Vol.2015-MPS-105 No.10
2015/9/29

近代書籍用 OCR のための 学習用特定フォントセットの自動生成手法

岩田彩† 上坂和美† 粟津妙華† 石川由羽† 高田雅美† 城和貴†

本稿では，近代書籍で用いられているフォントを用いた活字を自動で生成する手法を提案する．近代デジタルライブラリーで一般公開されている近代書籍のテキスト化に使用する多フォント活字認識手法の精度向上のため，学習データを十分に増やす必要がある．しかし，近代書籍に使用されているフォントは多種多様であるため，十分な学習データを収集することは困難である．そこで本稿では，学習用の近代書籍フォントセットを自動生成する変換フィルタを，遺伝的プログラミングを用いて生成する．

Automatic Training Multi-Font Set Generation Method For Early-Modern Japanese Printed Character Recognition

AYA IWATA† KAZUMI KOSAKA† TAEKA AWAZU†
YU ISHIKAWA† MASAMI TAKATA† JOE KAZUKI†

In this paper, we present an automatic training data generation method for automatic text extraction for “digital library from Meiji Era”. To Improve the accuracy of multi- font type recognition method that is used in the text of Early-modern Japanease printed books that are vieable at the public in digital library from meiji era Web site,it is necessary to increase the learning data suficiately. Because the font used in the books is a wide variety ,it is difficult to collect enough training data .So, we generate a conversion filter that automatically generating a early-modern book font set for learning using a genetic programming and shows the effectiveness of the font set.

1. はじめに

国立国会図書館[1]は，明治期から昭和初期にかけて刊行された近代書籍を近代デジタルライブラリー[2]として平成 14 年から Web 上で一般公開している．近代デジタルライブラリーでは，国立国会図書館が所蔵する書物のうち，著作権保護期間が終了したものや著作者の承諾を得たものからデジタル化し，公開している．公開されている書物は，現在絶版になっているものもあり，分野は哲学から産業，文学，芸術等多岐にわたるため当時の文化や風俗を知る上で学術的に貴重な資料である．インターネット上での一般公開は，国立国会図書館が所有する資料を図書館に足を運ぶことなく大勢の人が閲覧できるようにし，希少価値のある図書資料の利用性を高めることを目的としている．現在公開されている書籍の数は，図書およそ 35 万点，雑誌およそ 5 千点である．

近代デジタルライブラリーでは近代書籍をページごとにスキャンし，画像データとして一般公開している．近代デジタルライブラリーの Web サイトでは，タイトル・著作者の他，出版年や日本十進分類法（Nippon Decimal Classification，NDC）など詳細な項目を指定して蔵書検索を行うことが可能である．しかし，目次の一部分がテキストデータ化された書籍はあるものの，本文は未だテキスト化されておらず，書籍内容の文書検索ができない．そこで，貴重な資料をより手軽に便利に利用するため，早急に書籍内容をテキスト化し，検索可能にすることが望まれる．ただし，近代デジタルライブラリーで公開している数は，近代書籍 35 万冊，雑誌 5000 冊と膨大であるため，手作業によるテキスト化では効率が悪い．

そこで我々は，出版者や出版年代によって文字の形状が異なる近代書籍の活字に対応可能な近代書籍に特化した光学文字認識（Optional Character Recognition，OCR）を用いて，自動的に画像からテキスト化を行う多フォント活字認識手法[3][4][5]の研究を進めている．この手法は，外郭方向寄与度（Peripheral Direction Contributivity，PDC）特徴[6]を用いて文字画像から特徴を抽出し，特徴ベクトル化を行う．そして，特徴ベクトルに対して SVM（Support Vector Machine）を用いて学習を行い，文字を分類する．この多フォント活字認識の精度向上のためには，SVM に必要な学習データ数を増やす必要がある．

近代書籍には，出版者や時代ごとにおよそ 2 万種類のフォントが存在する．文字認識のための学習データ数は JIS コードの観点から，1 つのフォントにおいて最低でもおよそ 6000 種類の文字が必要である．さらに，文字の認識率を上げるためには，1 種類の文字につき最低 5 個の学習データがあることが望ましい[7].

