2005／CH／66（8）
2005／5／27

江戸期版本画像からの文字切り出しの試み

坪井 昭憲*　八村 広三郎†　吉村 ミツ‡

概要　おもに江戸期に出版された版本をデジタル化した画像から、それぞれの文字を切り出すための手法とその処理結果について報告する。文字切り出しは、汚れやシミの除去、２値化、行の切り出し、ラベリング処理による文字の分離と統合の処理などからなっている。ここでは、２値化の閾値は、基本的に大津の判別分析法によって行うが、頁全体、行単位、さらには局所的な数文字のブロック単位でという風に順次適応的に適用することにより、汚れやシミの影響をあまり受けずに文字切り出しの精度を向上させることができた。各文字の認識は今後の大きな課題であるが、切り出しの結果だけからでも、さらにパターンマッチングなどを利用して、文字の使用頻度を求めたり、版本全体の文字のインデックスを作成したりできると期待される。

Segmentation of Characters for Historical Woodblock-Printed Books

Akinori Tsuboi*　Kozaburo Hachimura†　Mitsu Yoshimura‡

Abstract　This paper describes a method and experimental results of character segmentation for digitized Japanese historical woodblock-printed books. The method includes a removal of stain and smear, binarization, extraction of character lines and extraction of characters by region labelling. The binarization has been done by recursively applying the Otsu's method which realized a good performance of character segmentation eliminating the influence of stain and smears. The result of character segmentation can be utilized, for instance, for making an index and a dictionary of characters used in the books.

1　はじめに

近年、人文科学分野へのコンピュータ技術の応用として、有形あるいは無形の歴史・文化情報資源をデジタル化、保存・流通することを可能にする、デジタルアーカイブの研究が注目を集めている。これにより、これらの情報資源を国内外の多くの研究者により共有することができ、これらの情報を利用する学術研究の発展と高度化に資することができる。さらに最新の情報処理技術によるさまざまな解析を行うことにより、新しい研究成果を生み出すことが期待される。また、現代の産業や科学分野の情報とはかなり性質の異なった歴史的文化的情報を対象とすることで、より高度な情報処理技術の開発を促進することも期待される。

このようなデジタルアーカイブの対象とする情報の中でも、歴史的な文書である古文書や古記録は最も基本的なものであって、これらを対象としたさまざまな試みがある。さらに、このような古文書をコンピュータにより読み取り、「翻刻」を自動化しようという、きわめて野心的な古文書 OCR の研究も行われている [1]。

古代から中世にかけて、わが国の書物の主流を占めたのは、手で文字を書き写した写本であった。その後、中国で木版印刷が盛んになり多くの版本がもたらされると、鎌倉時代後期からわが国でも木版印刷による版本の出版が盛んになった。さらにその後、16 世紀の末にヨーロッパと朝鮮から、活字を用いて書物を印刷する活版印刷の技術が伝来し、これによって古活字版と呼ばれる書物の流通を促し、近世初頭の出版ブームのさきがけとなったといわれている。江戸時代初期の約 50 年間に、この方法によって多くの書物が出版されたが、活字版をつくるのが面倒であり、また多くの部数を刷れないので、江戸中期以降には次第にすたれ、江戸の寛永期には、1 枚の板木に 2 頁分の版を彫って印刷する木版整版印刷による「版本」の出版が主流となった。このように、江戸時代から明治時代にかけて、日本の印刷技術は著しく発展し、数千の版元から、読本、

*立命館大学 テクノロジー・マネジメント研究科
*Graduate School of Technology Management, Ritsumeikan University
†立命館大学 理工学研究科
†Graduate School of Science and Engineering, Ritsumeikan University
‡立命館大学 COE 推進機構
‡Center for Promotion of the COE, Ritsumeikan University

