情報処理学会論文誌　データベース　Vol.6　No.2　53–62 (Aug. 2013)

# 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

粟津 妙華[1,a)]　高田 雅美[1,b)]　城 和貴[1,c)]

受付日 2012年11月8日，再受付日 2012年12月18日，
採録日 2013年1月27日

**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ルビである．現在のルビを直線的に除去する技術は，規格に沿った現在の書籍を対象としたものであるため，現在の書籍とは違う特性を持つ近代書籍には適用できない．そこで，本研究では，遺伝的プログラミングを用いて，曲線的に出版者・時代ごとの専用ルビ除去式の生成を行う．

**キーワード**：ルビ除去，近代書籍，遺伝的プログラミング，文字切り出し，テキスト化，ヒストグラム，文字認識

# Ruby Removal Filters Using Genetic Programming for Early-modern Japanese Printed Books

Taeka Awazu[1,a)]　Masami Takata[1,b)]　Kazuki Joe[1,c)]

Received: November 8, 2012, Revised: December 18, 2012,
Accepted: January 27, 2013

***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. However, ruby is a disturbing text conversion. Since existing techniques of linearly removing ruby had developed for books of the current standard, the techniques are inapplicable to early-modern Japanese books, which have a specific characteristic different from characters of current books. In this paper, we propose a method to remove ruby from early-modern Japanese books using Genetic Programming.

***Keywords:*** ruby remove, early-modern printed books, genetic programming, character segmentation, transforming text, histogram, recognition of characters

## 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57万冊を公開している．これらの近代書籍は，哲学・自然科学・文学など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に供するという目的のもと，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・著者名のほか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テキスト化は予算的

[1] 奈良女子大学
Nara Women’s University, 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

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 [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う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．我々が知る限りでは，近代書籍に特化したルビ除去は研究されていない．

ルビ除去の既存手法として，濃度ヒストグラムを用い直線的に分離する方法 [4], [7] や，外接矩形を用いて分離する方法 [8], [9] などが報告されている．しかし，親文字とルビが連結している部分が多いと，ヒストグラムの谷の部分が明瞭に出ない場合があり，良好な結果が得られないことがある．また，外接矩形を用いる方法では，親文字とルビが連結し 1 つの矩形になると除去することは困難である．そこで，本論文では，近代書籍に特化したルビ除去手法を提案する．近代書籍のルビは出版者・時代によって，それぞれ似た特性を持つという仮定のもとで，出版者・時代ごとに近代書籍を分類し，特定の出版者における特定の時代専用のルビ除去フィルタを，遺伝的プログラミング [10] を利用して曲線的に求める．

本論文の構成は，以下のとおりである．

2 章において現在の書籍と近代書籍におけるルビの特徴について説明し，3 章において既存の文字切り出しの研究と，ルビ除去への適用について述べる．4 章において遺伝的プログラミングによる曲線的なルビの分離方法を述べる．5 章において，提案手法の有効性を調べるための実験について述べる．提案手法とヒストグラムによる除去との結果を比較し，考察を行う．

## 2. 現在の書籍と近代書籍におけるルビの特徴

ルビとは，文章内の任意の文字に対し，ふりがな・説明・異なる読み方といった役割の文字を，より小さな文字で記されるものである．ルビをつける場合，その対象となる文字のことを親文字という．縦書きの日本語文書では，通常右側にルビをつける．

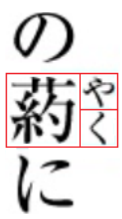



図 1　現在の規格によるルビの配置と文字のサイズ
**Fig. 1**　Current arrangement and size of ruby.

図 2　活字を組み合わせた活版
**Fig. 2**　Typography combined with printing type.

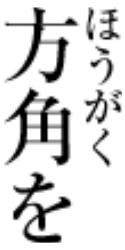



図 3　現在の書籍におけるルビ
**Fig. 3**　The ruby in the present books.

現在の書籍は，DTP によるデジタル製版であり，一般的に組版は JIS などによって規格が決まっている．この規格におけるルビを振る方法として，モノルビ・熟語ルビ・グループルビの 3 つが知られており，各方法によって，ルビの配置位置が異なる．しかし，どの方法においてもルビの親文字に対する配置位置は，親文字の外枠右側にルビ文字の外枠が接するように配置されている．ルビ文字のサイズは，親文字の 1/2 である．稀に三分ルビと呼ばれる親文字に 3 字付ける方法もあるが，使用例は少ない．通常は，親文字が 1 文字・ルビ 3 文字の場合，ルビの一部は親文字の前後の文字にかかる．**図 1** は，現在の規格に沿ったルビの配置と文字のサイズの例である．