学習データ収集支援のための Web アプリケーションを開発し，学習データの数を増加させている[8]．しかし，近代書籍から収集できる文字の種類や個数には限界がある．つまり，学習データとなる文字を近代書籍から収集するだ

†奈良女子大学
Nara Women’s University

けでは，十分な学習データを収集することは難しいといえる．それゆえ，近代書籍文字認識システムのための十分な学習データセットを近代書籍の冊数や文字数に依存せず，効率よく作成させるための手法が必要である．フォントセットを自動生成することで，文字収集にかかるコストパフォーマンスを改善し，多種類の文字からなる学習データを入手することが可能になる．

第 2 章では近代書籍に特化した文字認識システムについて詳しく述べる．第 3 章では近代書籍用 OCR のためのフォントセット自動生成手法について提案する．第 4 章では生成したフォントと近代書籍に使用されている文字の PDC 特徴の類似度を比較し，その有効性について検証する．

2. 近代書籍文字認識システム

2.1 多フォント活字認識手法

近代書籍の自動テキスト化には，近代書籍に特化した OCR である多フォント活字認識手法が用いられることが望ましい．近代書籍は，旧字体や異字体を多く含み，出版者や出版された時代により異なる種類のフォントが使用されており，その種類はおよそ 2 万にも及ぶ．そこで，このように多様な種類の活字からなる書籍にも対応可能な近代書籍の活字認識に特化した OCR を用いる．

多フォント活字認識の流れを以下に示す．

I. 前処理
  (i) 2 値化
  (ii) ノイズ除去
  (iii) 角度補正
  (iv) ルビ除去
II. 文字切り出し
III. PDC 特徴の抽出
IV. SVM を用いた識別

手順 I の前処理では，2 値化や角度修正，ノイズ除去，ルビ除去を行う．はじめに，PDC 特徴抽出のために書籍画像を 2 値化する．次にメディアンフィルタを施し，ノイズを除去する．ノイズ除去により，ノイズと文字領域の誤認知を防ぐことができる．その後，書籍ページのたわみや文書のずれを補正するため，角度補正を行う．ルビは文字切り出しの失敗原因の 1 つであり認識の妨げになるため，ルビ除去を行う[9][10]．近代書籍のルビは親文字との距離が非常に近いため，遺伝的プログラミング（Genetic Programing，GP）を用いて行ごとに任意の式を生成する．その式によりルビを除去する．

手順 II の文字切り出しでは，前処理を終えた書籍画像から文字を 1 文字ずつ切り出す．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施し，求めた外接矩

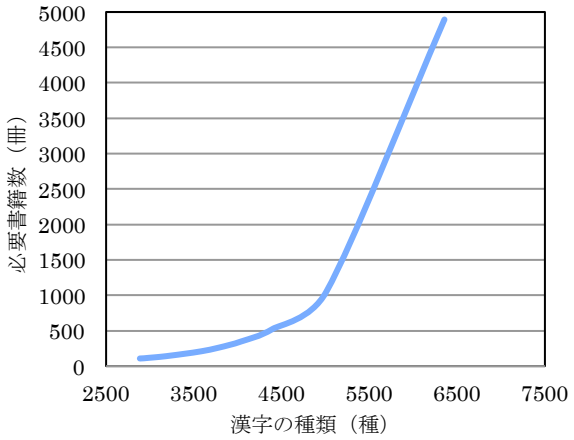

図 1 漢字の出現と書籍数の関係

表 1 出版者別公開書籍数

出版者	公開書籍数（冊）
春陽堂	1918
新潮社	880
大倉書店	496
日吉堂	456

形を用いて切り出しを行う．

手順 III の PDC 特徴の抽出では，切り出した文字ごとの PDC 特徴の抽出を行い，特徴ベクトルを計算する．PDC 特徴とは，パターン整合性に基づいた特徴であり，文字線構造情報を反映する．文字線構造情報とは，文字線の複雑さ，方向，接続関係，相対位置関係の 4 種類の情報から成り立つ．文字線の複雑さは文字線の本数を，文字線の方向と接続関係性は方向寄与度を用いて抽出する．文字線の相対関係は，文字の外郭形状を用いて表すことが可能である．方向寄与度は，文字線内の各黒点を 4 次元ベクトルで表す．黒点Pの方向寄与度d_Pを$d_P = (d_{1P}, d_{2P}, d_{3P}, d_{4P})$で表す．