浄瑠璃本、農書、算法書、教訓書、故実・礼法書など、さまざまタイプの膨大な書物が出版されたといわれる。

これらの版本による出版物は、手書きによる、情報の通信伝達を目的とする本来の「古文書」とは異なるものであるが、ここでは広い意味で古文書に含まれているものと考える。

本研究では、情報処理により、版本による古典籍の文字を認識し解読することを長期的・最終的な目標としているが。写本とは異なる版本といえども、くずし字、つづけ字などの存在のため、文字認識の実現は大変困難な課題である。このため、ここでは、版本画像からの文字の切り出しまでを当面の目標にする。手書き文書からの文字の切り出し処理は、文字認識のための基本的で必須の処理であるが、規格化された現代の印刷文書とは異なり、これ自体が必ずしも容易な処理ではない。文字の切り出しまででも実現できると、画像としての文字によるインデックスや文字画像の字書の作成が可能になり、これを元に、特定の時代、種類、出版元などの字形の分布を統計的に研究することなどへの応用も可能になる。たとえば、文献 [2] では、古文書におけるキャラクタスポッティングの研究が述べられているが、このような処理は、字形の比較的安定しており、また分量の大きい（頁数の多い）版本にこそむしろ望まれるものだと考えている。

2　古文書 OCR

現在、活字 OCR(Optical Character Reader) の技術は実用技術として確立されており、様々な場面で活用されている。しかし、手書き文字 OCR の技術は、文字帳票の読み取りなど、限られた用途での読み取りが実用化されているが、すべての手書き文字を読み取れるレベルまでには達していない。古文書 OCR はこの手書き文字 OCR の中の究極の目標として位置づけられる。

古文書の文字の基本的な特徴は、（1）毛筆で書かれていること、（2）つづけ字が多いこと、（3）くずし字が多いこと、であってこれらは、コンピュータによる認識処理にとって非常に困難な点であり、現代の手書き OCR の手法を必ずしもそのままは適用できない。

古文書 OCR の代表的研究として HCR(Historical Character Recognition) プロジェクトがある。HCR プロジェクトにおいては、古文書文字データベースの作成、古文書文字の切り出しと認識手法の研究、知識による翻刻支援、電子化古文書文字辞典の開発などの様々な成果があげられている [1]。

古文書 OCR には、大きく分けて、文字切り出しと、文字認識の 2 つの処理が必要である。文字切り出しは、入力画像から個々の文字を分離抽出することである。文字切り出しは、文字認識の前処理に位置づけられているが、認識率を左右する重要なプロセスである。

文字切り出しは、文字認識のための基本的な処理であるが、一方で、文字の切り出しを行おうとすると、文脈や前後の文字あるいは単語としての文字間の関連性などの情報が必要になる場合も多く、ある程度の認識が切り出し処理に必要となるという側面もある。したがって、最近では、文字切り出しを必要としない、認識手法なども研究されている [3][4]。

3　版本からの文字切り出し

3.1　木版刷り版本の画像

本研究では曲亭馬琴著の「椿説弓張月」の一部を対象とした。「椿説弓張月」は文章と挿絵で構成されているが、本研究では、文章の頁だけを対象とした。図 1 にその一部を示す。

木版刷りの文書であるため、毛筆文字のような文字のかすれはあまり見られず、文字が明瞭である。行は明確に分かれている。一方、活字を使用していないのでつづき字は多く現れる。また、歴史的出版物であることから、紙面の劣化、シミ、汚れが見られる。さらに、本文の漢字には振り仮名が振ってあり、本文部分の文字抽出の処理の際に影響を与える。

図 1: 椿説弓張月

3.2　処理手順の概要

前述したように、本研究では、古文書文字認識の一般的な手順における、2 値化と文字切り出しを行う部分について検討する。

文字切り出しの処理手順を図 2 に示す。まず入力された古文書画像に対して 2 値化処理を行い、文字切り

出しに適した白黒 2 値画像にする。次に、画像の縦方向の射影分布を求め、これによって本文と振り仮名の「行幹」の抽出を行う。そして、この「行幹」を利用して行を抽出し行単位にラベリングによる文字切り出しを行う。次に一定の面積より大きい文字矩形を対象にして、再処理を行う。この処理でも文字が分離されず、さらに 2 文字以上あると思われる縦に長い矩形に対して、その矩形領域の範囲内で、水平方向の射影分布を求め，その特徴量を利用して対象の文字矩形をさらに切り分ける処理を行う。以下、各手順の詳細を述べる。