近代書籍は，明治期から昭和前期における書籍で，活版印刷である．活版印刷とは，活字を組み合わせた活版による印刷技術である．**図 2** は，活字を組み合わせた活版である．近代書籍には，現在の書籍のように決まった規格はなく，現在の書籍とは違う特有の特性を持っている．近代書籍に共通する特性として，親文字とルビの近接度が非常に高いという点があげられる．この特性に加え，インクのにじみなどにより，親文字とルビが連結している部分も多く見られる．また，活版印刷で用いる活版そのものが粗雑なものがあるため，ルビ部分が歪んでいることもある．**図 3**

図 4　近代書籍におけるルビ
**Fig. 4**　Ruby of the early-modern printed books.

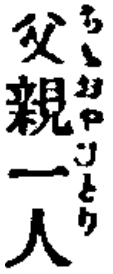


図 5　ルビの行が歪んでいる例
**Fig. 5**　An example of distorted lines.

は，現在の書籍におけるルビの例である．**図 4** は，近代書籍によく見られるルビの例である．**図 5** は，ルビの行が歪んでいる例である．

現在のルビは，親文字とルビの間に一定の間隔があり，行ごとに直線的に分離することができる．しかし，近代書籍におけるルビは，親文字とルビの間隔が狭く，またルビの歪みがあるため，直線的に分離することは難しいと考えられる．

## 3.　文字切り出しにおける既存研究

近代書籍におけるルビ除去についての研究はあまり多くない．そこで本章では，日本語文書における文字切り出しの既存研究から，ルビ除去に適用できると考えられる手法を説明する．

### 3.1　黒画素射影ヒストグラムによる文字切り出し手法

黒画素射影ヒストグラムによる文字切り出し手法では，まず黒画素部分にラベリング処理を行い，黒画素の射影ヒストグラムを算出する．次に黒画素射影ヒストグラムに平滑化処理を行い，ヒストグラムの谷の部分を切断位置に設定することで，接触・続け字の切り出しを行う．**図 6** の黒画素射影ヒストグラムは，文字が離れており，切り出すことができる例である．**図 7** は，つづけ字で前後の文字が連結しており，文字を切り出すことが困難な例である．

この手法の問題点は，すべての文字画像に対して適用可能な最適なパラメータの決定が困難なことである．黒画素射影ヒストグラムの平滑化幅が小さい場合，ヒストグラムの変化に対して過敏に処理することになるため，切り出しミスは減少するものの，漢字の部首によっては，わずかに

図 6　黒画素射影ヒストグラム
**Fig. 6**　Projection histogram by black pixels.

図 7　つづけ字の黒画素射影ヒストグラム
**Fig. 7**　Black pixel projection histogram of connected characters.

図 8　左：小さな矩形に分割された文字
右：1 つの矩形に統合した文字
**Fig. 8**　Left: A character divided by small rectangles, Right: A character unified to a rectangle.

存在する白画素部分で，上下に分割されてしまうという問題が発生する．一方，平滑化幅が大きい場合，黒画素射影ヒストグラムの変化を大局的にとらえるため，個別文字の過剰な切断を抑制できる反面，接触・入込み文字の切り出しミスが生じることになる．つまり，黒画素の分布が書き手に大きく依存するため，一定の制約条件によって連結・入込み文字の正確な切り出し位置を決定することは困難である．また，漢数字の「二」「三」などは，分割されてしまうため，1 つの文字として切り出すことは困難である．

この手法をルビ除去に用いる場合，行の横方向に黒画素射影ヒストグラムをとり，谷になっている部分で，直線的に親文字とルビを分離する．その結果，親文字部分が切れてしまうことがある．親文字の欠損は，文字認識の大きな障害となる．

### 3.2　外接矩形を用いた文字切り出し手法

外接矩形を用いた文字切り出し手法では，まず縦・横・斜めの 8 方向に連結した黒画素部分にラベリング処理を行い，外接矩形を求める．次に近接度の非常に高い小さな矩形を統合し，複数の矩形に分割されることが多い日本語の文字を 1 つの矩形とし，文字切り出しを行う．**図 8** は，小さな矩形に分割された文字と，それを 1 つの矩形に統合した文字である．

図 9　3 つの矩形に分割された漢数字
**Fig. 9**　A character divided by three rectangles.

図 10　親文字とルビが連結している場合の矩形
**Fig. 10**　A dividing rectangle includes the ruby.