各要素$d_{mP}(m = 1,2,3,4)$は，点 P から縦・横・斜めの 8 方向へ触手を伸ばして求まる黒点連結長$l_i(i = 1,2,\dots,8)$を用いて，

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

で定義される．外郭形状は，文字パターンを 45° おきに 8 方向から走査した場合に横切る文字線の 1 本目から n 本目までの外郭点をプロットしてできる形状を第n外郭点と呼ぶ．外郭深度$n = 3$まで取る．これにより，文字線のほぼ全ての輪郭点を含むこととなる．PDC 特徴の各要素を$P_{umn}(k)(k = 1,2,\dots,16)$とおくと，PDC 特徴ベクトル$P_N$は次のように$514\times N$次元ベクトルで表される．

$$P_N = (\quad P_{111}(1), P_{111}(2), \dots, P_{111}(16), P_{112}(1), \dots, P_{11N}(16), P_{211}(1), \dots, P_{tmn}(k), \dots, P_{84N}(16) \quad)$$

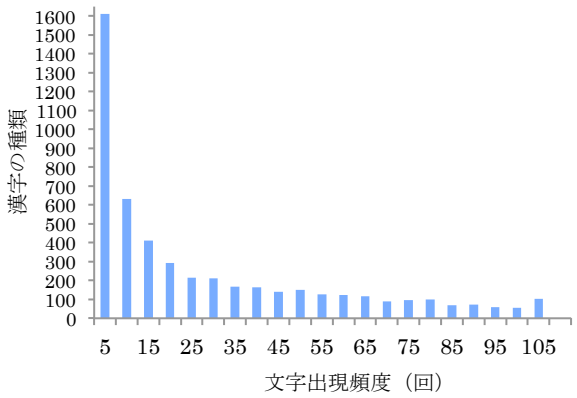

図 2　JIS 第一，第二水準漢字の出現頻度

手順Ⅳの SVM を用いた識別では，手順Ⅲで抽出した特徴ベクトルを, SVM を用いて学習させ, 文字の分類を行う．我々の研究においても, 漢字 256 種類に対して行った SVM を用いた認証実験において，認識率 92.5%という結果を残している[3].

2.2 データ収集手法

学習データ収集の目標として，システムの性能向上のために学習データとなる漢字の種類を増やすこと，認識率を上げるために同一漢字の個数を増やすことがあげられる．

多フォント活字認識手法の精度向上のためのフォント収集の主な手法として，Web アプリケーションを用いた収集方法がある [8]．Web アプリケーションを用いることで，総文字数 5482，311 種類の文字を収集するための時間が，手作業で 42 時間であるのに対して 10.5 時間で可能になる．

日本工業規格の 1 つとして選定された日本の文字コードである JIS 漢字コードは，第一水準から第四水準まで存在している．常用漢字表など多種の漢字表に登場し，使用頻度の高い文字を集めた JIS 第一水準と, JIS 第一水準と比較して使用頻度は低いが人名などで使用される漢字を集めた JIS 第二水準はそれぞれ 2965 種類, 3390 種類となっており，あわせて 6355 種類の漢字がある．図 1 は，漢字の出現率をグラフ化したものである．同一フォントにおいて収集した漢字の種類とその数の漢字を集めるために必要な近代書籍の冊数の関係を示す．グラフより，漢字を 6355 種類収集するには, およそ 4500 冊の書籍が必要になるものと考えられる．しかし，近代デジタルライブラリーで公開されている書籍のうち 1000 冊以上公開している出版者はほとんど存在しない．表 1 は主要な出版者とその公開冊数である．表 1 より，ほとんどの出版者が発行している書籍は 1000 冊以下であり，時代も考慮すると，中には 1 冊しか出版していない出版者も存在する．

