Vol.2012-MPS-87 No.34
2012/3/2

近代デジタルライブラリーの文字切り出しにおける実際的手法

福尾真実[†1]　高田雅美[†2]　城和貴[†2]

本研究では，近代デジタルライブラリーが所蔵する画像データから上手く文字を切り出す実際的手法の開発を行う．国立国会図書館では近代デジタルライブラリーとして，所蔵する書籍を Web 上で一般公開している．これらは，画像データとして公開されており，文書内容を用いた検索が行えないため，早急なテキスト化が求められている．そのため，近代書籍に特化した多フォント漢字認識手法が提案されている．しかし，ルビが振られた書籍からは上手く文字が切り出せず，認識ができない．そこで本稿では書籍の本文からルビを取り除く手法を開発する．

Practical technique in the Kanji character clipping of the Digital Library from the Meiji Era

Manami Fukuo,[†1] Masami Takata[†2] and Kazuki Joe[†2]

In this research, we develop a practical technique to clip the kanji character well from the image data that the digital library from meiji era houses. The national diet library in Japan is opened to the public as the digital library from meiji era on the Web. There are shown as image data. Since it is impossible to perform full text search, it should be converted to text data. Therefore, it has been proposed the multi-fonts kanji character recognition method for early-modern Japanese printed books. Kanji characters with rubi occur that the kanji character clipping and recognition are badly constructed. In this paper, we propose a technique to remove the rubi from body of the book.

†1 奈良女子大学理学部情報科学科
Dept. of Advanced Information and Computer Sciences, Nara Women's University
†2 奈良女子大学大学院人間文化研究科
Graduate School of Humanities and Sciences, Nara Women's University

1. はじめに

国立国会図書館[1] では，所蔵する明治から昭和初期にかけての近代書籍を近代デジタルライブラリー[2] として Web 上で一般に公開している．近代デジタルライブラリーで公開されている近代書籍は画像データとしてアーカイブ化されているため，文書内容を検索条件として利用することができない．そこで，資料をより便利で手軽に利用することができるようにするために，書籍の早急なテキストデータ化が望まれている．近代デジタルライブラリーに収録されている書籍は明治期から昭和前期までの資料で約 570,000 冊と膨大である．ただし，インターネット上で閲覧可能な書籍は約半数の約 240,000 冊である．また，著作権[3] が切れた作品が順次追加されていくため，今後蔵書はますます増加していく．ゆえに，手作業によるテキスト化は効率的ではない．そのため，自動テキスト化が望まれているが，一般的な光学文字認識（OCR）ソフトウェア[4] による文字認識は，旧字体や異字体が多く含まれる近代書籍には適していない．そこで，先行研究として，近代書籍に特化した多フォント活字認識手法が提案されている[5][6]．

この先行研究の手法では，与えられた書籍画像に対して，ノイズ除去などの前処理を行ってから文字切り出しを行い，外郭方向寄与度（Peripheral Direction Contributivity, PDC）特徴[7] 抽出を用いて文字の特徴ベクトル化を行う．この特徴ベクトルに対して，機械学習の 1 手法であるサポートベクターマシン（Support Vector Machine, SVM）[8] を用いて特徴ベクトルの学習を行っている．この際，切り出した漢字にルビが振られていると，特徴ベクトルの値がまったく別の値となるため誤認識を引き起こす原因となる．そこで，ルビの除去作業をする必要がある．しかし，現在の書籍と異なり，ルビに規格がないため OCR を用いたルビの除去作業はできない．そのため，新たにルビの除去方法を考える必要がある．

ルビの除去を行っている関連研究[9] として，新聞画像を用いたルビの除去方法がある．この新聞を用いたルビの除去では，異方性ガウシアンフィルタ[9] が用いられている．手順としては，まず，異方性ガウシアンフィルタを縦方向に適用し，2 値化を行うことで太い連結成分と細い連結成分ができる．この細い連結成分がルビに相当する部分である．しかし，この研究では，新聞に存在する罫線を除去する手法を用いて細い連結成分を除去しているため，罫線が存在しない近代書籍ではこの手法が使用できない．そのため，細い連結成分を消去する手法を新たに考えなくてはいけない．