この手法の問題点は，連結した文字の切り出しが困難なことである．漢字どうしの連結の場合，文字の縦幅はほぼ一定であるという仮定のもと，パラメータを決め，切り出すことができるが，漢字とひらがなやその他の文字の場合，文字幅が違うため，正確に文字を切り出すことは困難である．また，黒画素射影ヒストグラムを用いた手法と同様に，漢数字の「二」「三」などは分割されてしまうという問題点がある．**図 9** は，分割された漢数字の「三」である．

この手法を用いたルビ除去では，親文字とルビが連結している場合，大きな矩形として認識されるため，ルビを除去することはできない．**図 10** は，親文字とルビが連結している場合の矩形である．

## 4.　曲線によるルビ分離

本論文では，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．はじめに，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．

### 4.1　アルゴリズム

提案手法のフローチャートを**図 11** に示す．詳細は，以下のとおりである．

(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を推定
(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

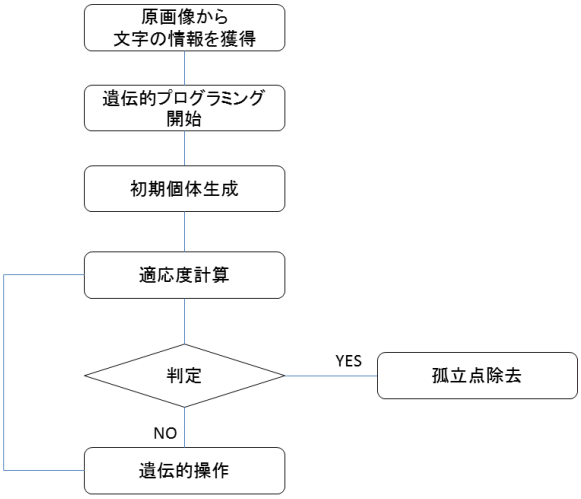



図 11　提案手法のフロー
**Fig. 11**　Flow of the proposed method.

(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．教師データは，原画像としてルビのある行と，目標画像としてルビを削除した行で構成されており，原画像・目標画像とも，二値化された画像とする．

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．初めに，原画像から最も左端の黒画素の座標を求め，そこから縦方向の直線を $x$ 軸とする．次に行に複数あるルビ付き文字列のそれぞれの上端を横方向にとった直線を，それぞれのルビ付き文字列の $y$ 軸とする．生成式は，$y =$ (終端要素を用いた曲線式) となり，式を適用するのは，ルビ付き文字列の部分だけである．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 $\pi$，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．文字の横幅は，行ごとに決まった値，それぞれのルビ付き文字列の縦方向の座標位置は $x$ で表し，ルビ付き文字列の上端を $x = 0$ とした変数である．**図 12** は，終端要素として与える変数 $x$ を示したものである．近代書籍では，ルビが長く，親文字の前後の文字にかかる場合も多く見られる．提案手法では，ルビとルビの左側の文字を 1 つの文字群とし，その文字群の上端位置を変数 $x$ における $x = 0$ の位置とする．ルビの左側に文字がない場合は，ルビだけで 1 つの文字群である．**図 13** は，ルビが親文字の前の文字にかかっている場合の $x = 0$ の位置である．

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．個体数は $N$ とする．

手順 (2b) では，適応度の計算を行う．適応度は，生成

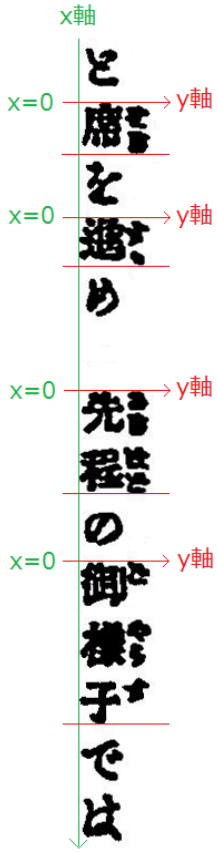



図 12　遺伝的プログラミングの終端要素として与える変数 $x$
**Fig. 12**　Variable x as termination element for GP.

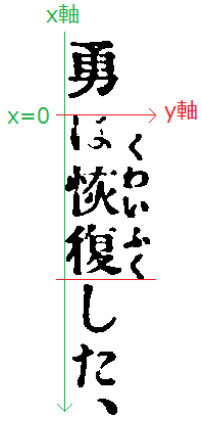



図 13　ルビが親文字の前の文字にかかる場合の変数 $x$
**Fig. 13**　Variable x in a case of ruby in front of a parent character.