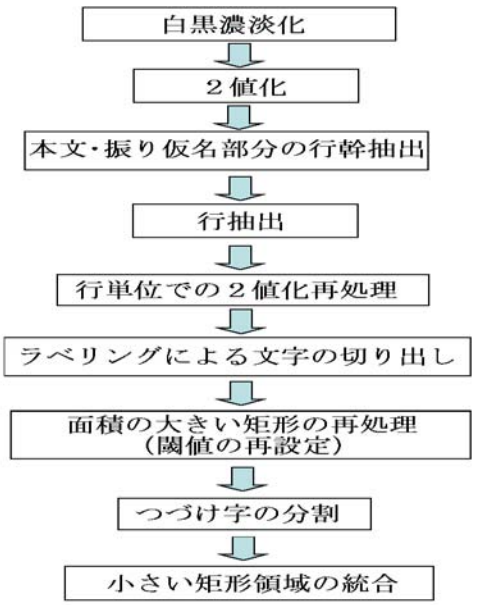

図 2: 文字切り出しの手順

3.3 2 値化

対象の版本はカラースキャナにより、カラー画像としてデジタル化する。入力したカラー画像を RGB 表色系から L*a*b*表色系 [5] へ変換し、明度 L*に対して大津の判別分析法 [6] をページ単位に適用して白黒２値化のための閾値を求める。

対象の版本は実際に何度も読まれたものなので、ページめくりのための手垢による汚れやシミが顕著な場合がある。このような場合、単純に明度情報だけに対して閾値処理を行うと、汚れやシミが文字部分と重なって以後の処理が困難になる。汚れやシミの色分布が、用紙そのものの色分布と若干でも異なる場合には、色彩に対する処理によりこれらの汚れやシミを取り除くことができる可能性がある。すなわち、L*a*b*表色系色相と彩度を表す平面を定義する a*、b*で判別分析を行い、シミ・汚れの部分を取り除くことができる。本論文執筆の時点では、この色情報によるシミ・汚れの処理は組み込んでおらず、以後の処理では、L*の明度情報を使った白黒濃淡画像およびこれを 2 値化した白黒 2 値画像をおもに取り扱う。

3.4 行幹と行の抽出

本研究では「行幹」と呼ぶ行の中心部分を求め、これを基準にして、以下の処理を行う。まず 2 値化した画像の垂直射影分布を求める。図 3 は、２値化した画像に、垂直射影分布を水平位置を合わせて重ねて示したものである。この図から、垂直射影分布のグラフは、本文部分と振り仮名部分に対応した山が周期的に現れていることが分かる。

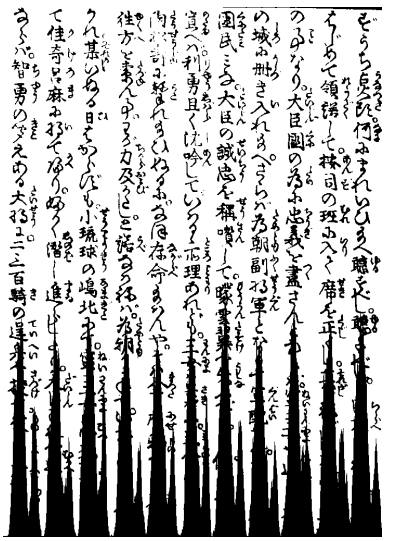

図 3: 垂直射影分布と文字行の関係

対象の版本は、現代の規格化した活字で作られたものではないので、それぞれの行で使われている各文字の大きさにはばらつきがあり、また、縦の行そのものにも水平方向に若干のゆがみや傾きがある。しかしながら、それぞれの本文行において、行内のほぼすべての文字の一部分が存在する、行の基幹部分が存在すると考えることができる。これをここでは「行幹」と定義して抽出し、以後の文字切り出し処理で、この行幹の情報を利用する (図 4)。