そこで，本稿では，この関連研究を改良することによって，近代デジタルライブラリーが所蔵する画像データからルビの除去を行う手法を開発する．

Vol.2012-MPS-87 No.34
2012/3/2

図 1 各書籍から切り出した文字画像

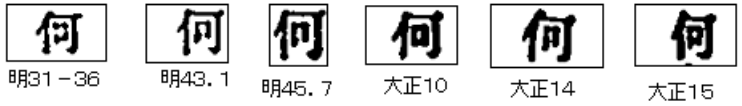

図 2 異なる時代の文字画像

先行研究で提案された手法について 2 章で述べる．3 章ではルビの除去の関連研究について述べ，4 章でルビの除去作業を行うために必要な処理について提案する．そして，5 章で実験と結果について述べ，最後に 6 章でまとめを述べる．

2. 先行研究における活字認識手法

図 1 は，出版者別に「何」という漢字のフォントを比較したものである．また，図 2 は春陽堂という出版者の年代別のフォントの比較を行ったものである．「何」という漢字 1 つをとっても線の太さや線と線との隙間などに違いが見られる．このように，出版者や年代が異なると使用されているフォントも異なる．そのため，近代書籍に特化した活字認識手法が必要である．

先行研究で提案されている近代書籍に特化した多フォント活字認識手法の手順は大きく分けて次の 3 通りである．

- 文字切り出し
- PDC 特徴抽出
- SVM を用いた特徴ベクトルの学習

まず，文字切り出しを行うが，その前に前処理として 3 つの処理を行う．書籍の画像データは，見開きで 1 つのデータである．そこで，1 つ目の処理として左右のページの分割作業を行う．次に，ノイズ除去[5] を行い，3 つ目の処理として，書籍をスキャンした際に傾いたページを修正するためのアフィン変換[10] を適用する．文字切り出しでは，黒画素射影ヒストグラムと背景領域射影ヒストグラムによる文字切り出し手法[11] が用いられている．PDC は，丁寧に書かれた楷書体手書き文字を正しく読み取る能力が高い特徴抽出方法である．この手法を用いて，漢字の特徴ベクトルを抽出する．そして，識別器として SVM を用いる．SVM は Vapnik らによって考案された機械学習の 1 つである．

図 3 誤認識を起こした画像　(a) ルビが原因 (b) 類似した構造が原因 (c) 線の欠落が原因

先行研究では，9 冊の書籍で共通して使用されている漢字 262 種類から 16 種，32 種，64 種，128 種，256 種と文字種を増やし評価実験を行っている．1 クラスにつき 9 個の画像データを用い，無作為に 5 個のデータを選び教師データとしている．また，残りのデータはテスト用の未知データとしている．結果，どの場合であっても 92%以上の認識率を得ている．

誤認識の原因は，ノイズ，類似した構造，線の欠落の 3 つに分けることができる．図 3 は，これらの原因で誤認識された漢字の例である．図 3(a) は，切り残されたルビがノイズとなっている．ノイズになる要因としては，この他に，書籍の劣化や裏抜けなどがある．図 3(b) は，類似した構造である「眼」と認識されている．これは，垂直方向や水平方向に似た構造の直線があることが原因である．図 3(c) は，書籍の劣化以外に，ノイズの除去の際に必要な線を削除することが原因となって誤認識される．

そこで，本研究では，ルビが引き起こしている誤認識に着目し，書籍画像からのルビの除去を目指す．

3. 罫線を利用したルビの除去

現在の書籍のルビは規格が決まっているため OCR を適用することができる．しかし，近代書籍のルビには規格がないため除去することはできない．関連研究として，古い新聞画像におけるルビの除去の研究[9] がある．この研究では，Scale Space の手法[12] を用いている．この手法は，欧文手書き文書の単語単位に連結させて文字切り出しを行うことを応用