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．その際，白画素部分は操作の対象ではないため，適応度の計算範囲は，行の全範囲ではなく，原画像のルビ付き文字列の半分より右側の黒画素の位置とし，生成式による除去後と目標画像の輝度値の一致率を適応度とする．図 14 は，適応度計算を行う原画像の範囲である．赤で囲ってある部分の中で，黒画素である位置が適応度計算を行う原画像の位置となる．1 行の中には複数のルビ付き文字列があり，その個数

図 14　適応度計算を行う原画像の範囲
**Fig. 14**　Range of the original image for fitness calculation.

を $K$ とし，1 行の中のルビ付き文字列を表す変数を $a$ とする．原画像のルビ付き文字列の縦の画素数を $X_a$，原画像のルビ付き文字列の $y = (1/2) * $ 文字の横幅の直線より右側の部分の横の画素数を $Y_a$ とし，$X_a * Y_a$ で表される領域を $S_a$ とする．このとき，領域 $S_a$ 内の原画像の輝度値を $o_{a(x,y)}$，領域 $S_a$ 内の生成式によって出力された画像の輝度値を $c_{a(x,y)}$，領域 $S_a$ 内の目標画像の輝度値を $t_{a(x,y)}$ とする．$x$，$y$ は，領域 $S_a$ 内の縦横の座標を表す変数である．これらを用い，$B_a(o_{a(x,y)})$，$E_a(o_{a(x,y)}, c_{a(x,y)}, t_{a(x,y)})$ を以下の式 (1)，式 (2) により定義する．

$$B_a(o_{a(x,y)}) = \begin{cases} 1 & (o_{a(x,y)} = 0) \\ 0 & (o_{a(x,y)} \neq 0) \end{cases} \tag{1}$$

$$E_a(o_{a(x,y)}, c_{a(x,y)}, t_{a(x,y)}) = \begin{cases} 1 & ( (o_{a(x,y)} = 0) \cap (c_{a(x,y)} = t_{a(x,y)}) ) \\ 0 & ( \overline{(o_{a(x,y)} = 0)} \cup \overline{(c_{a(x,y)} = t_{a(x,y)})} ) \end{cases} \tag{2}$$

個体 $i$ の適応度を $f_i$ とすると，$f_i$ は式 (3) で表される．

$$f_i = \frac{1}{K} \sum_{a=1}^{K} \sum_{x=0}^{X_a} \sum_{y=0}^{Y_a} \frac{E_a(o_{a(x,y)}, c_{a(x,y)}, t_{a(x,y)})}{B_a(o_{a(x,y)})} \tag{3}$$

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選び交叉させる．個体 $i$ を選ぶ確率 $p_i$ は，式 (4) により決定する．

$$p_i = \frac{f_i}{\sum_{k=1}^{N} f_k} \tag{4}$$

ルーレット選択は，残す遺伝子個体を選ぶときに，ある程度の適応度を持つものからランダムに選ぶため，個体の多様性が維持される．エリート保存選択では，多様性を失い局所的な最適解に収束する傾向があり，またランダム選択では個体の進化が進みにくくなるため，ルーレット選択で親個体を選ぶ．選んだそれぞれの親個体からランダムに 1 点を選び，その位置から下の部分の木構造を取り替えることで交叉を行う．

図 15　孤立点
**Fig. 15**　Isolated points.

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，残ったルビに対し孤立点除去を行う．生成式によるルビ除去で残った黒画素部分は小さいが，文字の誤認識を防ぐために必要である．縦・横・斜めの 8 方向に連結した黒画素部分にラベリング処理を行い，面積が極端に小さい部分を除去する．除去する際の閾値は，面積が 10 以下とする．**図 15** は，孤立点を示した画像である．

### 4.2　ルビ付き文字部分の位置と文字の横幅の推定

遺伝的プログラミングにおける終端要素とするため，ルビ付き文字列の上端と下端の位置と，文字の横幅の推定を行う．遺伝的プログラミングにおいて，文字の横幅は行ごとの決まった定数，文字列の縦方向の座標位置は変数として与えられる．

文字の縦横の比率はおよそ 1：1 であると仮定し，各文字の縦幅を求め，その平均値を文字の横幅値とする．初めに，行の縦方向に黒画素射影ヒストグラムをとり，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い．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，その他の文字の縦幅よりも大幅に小さい．ひらがなの「い」「つ」「へ」などは他の文字に比べ縦幅が小さい．繰返し符号の「同の字点」の縦幅は，かなり小さな値となり，現在ではあまり使われないが，近代書籍では散見される繰返し符号の「くの字点」の縦幅は，かなり大きな値となる．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，いったんすべ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．省いた後，残った縦幅の値から，もう一度平均値を求め，その値を文字の横幅とする．