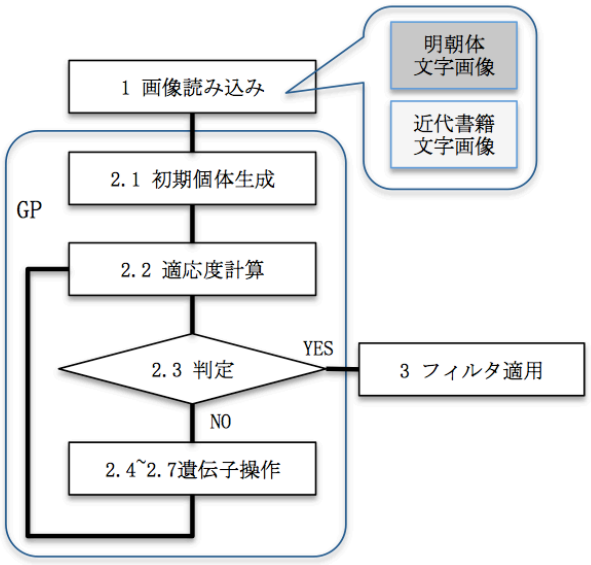

図 3　提案手法のフローチャート

そのため，近代書籍から十分な種類の学習データを収集することは困難であるといえる．

多フォント活字認識手法に用いる学習データは，十分な認識率を得るために文字 1 種類あたり 5 個以上必要である．図 2 は，近代デジタルライブラリーに存在する近代書籍と同タイトルの青空文庫[11]で公開されている書籍 105 冊における文字の出現頻度と種別の関係を示したグラフである．図 2 より，近代書籍における同一漢字の出現率は 5 回以下のものがほとんどであることがわかる．したがって，1 つの出版者が刊行した書籍から十分な個数の学習データを収集することは，限度があるといえる．

3. フォントセット生成手法

近代書籍のフォントセットを自動生成する手法を提案する．本稿では，GP を用いて明朝体を特定の文字へ変換するフィルタを生成する．明朝体を使用するのは，現在日本で広く普及しているフォントであり，どのような作業環境下でもフォントを入手することが容易だからである．生成されたフィルタによって明朝体のフォントセットから特定の出版者のフォントセットを自動生成する．フォントセットを自動生成することにより，書籍から文字を切り出す方法では十分な収集が困難な多く種類の文字を集めることが可能となる．

3.1 提案手法

提案手法の概要を図 3 に示す. 手順は以下の通りである.

1　教師データの原画像を読み込み
2　GP を用いたフォント変換式の生成
　2.1　初期個体の生成

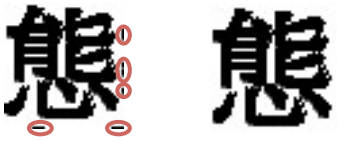

図 4　黒線除去前後の画像

図 5　明朝体と日吉堂・明治中期データセット例

2.2　適応度の計算
2.3　終了条件の確認
2.4　ルーレット選択による交叉
2.5　ランダム選択で選んだ個体の突然変異
2.6　適応度の計算
2.7　適応度の低い個体の削除，新たな個体の生成
2.8　手順 2.3 へ戻る

3　変換式で文字画像の変換・画像修正

手順 1 では，GP で使用する教師用データを読み込む．教師用データとは，画像データであり，原画像と目標画像の 2 枚セットで 1 組とする．原画像として手動で作成した明朝体の文字画像を，目標画像として近代デジタルライブラリーで公開されている書籍から切り出した出版者固有の文字画像を使用する．これらの画像は，二値化され，縦・横の大きさが同じものである．

手順 2 では，GP を用いて明朝体文字画像を変換するフィルタとなる変換式を生成する．手順 2.1 では初期個体の生成を行う．終端要素は，0〜9 の整数，原画像の輝度値を使用する．非終端要素には，四則演算子，絶対値，三角関数 sin・cos，絶対値，平方根を使用する．手順 2.2 では，適応度を計算する．適応度は，原画像を GP で生成したフィルタを用いて変換した出力画像と，目標画像の輝度値を比較することにより求める．適応度$fitness$を求める式は，以下の通りである．

$$fitness = 1 - \frac{\sum_{x=1}^{W_x}\sum_{y=1}^{W_y}|O(x,y) - T(x,y)|}{W_x * W_y * V_{max}}$$