Vol.2012-MPS-87 No.34
2012/3/2

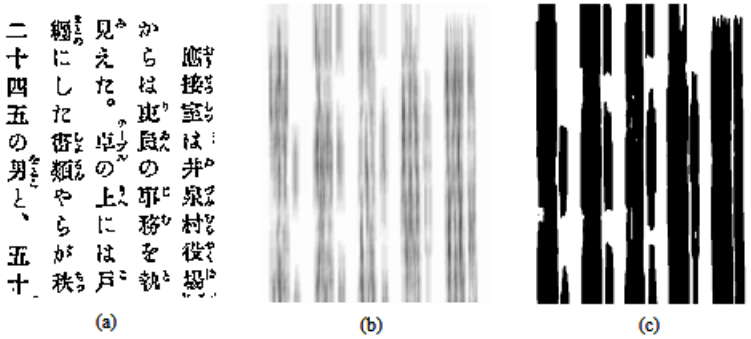

図 4 (a) 元の画像 (b) 異方性ガウシアンフィルタを適用した画像 (c) 図 (b) を 2 値化した画像

している．文字を連結させるには，対象画像に異方性ガウシアンフィルタを適用している．ガウス分布の関数は式 (1) で表される．

$$G(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}\right)\right] \quad (1)$$

欧文手書き文書では，横書きのため横方向に連結を行っている．しかし，新聞や近代書籍では文字は縦書きのため上から下へ辿っていく．そのため，縦方向に連結させる必要がある．異方性ガウシアンフィルタを用いて，対象画像の縦方向に強くぼかしをかけるが，縦方向に連結させるため，x 方向は $\sigma_x = 1$ とし，y 方向の σ_y の値を変化させていく．図 4 の (a) は近代デジタルライブラリーに所蔵されている書籍の一部分の画像である．そして，図 (a) の書籍画像を縦方向に異方性ガウシアンフィルタにかけたものが図 4 の (b) である．図 (b) で用いた σ_y の値は 127 である．その後，より鮮明に差異を表現するために 4 の (c) のように 2 値化を行う．2 値化により，太い連結成分と細い連結成分に分けられる．図（a）の元の書籍画像と見比べると，細い連結成分がルビの部分であると考えられる．

この細い連結成分を除去するために，関連研究では罫線の除去の手法を用いている．罫線の除去では，連結成分にラベリング処理を行い，おおよその行幅を推定する．そして，そこから大きくかけ離れた連結成分を罫線とみなし，削除している．しかし，そこで，4 章では細い連結成分の除去を異なる手法で行う．

4. ヒストグラムを用いたルビの除去

書籍には罫線は存在しないため，3 章の手法は使用できない．また，近代書籍に振られているルビは新聞のルビと異なり，本文に非常に接近している場合がある．その場合，図 4 の (c) の画像の 1 番右の行のように，処理を行うと太い連結成分と細い連結成分が一体化してしまうことがある．そのため，3 章の手法を改良する必要がある．

書籍からルビを除去するために，まずヒストグラムを取り，連結成分の 1 つの太さを調べる．その際，書籍には柱[⋆1]やページ番号が振られており，ヒストグラムを取得する際，処理の妨げになる．また，書籍をスキャンする際にノド[⋆2]の部分が太い黒の線となっている．このノドは，ページ分割の際に，縦線として残る．そのため，ガウシアンフィルタを適用した際にノドの部分が太く残ってしまう．ノド以外に，書籍をスキャンする際に，小口[⋆3]の周辺が紙の劣化のためか，ノイズが激しく入ってしまうことがある．また，厚みのある書籍は書籍をスキャンする際に小口が影のように入ってしまい，分割の際に残ってしまう．そのため，ルビの除去を行う前にページ分割の手法も改良する必要がある．

そこで，ページ分割の改良手法から考える．ページ分割の際,「ノド」「柱」「小口」が残るとルビ除去の際に妨げとなる．そのため，ページ分割の改良手法として，これらの除去を行う．まず，ノドの除去を行うには縦方向にヒストグラムを取り，最大値を求めてノドの中央部を求める．また，柱は横方向にヒストグラムを取ると，本文と柱ではヒストグラムのサイズが大幅に異なることを利用して柱の除去を行う．小口の除去は，切り出す場所を精査し求める際に開始位置を小口からずらすことで除去できる．

次に，本文のみになった書籍画像に対して，ノイズ消去を行う．まず，ある程度の長さがある連結成分をヒストグラムを用いて探し，連結成分の横幅を調べる．その結果，太い連結成分と細い連結成分がきれいに分かれている場合は，細い連結成分を消去する．逆に，太い連結成分と細い連結成分がくっついてしまっている時は，漢字の部分だけ残るようにルビが書かれている方向から削る．この際，他の太い連結成分と同等の太さまで削る．図 5 の (a) が図 4 の (c) の 2 値化を行った画像から細い連結成分を消去したものである．また，図 5 の