次にルビ付き文字部分の位置を推定する．ルビ付き文字部分は，ない部分に比べ，ルビの横幅の分だけ大きい．そこで，以下の式 (5) の基準でルビがあると推定し，それぞれのルビ付き文字の上端と下端の位置情報を保持する．

$$\text{ルビのある文字の横幅値} \geq \left(1+\frac{1}{4}\right) \times \text{文字の横幅値} \tag{5}$$

次に，ルビ付き文字が連続している場合は，それらを連結し，連続した文字列の上端と下端の位置情報を保持する．

遺伝的プログラミングに与える終端要素は，それぞれの連続したルビ付き文字列の上端位置を $x=0$ とした，縦方向の変数である．これは，1 文字ずつ切り分けて，ルビを除去するのではなく，インクのにじみなどによって，親文字が上下で連結している文字列にも対応できる除去式を生成するためである．近代書籍はにじみによる連結が多いため，連結した文字がない現在の書籍のように 1 文字を対象とするのではなく，複数個の親文字で構成される文字列を対象とする必要がある．

## 5.　実験

提案手法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

画像は，二値化した PGM 画像を用いる．教師データは，原画像としてルビのある行を，目標画像として原画像からルビを削除した行を用いる．

教師データは，それぞれの出版者・時代ごとに分類する．出版者は，春陽堂・日吉堂・駸々堂の 3 つ，時代は，明治中期（1883～1897）・明治後期（1898～1912）・大正（1912～1925）の 3 つである．各分類に対して，教師データを用意する．教師データとする行を，10 行・50 行・100 行・200 行・300 行・400 行と変化させ，教師データの個数による結果の違いを確認したところ，100 行以降は教師データ数を増加させても，結果に大きな差は見られなかった．そこで，教師データは 100 行とし，1 冊につき 10 行を 10 冊，計 100 行を使用する．

実験における遺伝的プログラミングのパラメータは，個体数，世代数の上限，交叉確率，突然変異確率がある．個体数は，1,000 から 5,000 まで 1,000 刻みで変化させた結果，3,000 個体以降は個体数を増加させても，結果に大きな差は見られないため，3,000 個体で固定とする．その他のパラメータは，世代数の上限 200，交叉確率 0.8，突然変異確率 0.2 と固定とする．

教師データ以外のサンプルにおいても，除去式が有効であるか検証するため，それぞれの出版者・時代において，教師データで用いた行とは異なる 300 行を用意し，求めた除去式を適用する．また，提案手法が有効であるか検証す

表 1　10 回中の曲線と直線の出現回数，適応度の平均値と最大値
**Table 1**　The number of appearances of curves and straight lines, average and the maximum values of fitness in 10 times.

| | | 曲線 | | | 直線 | | |
|---|---|---|---|---|---|---|---|
| | | 出現回数 | 平均適応度 | 最高適応度 | 出現回数 | 平均適応度 | 最高適応度 |
| 春陽堂 | 明治中期 | 7 | 0.9878 | 0.9881 | 3 | 0.9870 | 0.9874 |
| | 明治後期 | 8 | 0.9896 | 0.9893 | 2 | 0.9869 | 0.9876 |
| | 大正 | 9 | 0.9875 | 0.9887 | 1 | 0.9874 | 0.9874 |
| 日吉堂 | 明治中期 | 7 | 0.9752 | 0.9797 | 3 | 0.9757 | 0.9785 |
| | 明治後期 | 3 | 0.9822 | 0.9845 | 7 | 0.9836 | 0.9845 |
| | 大正 | 10 | 0.9751 | 0.9753 | - | - | - |
| 駸々堂 | 明治中期 | 7 | 0.9843 | 0.9849 | 3 | 0.9838 | 0.9846 |
| | 明治後期 | 9 | 0.9857 | 0.9857 | 1 | 0.9851 | 0.9851 |
| | 大正 | 9 | 0.9848 | 0.9842 | 1 | 0.9830 | 0.9830 |