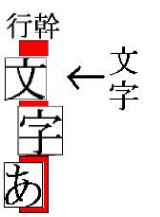

図 4: 文字行と行幹

２値画像全体の垂直射影分布の最大値を求め、その値の 1/2 を閾値として、閾値以上の領域を「行幹」として抽出する。「行幹」の抽出例を図 5 に示す。

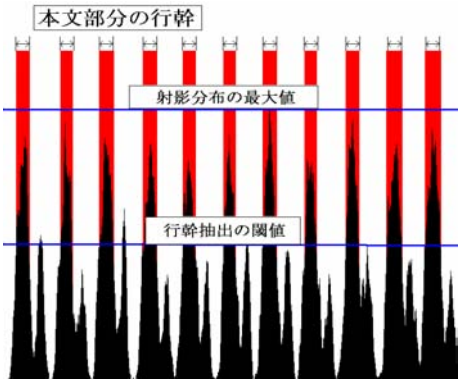

図 5: 垂直射影分布による本文部分の行幹抽出

本文部分の行幹を利用して、振り仮名部分の行幹も抽出する。図 5 からわかるように抽出された各行の本文の行幹の間には、振り仮名部分の小さな山がある。そこで、本文の行幹と行幹の間において振り仮名の行幹を求める処理を行う。この範囲内の垂直射影分布の最大値を求め、その値の 1/3 の値を閾値として設定し、この閾値以上の領域を振り仮名の行幹とする。

次に、求まった本文部分の行幹を利用して「行」の位置決めを行う。図 6 に示すように、隣合う本文部分の行幹間の中心位置を求め、それを行の端点とする。

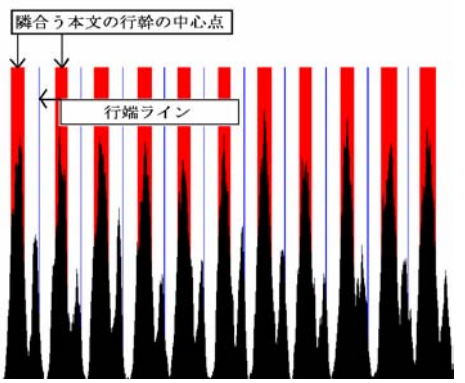

図 6: 行の位置決め

3.5　行単位での局所２値化

上の処理によりそれぞれの行を位置決めした後で、今度は、元の濃淡画像に対して行単位に判別分析法を適用して閾値を求め、この値で行内を再度２値化する。これにより、局所的な濃淡分布の変動を考慮した２値化処理が行われ、シミや汚れの影響を抑えることができる。

3.6　ラベリングによる文字の切り出し

２値化した行領域からの文字の切り出しには、領域のラベリング処理を利用する [7][8]。版本の対象画像では、文字のかすれが少なく、一文字一文字が明確であるため、文字のストローク部分が途切れることは少ない。２値化画像中の連結領域ごとにラベリングを行い、それぞれのラベルが付与された連結領域を囲む最小の外接矩形を求め、これで文字候補領域を切り出す。

求まった外接矩形のうち、行幹に重なるものは文字を形成する領域であるとみなし、これを文字矩形候補領域とする。さらに、相互に重なり合う文字矩形候補領域については、それらは一つの文字領域を構成するものとし、これらは一つに統合する。図 7 にその処理例を示す。

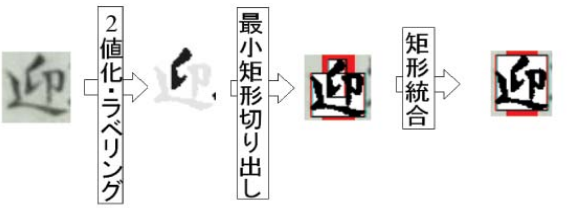

図 7: 領域ラベリングと矩形統合による文字切り出し

一方で、扁やつくりが別々の領域として抽出され、しかもこれらの矩形同士は相互に重なりあうことがない場合もある。このような、左右に分離した矩形同士であっても、これらの一部が行幹の領域に重なっておれば、これらは同じ文字の一部であるとみなして、これらをひとつに統合する。この処理の例を図 8 に示す。以上のようにして得られた文字矩形が、今後の処理での文字候補領域となる。