⋆1 書籍名や章題名，節の見出しなどを記したもの．
主に各ページの版画（読み方は「はんづら」である．これは，印刷される文字，画像などが入るスペースのこと.）の外に配置される．
⋆2 本の綴じ目に沿った部分で，見開いた時の中央部分．
⋆3 背の反対側

Vol.2012-MPS-87 No.34
2012/3/2

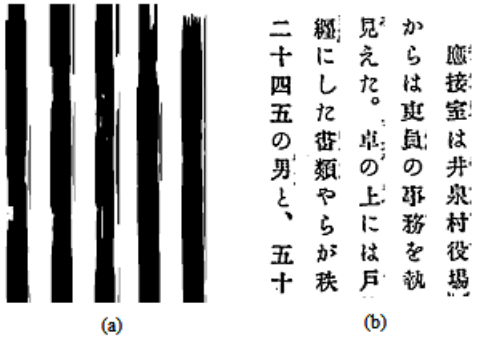

図 5 (a) 細い連結成分を消去した画像 (b) ルビを消去した画像

(b) は図 (a) の結果からルビの除去を行った画像である．

5. 実　　験

本研究では，ルビの除去を行うことで文字認識に使用できる文字画像が増加するかどうか調べる．5.1 節で実験に用いるデータについて述べ，5.2 節で実験結果と考察について述べる．

5.1 実験データ

本実験で用いたデータについて述べる．本実験では，近代デジタルライブラリーに所蔵されている書籍を用いる．タイトルは「田舎教師」，著者は「田山花袋」である．また，出版者は「左久良書房」で，出版年は明治 42 年である．ファイル形式は JP2 であり，見開き 1 ページで 1 つの画像データとなっている．

画像データ 1 つの大きさは，横 1214× 縦 925 ピクセルであり，ファイルサイズはおおよそ 120〜350KB の間である．書籍画像のページの構成は，中央にノドの黒の太い線があり，ページ外に中央線も存在する．柱には作品名とページ番号が振られている．柱は上部にあり，ページ番号が小口よりにある．小口はかなりはっきりと写っており，書籍が痛んでいるために，天[⋆1]の付近と小口の部分に汚れが目立っている．また，ページによっては前の持ち主の書き込みらしきものも所々に見られるが，極一部であるため，書き込みがあるページも文字切り出しに用いる．

次に，手順を説明する．先行研究の手順は 2 章で説明している通りに，3 つの前処理を行い，その後文字切り出しを行う．ルビの除去を行う文字切り出しの手順は，まずページの分割を行い，ノイズ除去を行う．その後，ページの角度補正を行うために，アフィン変換を適用する．ここまでは先行研究の手順と同じである．ここまでの手順の次に，ルビの消去を行い，文字切り出しを行っていく．

実験環境として，CPU は Intel(R) Core(TM) i5-2500k CPU @ 3.30GHz 3.30GHz，メモリは 4.00GB の計算機を用いる．また，2 値化などの画像処理には OpenCV を用いるが，他の処理との兼ね合いにより OpenCV2.1 を用いる．これは，OpenCV2.2 以降は 2.1 以前に比べライブラリ構成が大きく変わっているためである．

5.2 実験結果と考察

実験はページ分割の改良手法が有効であるか調べるために先行研究と比較を行う．また，ルビを除去した書籍画像から文字切り出しを行うと先行研究の手法と比べ，文字数や文字種が増加するか調べる．5.2.1 節でページ分割の実験を行い，5.2.2 節でルビの除去に関しての実験を行う．

5.2.1 ページ分割の改良

ページ分割の改良の有効性について調べるために，ヒストグラムを用いた分割手法に変更し，先行研究の手法を用いてページの分割したものと比較する．先行研究の手法では，分割に失敗した画像は 200 枚中 10 枚ほどである．失敗した原因は，小口の部分が色濃く写っており，その部分を中央であると誤認識したと思われる．改良した手法では，分割できなかった画像は 1 枚のみである．近代デジタルライブラリーでは，デジタル化を行う際に初期の頃はフィルムに撮影している．そのため，書籍を開いたままの状態でページにたるみができやすく本文が傾いていることが多い．改良手法でも分割できなかった画像は，書籍をスキャンする際にページがたるみ，ノドの黒い境界部分が斜めになっているためであると考えられる．しかしながら，多くの画像で，先行研究の方法と異なり，分割が可能となっている．そのため，書籍のページ分割の手法としてヒストグラムを用いた分割手法の改良は有効であると分かる．