るため，文字切り出しにおける黒画素射影ヒストグラムをルビ除去に適用した場合と比較する．

### 5.2　結果

実行時間は，Intel Xeon Processor，メモリ 8 GB の環境で，およそ 3 日間必要であった．

それぞれの出版者・時代ごとに 10 回の実験を行い，生成式が曲線もしくは直線となる回数，それぞれにおける平均適応度，最高適応度を**表 1** に示す．日吉堂の明治中期・明治後期以外は，すべてで曲線における平均適応度・最高適応度の方が高くなっている．日吉堂の明治中期・明治後期における平均適応度は直線の方が高いが，最高適応度は，明治中期では曲線の方が高く，明治後期では曲線・直線は同じ値である．これは，2 つの分類においての教師データとして用いた行の中に，他の分類に比べて親文字とルビの近接度が低い行が含まれていたため，直線式が生成されやすく，平均適応度が高くなったと考えられる．しかし，現在の書籍のようにすべての行において親文字とルビに一定の間隔があるわけではなく，近接度が親文字とルビの組合せによって異なることが多い．最もルビがとれる式は，曲線となっている．そのため，生成された曲線式のうち 91.3%に周期関数である sin・cos が含まれている．また，それぞれの分類における最高適応度を示す生成式はすべて sin・cos が含まれている．これは，ルビがどの出版者・時代においても親文字に対してランダムな位置に配置されているのではなく，それぞれ何かしらの決まりのもと，周期的な位置に配置されているため，周期関数が多く含まれると考えられる．

遺伝的プログラミングで生成された最高適応度の式を適用し，孤立点除去を行った場合の目標画像との輝度値の一致率を**表 2** に示す．一致率を計算する際の画像の範囲は，4.1 節の手順 (2b) と同じである．表 2 より，すべてで一致率が 99%を超えていることが分かる．特に，日吉堂の明治中期と大正時代では，孤立点除去の効果が高い．この 2 つの分類では，他の分類に比べ，文字が縦長であるものが

表 2　出版者・時代ごとの目標画像との一致率（%）
**Table 2**　The coincidence rate by publisher and era.

| | 明治中期 | 明治後期 | 大正 |
|---|---|---|---|
| 春陽堂 | 99.67 | 99.64 | 99.32 |
| 日吉堂 | 99.33 | 99.60 | 99.54 |
| 駸々堂 | 99.67 | 99.77 | 99.75 |

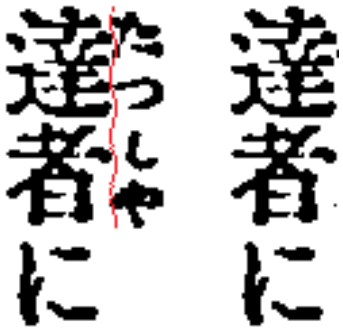


図 16　式 (6) で表される曲線とルビ除去後
**Fig. 16**　The curve denoted by (6) and the result.

含まれているため，文字の横幅の推定において，実際の横幅よりも大きい値となってしまい，ルビ部分が残ったと考えられる．しかし，孤立点除去を行うことで，良好な結果が得られる．式 (6) は，春陽堂・明治中期において生成された式の一例である．式 (7) は，日吉堂・明治中期において生成された式の一例である．**図 16** は，式 (6) の除去式で表される曲線とルビ除去後の画像，**図 17** は，式 (7) の除去式で表される曲線とルビ除去後の画像である．式中の $x$ は，それぞれのルビ付き文字列の上端を $x = 0$ とする縦方向の変数である．$y$ は，行全体において最も左端の黒画素の座標を，$y = 0$ とした横方向の座標位置である．

$$y = ((8/3) + ((\text{文字の横幅} - (\cos((2 * \pi * x/(((4 - (\cos((2 * \pi * x/((\sin((2 * \pi * x/(((5 + 3)/2)) - \pi))/2)) - \pi/2))/1)))/2)) - \pi/2))/(8/3))) - (\cos((2 * \pi * x/(((\text{文字の横幅} + 4)/2)) - \pi/2))/(7/5)))) \quad (6)$$

図 17　式 (7) で表される曲線とルビ除去後
**Fig. 17**　The curve denoted by (7) and result.

表 3　既存手法と提案手法の比較：除去成功率（%）
**Table 3**　Removal success rate of the existing and the proposal method.

| | | ヒストグラム | ヒストグラム判別分析法 | 提案手法 |
|---|---|---|---|---|
| 春陽堂 | 明治中期 | 82.3 | 79.0 | 99.0 |
| | 明治後期 | 92.7 | 81.7 | 99.3 |
| | 大正 | 90.7 | 62.7 | 96.7 |
| 日吉堂 | 明治中期 | 84.3 | 76.7 | 97.3 |
| | 明治後期 | 86.0 | 82.0 | 99.3 |
| | 大正 | 95.7 | 88.3 | 99.0 |
| 駸々堂 | 明治中期 | 96.3 | 93.3 | 99.0 |
| | 明治後期 | 93.3 | 91.7 | 99.0 |
| | 大正 | 94.3 | 91.0 | 98.7 |

$$y = ((\text{文字の横幅} - \cos((2 * \pi * x/(((x * (\cos((2 * \pi * x /(((1 * (x * (((8 + 7)/((5 * ((\text{文字の横幅} * |(6 + (\text{文字の横幅}))|)/(\text{文字の横幅})) * 8)))/2)) -\pi/2)) * 8))/2)) - \pi/2))) \tag{7}$$