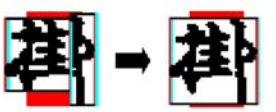

図 8: 左右に分離した文字の統合

3.7　大きな文字矩形領域内の再処理

以上のようにして抽出された文字矩形（文字候補領域）の中には、複数の文字がつながって大きな一つの矩形となっているものもある。このような、このような大きい矩形に対しては以下のような再処理を行う。

まずここまでの処理によって切り出された文字矩形の面積の平均値を求める。面積が、この値の２倍以上となっている矩形について、その矩形に対応する部分の白黒濃淡画像を切り出し、これに対して、再び大津の判別分析法を用いた２値化、ラベリングによる文字切り出し、矩形統合処理の処理を行う。このようにすることで、その矩形範囲内で適切な閾値が求められる。その手順と例を図 9 に示す。

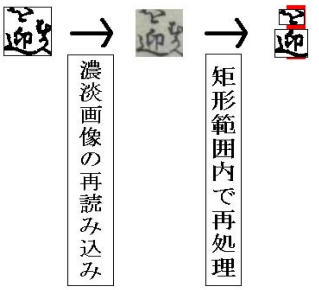

図 9: 指定面積以上の矩形領域内の再処理

3.8　つづけ字の分離

以上の処理を行った後、再度、文字矩形の大きさの平均値を求める。大きさが平均値の 1.5 倍以上あり、かつ、縦の長さがある閾値以上の縦長の矩形、すなわち、なお複数の文字をつづけ字として含んでいると推測される矩形に対して、これをそれぞれの文字に分離する処理を行う。まず、その矩形範囲内での水平方向の射影分布を求める。ページ単位で求めた文字の平均高の値で決まる一定の縦座標の範囲内において射影分布の値が最小となる箇所で、矩形を分割する。これを上から順に行い、分割された下方の矩形に対しても、さらに上の条件を満たしておれば、再帰的に分割処理を行う。図 10 に処理結果を示す。

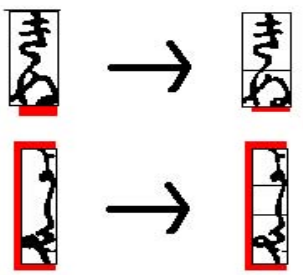

図 10: 指定面積以上の矩形領域内のつづけ字の分離

3.9　面積の小さい矩形領域の統合

平仮名の「う」や漢字の「三」などのように、文字の画（ストローク）が上下に分かれている文字では、ラベリング処理により、上下が別々の矩形として扱われ、独立した文字候補領域として抽出されてしまうことがある。ここではこのような状況に対処するため、矩形領域の統合を行う。文字矩形の大きさの平均値の 3 分の 1 を閾値として設定し、これより小さい文字矩形に対しては、矩形の統合処理を行う。その対象となっている矩形の重心と、その上下の矩形の重心を求め、対象矩形を、重心が近いほうの矩形と統合する。この処理の例を図 11 に示す。

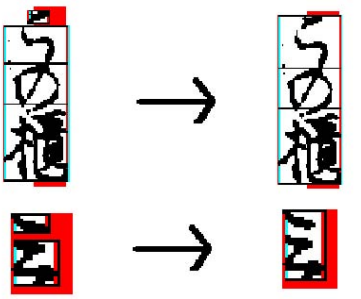

図 11: 面積の小さい矩形領域の統合処理

4　処理例

木版本「椿説弓張月拾遺巻之三、第五十一回」からディジタル化して得た画像のうち、挿絵のないページの、NO.1 から NO.11 の画像データに対して行った処理結果を示す。以下で、NO.3 の画像を対象として行った処理結果を示す。図 12 が NO.3 の画像データである。この画像の左下部分には、ページを繰り返しめくる際についたと思われる手垢による汚れがある。