ここでW_x，W_y，V_{max}は，画像の縦幅，横幅，輝度値の最大値である．$O(x,y)$と$T(x,y)$はそれぞれ，出力画像と目標画像の輝度値である．適応度は出力画像と目標画像の画素の差が小さいほど適応度は高いことになり，最大で 1 となる．手順 2.3 では，終了条件を満たしているか判断する．適応度が 1 になるか，パラメータを変化させても同じ適応度が一定世代続く場合は，計算を終了する．手順 2.4，手順 2.5 では，遺伝子操作を行う．ルーレット選択で選んだ個体を交叉させた後，ランダム選択で選んだ個体を突然変異させる．手順 2.6 では，手順 2.2 と同様に適応度を計算する．以下，終了条件を満たすまで手順 2.3 から手順 2.8 を繰り返す．

表 2　GP の実験条件

実験	教師データ	交叉率	突然変異率
(a)	1-300	0.7	0.1
(b)	301-600	0.7	0.1
(c)	601-900	0.7~0.9	0.1~0.3
(d)	901-1000,1-200	0.7~0.9	0.1~0.3

手順 3 では，フィルタを用いた画像生成と生成画像の修正を行なう．画像生成では，明朝体の文字画像を手順 2 で生成されたフィルタを用いて変換し，近代書籍用 OCR のための特定フォント画像を生成する．画像修正では，生成した画像の 4 辺に現れる黒線を除去する．目標画素周辺の画素情報を基に文字領域以外の独立した黒点部分を判別し，除去する．図 4 は，黒線が残った状態と黒線除去後の画像である．

4. 実験

第 3 章で提案した手法を用いて学習データ用フォントセットを自動生成し，それらの PDC 特徴を抽出し，生成したフォントセットの有効性を確認する．

4.1 実験条件

実験に使用するデータセットには，原画像として目標画像と同じ大きさの明朝体画像を，目標画像として近代書籍から切り出した文字画像を使用する．近代書籍の文字画像は，近代デジタルライブラリーで公開されている同一のフォントが使われている書籍の本文画像から切り出して取得する．同一のフォントとは，時代・出版者が同じ書籍に使用されているフォントのことである．切り出した文字画像を取り扱いが容易である PGM 画像に変換し，余分なシミを削除するために二値化する．明朝体の画像は，目標画像の漢字と同じ漢字の画像を，Gimp を用いて手作業で作成したものを使用する．同じ文字の明朝体画像と近代書籍の画像 2 枚で 1 組のデータセットとする．図 5 は，データセットの一例である．データセットを 1000 組用意し，1 から 1000 までの番号を付けて区別する．番号は JIS コードの順番（音読み優先）を採用し，順につける．したがって，文字の複雑さなどは考慮されず，比較的ランダムに番号付けされている．

実験で使用する近代書籍のフォントとして，日吉堂から明治中期（1883-1897）に出版された書籍に使用されている

表 3　実験ごとの一致率

実験	平均	最大値	最小値	分散 (10^2)	実行時間
(a)	0.708	0.888	0.446	0.4683	2h21m32s
(b)	0.712	0.884	0.485	0.4418	1h04m35s
(c)	0.706	0.882	0.443	0.4705	10h51m11s
(d)	0.688	0.853	0.438	0.4509	10h18m53s

表 4　ユークリッド距離

実験	平均	最大値	最小値	分散	標準偏差
(a)	542.46	92.63	189.84	9629.44	98.13
(b)	548.69	961.96	201.69	10717.56	103.53
(c)	536.64	953.06	203.42	9505.59	97.50
(d)	535.52	956.91	208.09	10925.04	104.52
(e)	342.58	367.85	305.21	306.90	19.92

ものを用いる．日吉堂の明治中期に出版された書籍は，まとまった冊数存在する．そのため，Web アプリケーションにより，多くの種類の漢字の収集が完了している．教師データとして使用する漢字は，収集済の日吉堂・明治中期漢字のうち 2 個以上収集されている文字を選択する．

1000 種類のデータセットのうち，GP に使用する教師データとして 300 種類分の画像データを番号順に選択し，提案手法を実行する．教師データの選択は 4 回行う．異なる教師データを用いることによって，生成されるフィルタの差異を確認する．表 2 は実験（a）-（d）における教師データの組み合わせや交叉率・突然変異率などの条件を示す．GP のパラメータである交叉率，突然変異率は，実験（a），（b）では固定し，実験（c），（d）では一定世代同じ適応度が続く場合，交叉確立率を 0.7-0.9 に，突然変異率を 0.1-0.3 にランダムに変更する．初期個体は 200，教師データ数は 300 に固定する．実験環境として，CPU は Intel® Core™ i7-4770K CPU @ 3.50GHz × 8，メモリは 15.5GB の計算機を使用する．