次に，文字切り出しの既存手法である黒画素射影ヒストグラムをルビ除去に適用した場合と提案手法を比較する．それぞれの出版者・時代において，教師データ以外の 300 行で比較する．ヒストグラムの閾値を推定する際，2 通りの方法で行う．1 つ目は，閾値を 10 から 200 まで 10 刻みで変化させ，最もルビがとれた閾値の結果とする．2 つ目は，判別分析法を用いて，閾値を自動で求めた結果である．**表 3** は，それぞれの手法における除去成功率である．表 3 より，すべての場合で既存手法に比べ提案手法の方が良好な結果が得られることが分かる．判別分析法の除去成功率が最も低い．これは，ヒストグラムにおいて大きな山が 2 つ以上ある場合，適切な閾値が得られないことが原因であると考えられる．

提案手法を用いてルビ除去が正確にできなかった画像を，**図 18** に示す．これは行の途中で傾き方が変化していることが失敗の原因であると考えられる．

**図 19** は，本研究で対象としている近代デジタルライブラリで公開されているデジタルデータであり，ページを開いた状態で上から撮った写真のデータである．図 19 から

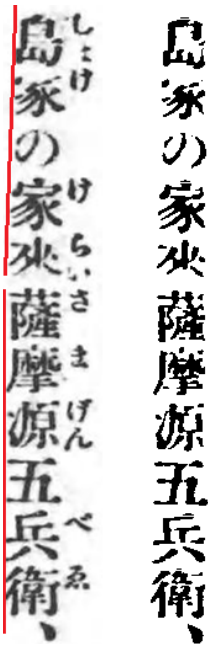


図 18　ルビ除去に失敗した例
**Fig. 18**　Example of ruby removal failure.

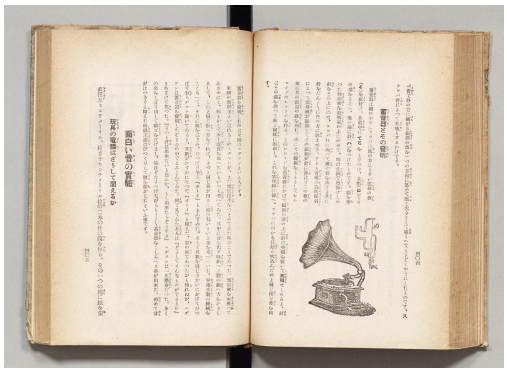
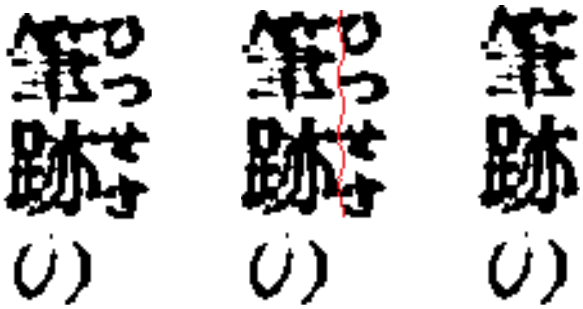


図 19　近代デジタルライブラリーで公開されているデジタルデータ
**Fig. 19**　Digital data of Digital Library from the Meiji Era.

分かるように，近代デジタルライブラリで公開されているデジタルデータは，写真データであり，ページを開いた書籍を上から撮影したものであるため，中央付近の書籍を綴じた部分でページがたわんでいる．また，書籍の上下でたわみ方も異なる．そのため，傾き方が行の上下で異なっていると考えられる．これを解決するには，近代デジタルライブラリにおける書籍の写真の撮り方を修正する必要がある．

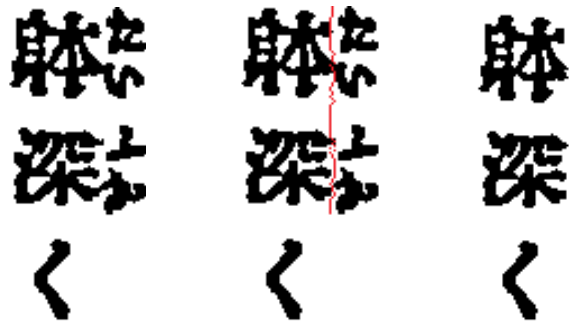
1 行の中に親文字とルビが連結している部分が多い場合，既存の黒画素射影ヒストグラムや外接矩形を用いる方法では，良好な結果を得ることはできない．**図 20** (a) は親文字とルビが連結している原画像，(b) は式 (6) を表示した画像，(c) はルビを除去した画像である．**図 21** (a) は親文字とルビが連結している原画像，(b) は式 (7) を表示した画像，(c) はルビを除去した画像である．図 20 と図 21 より，提案手法では，親文字とルビの連結に関係なく除去す