以下の処理例では、上述の手垢による汚れの部分を含む画像の左下の部分のみを拡大して示すことにする。

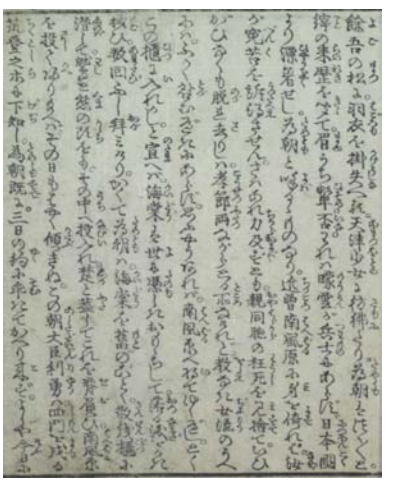

図 12: 対象画像 (NO.3)

4.1　行幹と行の抽出結果

本文と振り仮名の行幹の抽出結果を図 13 に、行の抽出結果を図 14 に示す。

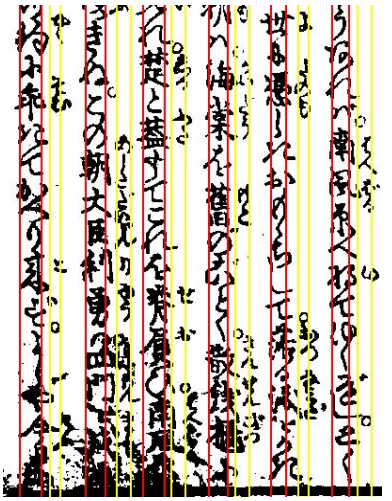

図 13: 垂直射影分布による行幹の抽出結果

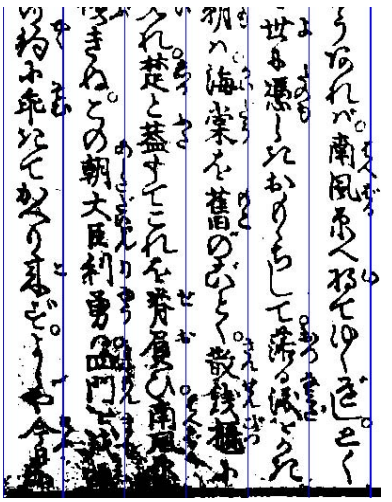

図 14: 行の抽出結果

4.2　行単位での２値化再処理の結果

行単位での判別分析法による 2 値化の再処理の結果を図 15 に示す。行単位での再処理の結果、図 15 を図 15 を比べると、下部の、手垢による汚れの影響が軽減されていることが分かる。

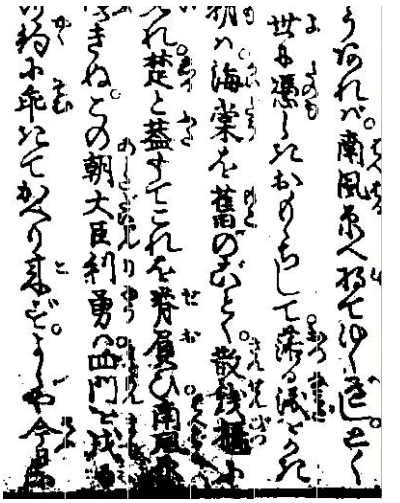

図 15: 行ごとの 2 値化処理の結果

4.3　文字候補領域の切り出し結果

ラベリング処理と矩形統合を利用した文字の切り出し処理の結果を図 16 に示す。左下の汚れ部分の影響が見られ、また、複数の文字を含んだ矩形が多く残っていることが分かる。