5.2.2 ルビの除去を用いた文字切り出し

ルビを除去した画像から行切り出しを行うとルビの除去を行わないで行切り出しを行った

⋆1 本を立てた場合に上に見える切り口（各ページの最上部）

Vol.2012-MPS-87 No.34
2012/3/2

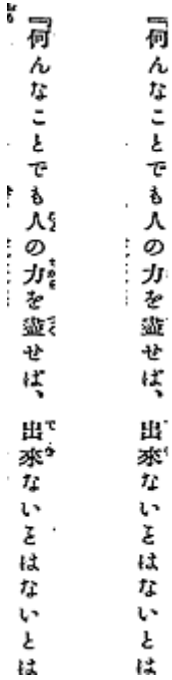

図 6 （左）元画像のまま行切り出しを行った画像（右）ルビを除去した後行切り出しを行った画像

場合とどのように違いが出るか比較する．図 6 の左側が元の画像のまま行切り出しを行った画像，右側がルビの除去作業を行った後，行切り出しを行った行画像の一例である．文字切り出しの手順としては，行切り出しを行った後に，文字を切り出していくため，図 6 の左側のような画像ではほとんどの漢字にルビがついてきてしまう．そのため，文字認識の際には使用できない文字画像となってしまう．そのため，行切り出しの時点ではルビの除去を行った画像は有効である．

先行研究の手法で得た画像データとルビを除去したのちに文字切り出しを行った画像データの例を図 7 に，先行研究の手法とルビを除去したのちに文字切り出しを行った結果を表 1 に示す．

図 7 の左側が先行研究の手法で文字切り出しを行った画像で，右側がルビを消去したのち文字切り出しを行った画像である．先行研究の手法では，文字切り出しを行っても図 7 の

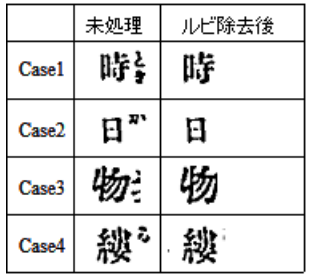

図 7 ルビを除去した文字画像

表 1 先行研究との比較

	先行研究	ルビ除去
文字画像数	194 個	369 個
文字種	118 種類	210 種類

Case1，2，4 の左側の画像ようにルビが残ってしまったり，Case3 のように中途半端にルビが残ってしまっている．そのため，文字認識の妨げとなる．一方，提案手法を用いてルビを消去し，文字切り出しを行うと図 7 の右側の画像のように認識に使用できる画像となる．

表 1 より，文字認識に使用することができる画像数も先行研究の手法では 194 個であったが，ルビの除去を行ってから文字切り出しを行う手法では，369 個に増加している．文字種も先行研究の手法を用いた 118 種類から 210 種類とほぼ倍増している．ただし，書籍に含まれている文字の数と比較すると，ルビ除去を適用しても画像数は少ない．これは，アフィン変換が適切に用いられていないため，多くが傾いており，うまく文字切り出しが行えていないことが原因と考えられる．一方，ルビを除去することで地名や難読漢字に振られていたルビが取り除かれたことにより，切り出した文字画像にルビが含まれず文字認識に使用できる文字種が増加したと考えられる．実際，切り出した文字画像には「琵（び）」「琶（わ）」「廉（やす）」などの複雑な漢字が新たに増加している．

6. ま と め

本稿では，近代デジタルライブラリーの文字切り出しにおける実際的手法としてルビの除去法を提案した．

先行研究では文字切り出しを行い，PDC 特徴抽出で文字の特徴ベクトルを抽出した後，

SVM を用いてこのベクトルの機械学習を行うことで文字認識を行っている．しかし，文字切り出しを行った際にルビが画像データに残ってしまい，文字認識の妨げとなるため多くの文字画像が使用できなかった．また，文字切り出しを行う前に前処理として，与えられた見開き 1 ページの書籍画像を分割し，ノイズ除去を行い，スキャンの際に生じた傾きの補正を行っているが，先行研究の手法ではページの分割が正しくできずに，小口の部分で分断されてしまうものもある．そこで，まずルビを消去する前に書籍のページ分割の部分を改良している．ページの分割には縦方向のヒストグラムを用いることでノドの部分で正確に分割できる．また，横方向にヒストグラムを取ることで柱の部分を除去することができている．