教師データ以外の文字に対して，生成されたフォント変換式の有効性を検証するため，生成された式を適用して残り 700 種類の明朝体漢字データを変換する．変換した漢字データと目標とした出版者・時代の画像データの類似度を比較する．

教師データに用いた漢字 300 種類と教師データと異なる漢字 700 種類の合計 1000 種類をフィルタで変換し，生成したフォント画像が多フォント活字認識手法の学習データとして有効かどうかを検証する．生成したフォント画像 1000 種類と目標画像として使用した日吉堂明治中期漢字の画像 1000 種類の PDC 特徴を抽出し，ユークリッド距離を計算し，類似度を調べる．近代書籍に特化した多フォント活字認識手法は，PDC 特徴を用いて文字の特徴ベクトル化を行い，SVM により特徴ベクトル学習をすることで文字認識を

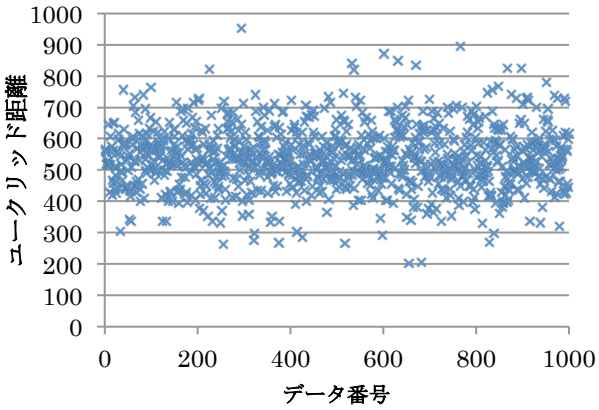

図 6　データセット番号毎のデータ分布

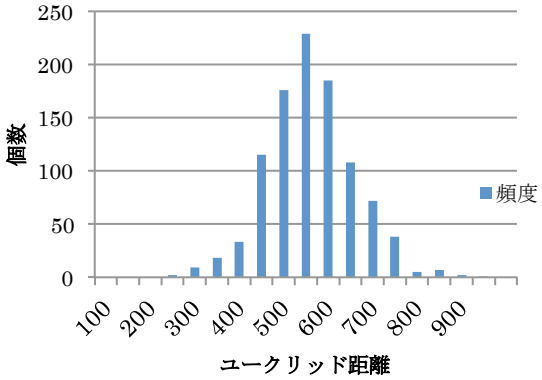

図 7　ユークリッド距離毎の個数

行う．そのため，画素の差で比較する画像の類似性より PDC 特徴の類似性があることが望まれる．

4.2 実験結果

提案手法の実験結果を表 3 に示す．一致率は，教師データに用いた 300 文字を含めた目標画像と生成したフォント画像 1000 字の黒線除去前後の一致率である．表 3 より，生成されたフィルタで変換した漢字全てにおいて 70%近い一致率を得ていることが分かる．実行時間は，条件によるがおよそ 1-10 時間必要である．特定の出版者のフォントを生成するフィルタを 1 度作成すれば，そのフィルタを用いて明朝体画像を変換することで何千種類も学習データを生成することができる．本稿では，300 種類の文字の収集にかかる 10.5 時間と GP 実行時間のおよそ 10 時間，1000 種類の明朝体へのフィルタの適用にかかる数秒と全ての工程を合わせると 20 時間で 1000 種類の漢字からなる学習データセットを生成することができる．明朝体の変換にかかる時間は，わずか数秒であるため，文字数が増えてもフォントセット生成にかかる時間は変わらないと考えられる．従来の手法では，およそ 311 種類の文字を収集するには 10.5 時間必要であることから，提案手法では同じ時間で 1.6 倍の

種類の漢字を収集できることがわかる．したがって，提案手法によるデータセットを生成した場合，処理にかかる時間と収集可能な文字種や文字数を考慮すると，この手法は有効であるといえる．

