籍に付加された外部情報を用いて，手動で分類することは難しく，この分類方法の実現可能性は低いと言わざるを得ない．そのため，近代書籍の自動テキスト化には，活字画像から直接得られるデータを元に自動で分類し，ルビ除去式を生成することが必要である．

4. 活字データによる分類

出版者・時代という外部情報を元に手動で分類するのではなく，画像から直接得られるデータによって自動で分類し，それらを教師データとして遺伝的プログラミングを用いてルビの除去式を生成する．

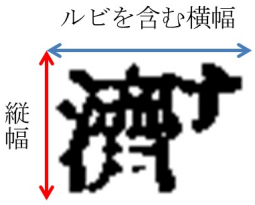

図 2 アスペクト比に用いる縦幅とルビを含む横幅
Fig. 2 Width with ruby and height for aspect ratio

4.1 活字のアスペクト比

本論文で提案する分類方法は，活字のアスペクト比を用いるものである．[7] では，フォントは出版者・時代ごとに特徴があると仮定したが，実際のフォントは，活版の数，つまり印刷所の数だけ存在していると考えられる．これは出版者・時代という括り以上に，手動での分類は不可能である．そのため，実際に印字された活字データから特徴値を求めることが最善であると考える．そこで，行ごとのルビを含む活字の平均アスペクト比を用いて分類し，遺伝的プログラミングによってルビ除去式を生成する．

活字のアスペクト比を求めるために，まず行の中の親文字の縦幅とルビを含む横幅を算出する．初めに，行の縦方向に黒画素射影ヒストグラムを取り，谷の部分で分離し，その縦幅の平均を求める．その際，求めた縦幅が，実際の縦幅と大きな差が出ることがある．インクのにじみなどにより親文字が上下で連結した文字は，その他の文字の縦幅よりも大幅に大きくなり，漢数字の「二」「三」のように小さく分離されてしまう文字や句読点は，その他の文字よりも非常に小さくなる．そのため，平均値を求める際には，上記の実際の縦幅の値と大きく異なる縦幅の値をを省く必要がある．これにより，実際の縦幅と平均値の差異が小さくなることが期待される．省く値は，一旦全ての縦幅の値から平均を求め，その平均値から大きく離れた縦幅の値とする．次にルビを含む横幅の平均を求める．縦幅の平均を求める際に省いた文字と，縦幅の 1.2 倍以下の横幅となる文字は除き，横幅の平均を算出する．つまり，ルビのついている親文字を対象として，その縦幅とルビを含む横幅の比を求める．アスペクト比を f とすると，

$$f = \frac{\text{ルビを含む横幅}}{\text{縦幅}}$$

と表される．**図 2** は，アスペクト比に用いる縦幅とルビを含む横幅を示している．

4.2 アスペクト比によるルビ除去式の生成

[7] で用いた，近代デジタルライブラリで公開されている近代書籍 900 冊を対象にアスペクト比 f を求めたところ，およそ 1.4 から 1.8 の間となり，大半は 1.5 から 1.7 である．

Vol.2014-MPS-98 No.20
Vol.2014-BIO-38 No.20
2014/6/27

表 1　各グループにおける目標画像の輝度値との一致率 (%)
Table 1　The best agreement rates for each class

f	目標画像との一致率
[-:1.4]	99.022
[1.35:1.45]	99.018
[1.4:1.5]	99.021
[1.45:1.55]	99.021
[1.5:1.6]	99.013
[1.55:1.65]	99.022
[1.6:1.7]	99.001
[1.65:1.75]	99.212
[1.7:1.8]	99.207
[1.75:1.85]	99.018
[1.8:-]	99.022

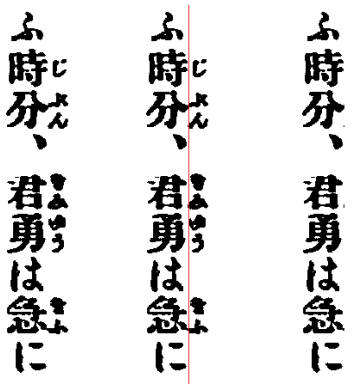

図 3　f 値が 1.65 以下の式を適用した行の一部
Fig. 3　Part of the line to which is applied a formula of 1.65 or less

そこで，f の値が 1.4 以下，1.35-1.45，1.4-1.5，1.45-1.55，1.5-1.6，1.55-1.65，1.6-1.7，1.65-1.75，1.7-1.8，1.75-1.85，1.8 以上の 11 に分類する．1.4 以下と 1.35-1.45，1.7-1.8，1.75-1.85，1.8 以上のグループは，教師データが 100 行集まらなかったが，50 行以上あるので，それを教師データとする．それ以外のグループは教師データ 100 行である．実験はグループごとに 10 回行い，各グループでの生成式によるルビ除去後と目標画像の輝度値の最も良い一致率を求める．結果を**表 1** に示す．