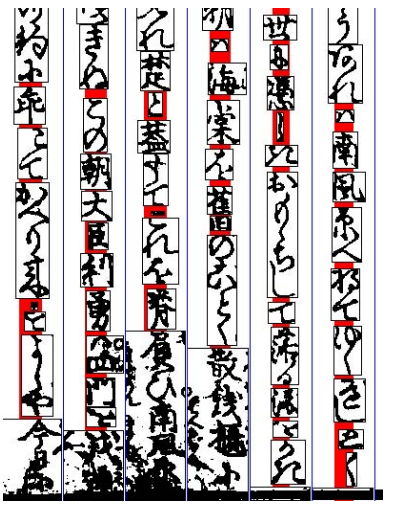

図 16: 領域ラベリングと矩形統合による処理の結果

4.4　指定面積以上の矩形領域内再処理の結果

複数の文字を含んでいると思われる、閾値以上の面積を持つ文字矩形の再処理後の結果を図 17 に示す。図 16 と比べて、大きな文字矩形の数が減少していることが分かる。

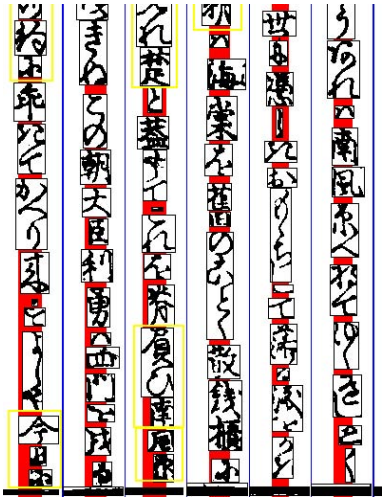

図 17: 指定面積以上の矩形領域内の処理の結果

4.5 水平射影分布を利用したつづけ字の分離処理の結果

水平射影分布を利用したつづけ字の分離処理の結果を図 18 に示す。つづけ字により複数の文字が含まれていた矩形の多くが処理を施され、大きな文字矩形がさらに減少していることが分かる。

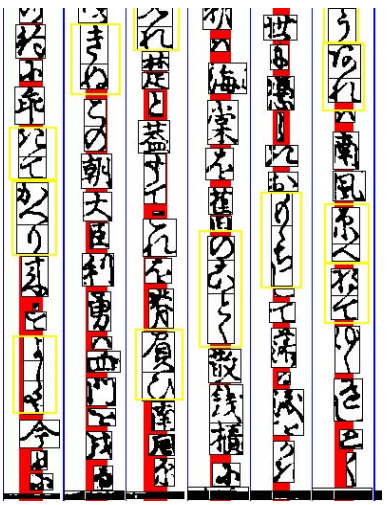

図 18: つづけ字分離処理の結果

4.6 小矩形領域の再統合の処理結果

上下に分割されている、小矩形領域の再統合処理の結果を図 19 に示す。いくつかの漢字の部首や、ひらがなの画が統合されている。

図 19: 小矩形の統合処理の結果

4.7 振り仮名部分の処理結果

以上のような本文部分の処理と同様の処理を振り仮名部分についても行う。振り仮名部分の処理の結果と本文部分の処理結果を合成したものを図 20 に示す。

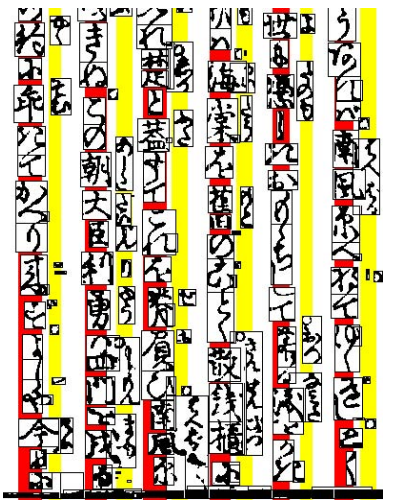

図 20: 本文分と振り仮名部分の文字切り出し結果

5 評価と考察

5.1 文字切り出し結果の評価

今回処理実験の対象とした全部で 11 ページの画像に対して処理を行った。この結果、本文部分のそれぞれの文字が正しく切り出せている文字を「成功文字」とし、それ以外のものを「失敗文字」として、それらの割合を算出した。文字切り出しの成否は [9] を参考