（a）原画像　（b）生成式を表示した原画像　（c）ルビ除去後の画像

図 20　式 (6) を適用した原画像とルビ除去後の画像

**Fig. 20**　The original image and the result by applying (6) for ruby removal.



（a）原画像　（b）生成式を表示した原画像　（c）ルビ除去後の画像

図 21　式 (7) を適用した原画像とルビ除去後の画像

**Fig. 21**　The original image and the result by applying (7) for ruby removal.

ることができることが分かる．これは，ルビ除去式の生成において，親文字とルビの近接度の情報を用いていないからである．

本実験では，すべての分類において既存の黒画素射影ヒストグラムより良好な結果を得ることができる．遺伝的プログラミングで生成された式で除去しきれなかった部分も，孤立点除去を行うことで，除去成功率は上がる．ゆえに，近代書籍からのルビ除去において，遺伝的プログラミングを用い，曲線的にルビ除去を行う本手法は，有効である．

## 6.　まとめ

本論文では，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の手法を提案した．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，近代書籍を出版者・時代ごとに分け，遺伝的プログラミングを用いて，それぞれにおける専用の除去式を生成する．遺伝的プログラミングを用い，式を木構造で表し，100 行の教師データから自動で除去式を生成し，除去しきれず残った部分は孤立点除去を行う．教師データを用いた場合の目標画像との一致率は，すべての分類において，99%を超えている．現在の書籍のための文字切り出しを改良したルビ除去との性能を比較するため，出版者・時代ごとに 300 行を用意し除去式を適用し実験を行った．比較実験には，黒画素射影ヒストグラムを用い，閾値決定法には，2 種類の方法を用いた．黒画素射影ヒストグラムを用いた手法に比べ，除去成功率は上がっており，提案手法は良好な結果を得ている．

提案手法は，傾きがない行を対象とした方法である．そのため，大きな傾きのある行では適用できない．しかし，国立国会図書館で公開されている写真データは，ページが傾いているものが多い．そこで，今後の課題としてルビ除去を行い，テキスト化を進めるためには，近代書籍の特性を考慮した傾きを補正する手法の開発が必要である．

**謝辞**　実験用のデータを提供していただいた国立国会図書館関西館電子図書館課に感謝します．本研究の一部は科研基盤研究（C）21500237 による．

### 参考文献

[1] 国立国会図書館 (online), 入手先 ⟨http://www.ndl.go.jp/⟩（参照 2012-11-8）.
[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, *Proc. 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, *Proc. 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] 秋山照雄，内藤誠一郎，増田　功：非接触文字優先切出しによる印刷物からの文字切出し法，電子通信学会論文誌（D），Vol.J67-D, No.10, pp.1194–1201 (1984).
[8] 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).
[9] 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).
[10] 伊庭斎志：遺伝的プログラミング入門，東京大学出版会 (2001).

**粟津 妙華**　（学生会員）

2012 年奈良女子大学理学部情報科学科卒業．2013 年同大学大学院人間文化研究科情報科学科修士課程修了．修士（理学）を同大学より取得．2013 年同大学院人間文化研究科複合現象科学専攻博士後期課程進学，現在に至る．パターン認識に関する研究に従事．

**高田 雅美**　（正会員）

2004 年奈良女子大学大学院人間文化研究科複合領域科学専攻修了．博士（理学）を同大学より取得．2004 年独立行政法人 JST 戦略的創造研究推進事業において，京都大学大学院情報学研究科にて委嘱研究員．2006 年奈良女子大学大学院人間文化研究科助手．2007 年奈良女子大学大学院人間文化研究科助教．2013 年奈良女子大学理学部講師．数値計算ライブラリの開発，分散メモリ環境を対象とする並列プログラムの開発に関する研究に従事．

**城 和貴**　（正会員）

大阪大学理学部数学科卒業．日本 DEC，ATR 視聴覚研究所（日本 DEC より出向），（株）クボタ・コンピュータ事業推進室で勤務の後，1993 年奈良先端科学技術大学院大学情報科学研究科博士前期課程入学，1996 年同研究科後期課程修了，同年同研究科助手．1997 年和歌山大学システム工学部講師，1998 年同助教授．1999 年奈良女子大学理学部情報科学科教授，現在に至る，博士（工学博士）．情報処理学会論文誌数理モデル化と応用編集委員長．