全てのグループで 99%以上の輝度値の一致率となっている．生成された式を精査したところ，1.4 以下と 1.35-1.45，1.4-1.5，1.45-1.55，1.5-1.6，1.55-1.65，1.6-1.7 の各グループで生成された式は，全て同じ式となる．つまり 1.65 以下は同一の式でルビを除去できる．生成された式を以下に示す．

$$y = \text{縦幅の平均値} + 6$$

図 3 は上記の除去式を適用し，ルビを除去したものである．f の値が 1.65-1.75，1.7-1.8，1.75-1.85，1.8 以上のグループで生成された式は，それぞれ異なっている．例として，

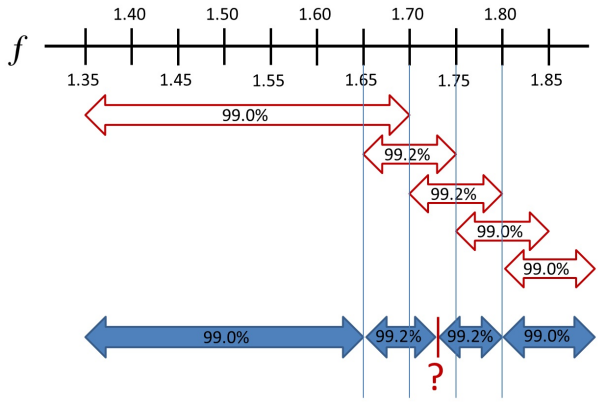

図 4　f 値の範囲と目標画像との輝度値の一致率
Fig. 4　The best agreement rates and range of value f

1.7-1.8 で生成された式は以下の通りである．

$$\begin{aligned} y = {} & (8 - ((\text{縦幅の平均値}/8) - (\text{縦幅の平均値} - \\ & (5/(\cos((2 * \pi * x/(((8 * 5)/2)) - \pi/2)) - \\ & (\text{縦幅の平均値} - ((4/(6 * \cos((2 * \pi * x/ \\ & ((\cos((2 * \pi * x/(((8 * 5)/2)) - \pi/2))/2)) \\ & -\pi/2))))/\text{縦幅の平均値})))))) \end{aligned}$$

この結果より，f 値が

- 1.65 以下
- 1.65-1.75
- 1.7-1.8
- 1.8 以上

の 4 つのグループで求まった除去式を用いることで，出版者・時代というグループ分けよりも良好な結果を得られることがわかる．**図 4** は，f 値の範囲と目標画像との輝度値の一致率である．

1.7-1.75 の範囲では，1.65-1.75 と 1.7-1.8 で生成されたどちらの除去式を用いても，およそ 99.2%の目標画像との一致率となっている．そこで，どの範囲でどちらの式を用いるのが適切か検証する．1.7-1.75 の行を 0.01 刻みで 15 行ずつ用意し，1.65-1.75 と 1.7-1.8 の除去式を適用する．f 値を 0.01 刻みにした場合，使用した教師データでは 15 行ずつしか用意できないため，15 行で実験を行う．

結果，1.7-1.71 では 1.65-1.75 の除去式を適用すると，目標画像との一致率は 99.193%であったが，1.7-1.8 の除去式を適用すると，99.016%となる．そのほかの範囲でも同様に，両方の除去式を適用したところ，1.7-1.71 では 1.65-1.75 の除去式，1.71-1.75 では 1.7-1.8 の除去式を適用した方が一致率が高くなる．

これより，f の値を 1.65 以下，1.65-1.71，1.71-1.8，1.8 以上に分け，求めた除去式を適用することで高いルビ除去率となることがわかる．この分類方法は，読み込んだ行の画像から直接得られた特徴値を用いることから，出版者・時代という外部からの付加情報を用い手動で分類する方法

表 2 f 値による手法とヒストグラムによる手法のルビ除去率 (%)
Table 2 Removal success rate of the histogram and the proposal method

f 値	提案手法	ヒストグラム
[-:1.65]	96.2	74.3
[1.65:1.71]	95.5	82.4
[1.71:1.8]	93.9	89.1
[1.8:-]	91.8	90.4

に比べ，効率良くルビを除去することができる．

5. 有効性の検証

提案の分類方法を用いて生成されたルビ除去式の有効性を調べるため，生成式を用いてルビ除去の実験を行う．

5.1 実験条件

本研究は，近代書籍全般を対象としたものであるが，近代書籍の数は膨大であり，それら全てを実験対象とすることは困難である．そこで，近代デジタルライブラリで公開されている約 34 万冊を対象として実験を行う．まず，これらの書籍の中でルビの付いている書籍の数を調べる．