表 4 は，目標とした日吉堂・明治中期漢字と明朝体をフィルタで変換した生成したフォント画像の PDC 特徴の類似度を，ユークリッド距離を比較することで求めた結果である．1000 種類の目標画像と生成したフォント画像の類似度の最大値，最小値，平均，標準偏差を表に示す．実験（e）は，日吉堂・明治中期の漢字「愛」4 個の PDC 特徴を抽出し，総当りでユークリッド距離を求めた結果である．同出版社の同時代の漢字の PDC 特徴のユークリッド距離は，平均して 300 前後である．生成したフォント画像群では平均 500 台である．つまり，GP で生成したフィルタ書籍から切り出した漢字同士のユークリッド距離と比較すると，フィルタで変換し作成したフォントは，近代書籍に使用されたフォントの完全な再現には至っていないといえる．図 6 は，1000 種類のデータのユークリッド距離の散らばりを表している．図 7 は，データの個数とユークリッド距離の関係をヒストグラムで表したものある．図 6 のユークリッド距離の分布より，550 前後にデータが集まっていることがわかる．図 7 より，500-600 周辺に全データの半数が集まっている．ゆえに，散らばりの少ないまとまったデータセットが生成されたと考えられる．

5. まとめ

本稿では，近代書籍に特化した OCR のための十分な学習データを，変換フィルタを用いて生成する手法を提案した．文字変換フィルタを用いて明朝体の文字画像データを特定の出版者固有のフォントへ変換する．明朝体文字画像を特定のフォント画像に変換するフィルタは，GP により生成する．

実験では，データセット 300 種類の組み合わせを 1-300，301-600，601-900，901-1000 と 1-200 と変更し，交叉確立や突然変異率も進化過程に応じて変更するように指定した．使用した書籍は，出版者は日吉堂で，時代は明治中期である．実験より，フィルタ生成のために学習データとして使用した漢字以外を変換した場合でも目標とする近代書籍から抽出した文字画像との一致率は 7 割を超えることが分かり，類似した文字を生成するという観点からフィルタは有効性であると考えられる．提案手法で生成した学習データセットは，ユークリッド距離が 500-600 前後に集中しており，ばらつきが少ない学習データセットが生成された．これより，近代書籍用 OCR のための学習データとして使用することに適しているといえる．

本稿では 1000 種類の漢字を生成したが，明朝体の漢字画像 6000 種類，GP の教師データとして特定出版者の文字 300 種類の用意さえすれば，提案手法でフィルタを生成することができ，生成したフィルタを使用することで，特定の出版者の 6000 種類の文字からなる近代書籍用 OCR のための学習データセットを生成することが可能である．また，近代書籍の数が十分でないため，同一文字において十分な教師データの個数が収集不可能な場合でも，明朝体文字をフィルタで変換し生成することで，教師データの個数を増やすことができる．

今後は，より最適な教師データの組み合わせや GP の条件を選定することが課題として考えられる．

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．また，研究データの作成にご協力いただいた奈良女子大学大学院生である國松香苗さん，舩阪真生子さん，馬瀬春香さんに感謝いたします．

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) 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).
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 (PDPTA’09)，Vol.Ⅱ，pp．728-734(2009).
5) Awazu,T. , Fukuo,M., Takata,M. and Joe,K. : A Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books with Ruby Characters, International Conference on Pattern Recognition Applications and Methods (ICPRAM 2014), 637-645 (2014.3)
6) 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)
7) 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，研究報告数理モデル化と問題解決（MPS）研究報告(2012)
8) Kosaka.K.,Awazu.T.,Ishikawa.Y.,Takata.M. and Joe.K.: An Effective and Interactive Training Data Collection Method for Early-Modern Japanese Printed Character Recognition,Proceedings of 2015 International Conference onParallel and Distributed Processing Techniques and Applications(PDPTA 2015)
9) 栗津妙華，高田雅美，城和貴.:活字データ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），8(1)，72-79（2015-03-30)，1882-7780
10)　栗津妙華，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，研究報告バイオ情報学（BIO），2014-BIO-38(20)，1-6（2014-06-18）
11)　青空文庫
http://www.aozora.gr.jp/