次に，ルビの除去であるが，異方性ガウシアンフィルタを用いることでルビの消去が行える．手順は，まず縦方向に連結させたいため異方性ガウシアンフィルタを縦方向に適用し，その後に 2 値化を行う．その結果，細い連結成分と太い連結成分に分けることができ，細い連結成分がルビに相当する部分になる．そのため，この細い連結成分を消去すればルビの除去が行える．太い連結成分と細い連結成分が離れている場合は太い連結成分を基準とし，その太さより細い連結成分を消去する．また，ルビと本文が非常に近い位置にある場合には太い連結成分と細い連結成分がくっついてしまうことがある．その場合は，外側からルビの部分を削っていく．

実験では，先行研究の手法と改良したページ分割の手法の比較，ルビの除去を行ってから文字切り出しを行った場合を比較した．使用した書籍の画像データは，明治期のものである．書籍のページ分割では，書籍のノドなどが残ってしまう先行研究の手法では，10 枚の画像が分割に失敗しているが，改良したヒストグラムを用いた手法では 1 枚のみである．そのため，ヒストグラムを利用したこの手法はページの分割に有効である．また，ルビを除去した後に文字切り出しを行った場合，先行研究の手法と比較して文字認識に使用できる文字画像の数は 194 個から 369 個に増加している．また，文字種も 118 種類から 210 種類にほぼ倍増している．これは，もともとルビが振られていない漢字が漢数字などの簡単な構造のものであり，取得できる文字画像の数にも偏りがあったため，ルビの除去をして地名や難読漢字のルビが除去されることによって文字数や文字種が増加している．

現在のルビ除去の手法では書籍の文章が傾いていないことが前提となっているため，書籍の本文が大きく傾いていると上手くルビの除去ができない．そこで今後の課題としては，本文を基準としたより正確な書籍の傾き補正があげられる．

参 考 文 献

1) 国立国会図書館
http://www.ndl.go.jp/index.html
2) 近代デジタルライブラリー
http://kindai.ndl.go.jp/index.html
3) 著作権法
http://www.cric.or.jp/db/fr/a1_index.html
4) 中村 洋治，除村 健俊，豊川 和治，北山 友.: PC 上で動く印刷文字 OCR，情報処理学会第 33 回（昭和 61 年後期）全国大会，pp.1635-1636，1986.
5) Chisato Ishikawa, Naomi Ashida, Yurie Enomoto, Masami Takata, Tsukasa Kimesawa, and Kazuki Joe.: Recognition of Multi-Fonts Character in Early-Modern Printed Books, In Proceeding of International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA ’09), Vol.2, pp.728-734, 2009.
6) Manami Fukuo, Yurie Enomoto, Naoko Yoshii, Masami Takata, Tsukasa Kimesawa, and Kazuki Joe.: Evaluation of the SVM based Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books, In Proceedings of 2011 International Conference on Parallel and Distributed Processing Techniques and Applications(PDPTA ’11), Vol.II, pp.727-732, 2011.
7) 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，Vol.J66-D No.10，pp.1185-1192，1983.
8) Nello Cristianini and John Shawe-Taylor.: サポートベクターマシン入門，共立出版，2005.
9) 島 貴宏，寺沢 憲吾，川嶋 稔夫.: 新聞画像アーカイブのための画像処理技術の研究，電子情報通信学会技術研究報告，Vol.110 No.467，pp.1-6，2011.
10) 奈良先端科学技術大学院大学 OpenCV プログラミングブック制作チーム.: OpenCV プログラミングブック，毎日コミュニケーションズ，2007.
11) 中山 英久，藤原 勇太，加藤 寧.: 背景領域細線化を用いた手書き文字切出しの改良手法，情報科学技術フォーラム講演論文集，Vol.8 No.3，pp.365-370，2009.
12) Lewis D.Griffin, Martin Lillholm.: Scale Space Methods in Computer Vision，Spinger，2003.