近代デジタルライブラリでは，書籍を分野ごとに分けている．例えば，哲学（8153 冊），歴史（32826 冊），社会科学（105861 冊），言語（11125 冊），文学（36698 冊）など全部で 10 に分類されている．それぞれの分野ごとに 1%の書籍をランダムに選び，ルビの有無を調べる．結果，ルビのある割合が 5%以下の分野が 4 つ，およそ 10%前後の分野が 2 つである．そのほか総記と言語で約 20%，哲学でおよそ 50%，文学でおよそ 80%の書籍にルビが存在する．この結果から，近代デジタルライブラリの書籍の中でルビのある書籍の総数は 64304 冊と推定する．

このルビのある書籍の中から信頼度 95%，誤差 5%で標本数を求めると，およそ 1537 冊となる．1537 冊を分野ごとの割合に分けると，哲学で 148 冊，文学で 166 冊などとなる．分野ごとにルビのある書籍を所定数ランダムに選び出し，その書籍から 1 行切り出しルビ除去の実験を行う．1 冊の書籍では，当然同じフォントを使用しており，ルビの特徴も同じであるため，今回の実験ではルビが複数ついている行を 1 行切り出す．実験では，行は分野に関係なく f の値だけで分類し実験を行い，さらに判別分析法を用いた黒画素射影ヒストグラムによる手法との比較も行う．除去の成否は目視である．

5.2 結果

f の値が 1.65 以下の行は 853 行，1.65-1.71 は 245 行，1.71-1.8 は 366 行，1.8 以上は 73 行である．ルビ除去の結果を**表 2** に示す．

全ての f 値グループで提案手法の除去率は 90%以上の除去率となり，1.65 以下と 1.65-1.71 では，95%を超えており，非常に良好な結果である．それに対し，f 値が 1.8 以上のグループでは，提案手法とヒストグラム法を比較した場合，除去できた行数は 1 本しか異ならないという結果となっている．f 値が大きいということは，ルビを含めた横幅が縦幅に比べ，非常に大きいことを意味する．活版印刷の場合，ルビは通常親文字の活字の縦横半分の大きさの活字であることが多いが，活字の文字の周りの余白部分の大きさが活版によって異なり，これが f 値の違いとなる．余白が大きければ，4 章で求めた活字の縦幅は小さく，またルビを含めた横幅は余白分だけ大きくなり，結果 f 値は大きくなる．つまり，f 値が大きいものは，親文字・ルビともに余白が大きく，親文字とルビの近接度が低いものが多いと考えられる．そのため，判別分析法を用いた黒画素射影ヒストグラムによる除去法とほとんど差のない結果となっていると考えられる．

この結果から，f 値が 1.8 以下であれば，提案手法は非常に有効な手法であると言える．また f 値が 1.8 以上でも，ヒストグラムを用いた手法とほとんど変わらないが，90%以上の除去率となり良好といえる．出版者・時代という分類による除去式の生成では，98%前後の除去率であったが，これは書籍に付加された外部情報を元に手動で分類せねばならず，実現不可能である．それに対し，提案手法は活字画像から直接データを求め，それを元に分類するため自動で行うことができる．同じように自動で除去できる判別分析法を用いた黒画素射影ヒストグラム法より良好な結果となっており，提案手法は近代書籍のルビ除去に有効であるといえる．

6. まとめ

本論文では，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手法を提案した．本手法を用いることにより，現在の書籍を対象としたルビ除去手法には適さない近代書籍において，ルビを自動で除去することができ，近代書籍の自動テキスト化が進むことが期待される．

提案手法では，画像データからルビのついている活字の縦幅とルビを含む横幅を求める．そして，それら用いたアスペクト比を元に分類し，遺伝的プログラミングを用いてルビ除去式を生成する．結果，教師データから求まった 4 つの除去式によって，目標画像との一致率は 99%を超えている．これらの式を使い，近代デジタルライブラリで公開されている書籍画像から自動でルビを除去する実験を行ったところ，同様に自動でルビを除去する判別分析法を用いた黒画素射影ヒストグラム法より良好な結果を得られた．本手法は，[7] の手法とは異なり，読み込んだ活字データを使い自動で分類するため，近代書籍のためのルビ除去手法として，非常に有効である．

今後は，レイアウト解析が重要であると考える．近代書

籍を既存のレイアウト解析手法によって解析した場合，精度は非常に低く実用性に乏しい．近代デジタルライブラリで公開されている自然科学や技術などの分野では，文章と図や表などが複雑に配置されているページも多く，近代書籍の自動テキスト化のためには，文章や図・表をそれぞれ正確に切り出さなければならない．この問題を解決するために，近代書籍に特化したレイアウト解析技術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 2014-05-28).
[2] 城和貴, 高田雅美: 近代デジタルライブラリの自動テキスト化，科研基盤研究 (C)，21500237 (2009-2011).
[3] 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).
[4] 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).
[5] 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).
[6] 曹宇, 佐藤匡正: 文字寸法の違いに着目した OCR 認字率の改善法, 電子情報通信学会技術研究報告. SS, ソフトウェアサイエンス Vol. 100, No. 678, pp. 17–22 (2001).
[7] 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, Vol. 6, No. 2, pp. 53–62 (2013).