にして判定した。その結果を表 1 に示す。

表 1: 本文の文字切り出しの抽出率

ページ	行数	総数	成功	失敗	抽出率
No.1	7	212	162	50	76.4%
No.2	11	332	285	66	85.8%
No.3	11	330	286	61	86.7%
No.4	11	343	260	93	75.8%
No.5	11	326	265	70	81.3%
No.6	11	329	261	90	79.3%
No.7	11	343	270	98	78.7%
No.8	11	319	257	72	80.6%
No.9	11	320	251	75	78.4%
No.10	11	321	271	64	84.4%
No.11	11	319	260	59	81.5%
total	197	3494	2828	79	80.9%

5.2 考察

大津の判別分析法を順次繰り返し適用し、局所的な閾値設定を行ったため、比較的良好な文字抽出率を得ることができた。

文字切り出しの失敗した場合の例を図 21 に示す。(a)、(b) では上下に矩形が分離され、これらの矩形の統合ができていない。(c) は振り仮名が混入しており、このため、水平射影分布による分離に失敗している。

(a) (b) (c)

図 21: 文字切り出しの失敗例

このような現象を解消し、さらに抽出率を向上させる必要があると考えている。

本文部分の処理結果を利用して、さらに同様の処理により、振り仮名部分の行幹抽出および文字抽出処理もできた。しかし、振り仮名部分にはつづけ字が多く、文字の分離切り出しはいまだ不十分である。

6 むすび

本報告では、木版本のデジタル画像から文字を切り出す手法の提案とその評価について述べた。課題としては、切り出し手法、特に手垢などの汚れに影響されにくい２値化手法の検討を行うことが必要である。このため、色彩情報を用いた処理を導入して、文字の切り出しをより安定に行うことを検討する。また、分離した文字の統合、さらには、つづけ字の分離の処理ついては、より信頼性の高い手法を考案する。

一方、今後の展開としては、振り仮名の情報を積極活用して、これを元に解読の支援を行うこともできると考えている。また、切り出した文字を画像として扱い、これの出現頻度や出現ページを記したインデックスとして表現する文字字書への応用、および、それを用いた教育等への応用を検討している。

謝辞 本研究を進めるにあたり、ご指導ご助言を頂いた本学文学部の赤間亮教授に厚く御礼申し上げます。なお、本研究は、一部、文部科学省オープンリサーチセンター補助事業の支援によって行われた。」

参考文献

[1] 山田奨治, 柴山守:日本学術振興会科学研究費補助金　研究成果報告書　古文書翻刻支援システムの研究 (3), 2004.

[2] 梅田三千雄，橋本智広：認識処理を援用した文字切り出しによる古文書のキャラクタスポッティング、電気学会論文誌 C、Vol.112, No.3, pp.1876-1884, 2002.

[3] 近藤博人, 松本隆一, 柴山　守, 山田奨治, 荒木義彦：文字切出しを前提としない古文書標題認識, 情報処理学会研究報告 2003-CH-57, Vol.2003, No.5, pp.1 － 8, 2003.

[4] 寺沢憲吾, 長崎健, 川嶋稔夫：文字切り出しによらない毛筆手書き文字検索のための部分空間法，信学技報，PRMU2004-172，pp.51-56，2005.

[5] 日本色彩学会：新編　色彩科学ハンドブック［第２版］，東京大学出版会，1998.

[6] 鳥脇純一郎：画像理解のためのディジタル画像処理，昭晃堂，1996.

[7] 柴山守：古文書の文字切出しを考える，人文学と情報処理，No.18，pp.57-63，1998.

[8] 尾崎浩司，柴山守，山田奨治，荒木義彦，古文書画像の標題文字セグメンテーション，人文科学とコンピュータシンポジウム，NO.17，pp279-286、2000.

[9] 後藤丹治校注：日本古典文学大系 61「椿説弓張月　下」、岩波書店、pp.193-205、1962.