人文科学とコンピュータ研究会　47－7
（２０００．７．２１）

古文書画像のレイアウト認識と標題抽出

尾崎 浩司†　柴山 守‡　荒木 義彦†
立命館大学†　大阪市立大学‡

古文書の文字切り出し、及び文字認識の基礎的研究を行うために、古文書の標題のみを対象とした文字パターン辞書の構築と関連するユーザインターフェイスの開発を行っている。本稿では、古文書の原画像からピラミッド構造により抽象化された概略画像を抽出し、その概略画像上での標題の抽出、及びレイアウトを認識する手法について述べる。標題の抽出では、概略画像より射影ヒストグラム法、及びラベリング法の併用による手法を試みた。実験結果では、994 文書中約 78. 1%が正しく抽出された。これらの特徴と問題点について考察する。また、レイアウトの認識では、標題、本文、日付、差出人、受取人等を認識するルール、及びその実現する手法について考察する。

Layout Recognition and Title Extraction for Historical Document Images

Kouji OZAKI†　Mamoru SHIBAYAMA‡　Yoshihiko ARAKI†
Ritsumeikan University†　Osaka City University‡

As a part of an ongoing research on character segmentation and recognition for historical documents, we have been developed a character pattern dictionary focused on title of document and an user interface for segmenting characters. This paper describes the generation of outline image using the pyramid structure, extraction of title, and layout recognition for the images. In the title extraction, both histogram projection and labeling methods are used. The ratio of accuracy in extraction is estimated to be 78.1% for 994 documents. In the layout recognition, a rule for identifying title, body, date, sender, and receiver of each document, and a method for implementation is discussed.

1.はじめに

計算機技術の進歩に伴い、人文学分野においても工学的手法が取り入れられ、研究が進められている。その一つとして古文書画像のデータベース化が挙げられる。古文書画像データベースの検索においては、標題、発信人、受取人、年代などの目録を作成し、その目録より対象とする画像を検索するのが一般的である。さらに全文検索を行うには、翻刻、解題、読み下し文のテキストが必要となる。しかしながら目録作成等をすべて手作業で行うには膨大な時間と費用、専門的知識を必要とする。古文書文字の切出し、認識の研究は、それらの作業を軽減するのに大いに貢献するに違いない。

本研究は、古文書文字の切出し、及び文字認識の基礎的研究を行うために、古文書標題のみを対象とした文字パターン辞書のデータベース構築と関連するユーザインターフェイスの開発を目的にしている。

古文書の形態は縦横の長さ、大きさが一様でないため、古文書レイアウトの把握や他の古文書との比較が容易にできない。そのため古文書の概略画像をピラミッド型の上位層で抽出し、その抽出した抽象化レベルのレイアウトから標題部分だけに着目して原画像から標題部分の抽出を行う。

古文書画像は「伏見屋善兵衛文書」(大阪市立大学学術情報総合センター所蔵)の約1,300文書、2,000画像を対象にする。

2.古文書画像の抽象化

古文書の原画像をピラミッド構造により、抽象化して概略画像を得る。ピラミッド構造とは、原画像に対してピラミッドの上位層で画像を抽出する方法である。

概略画像の抽出にピラミッド構造を用いる理由は、効率的なメモリーの利用のためである。古文書画像は大きさが一様でなく、大きなものでは場合によってはメモリーに蓄えられない。

ピラミッド構造の場合、メモリーに蓄える量は概略画像分だけで済み、幾何学的変換による縮小処理と比較して少ないメモリーで処理することができる。

概略画像を抽出する理由は

(1) 縦または横に長い古文書画像のレイアウトの把握

(2) 文字列の位置関係、様式、形態の把握

(3) レイアウト特徴による文書の分類

が容易にできるためである。

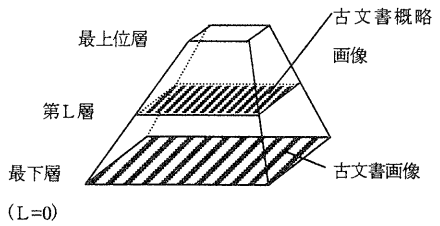


図1 ピラミッド型抽出

ピラミッド構造を用いて抽出された画像は、カラー画像である。よってYIQ変換による濃淡化処理を行い、2値化処理を行う。

3.射影ヒストグラム法による標題抽出

3-1 ヒストグラム

つぎに概略画像からの行、及び文字列の抽出の概要を図2に示す。

抽出された概略画像($m\times n$)より垂直射影ヒストグラムv_iをとる。v_iは、

$$v_i=\sum_{j=0}^{n-1}p(i,j)\qquad(i=0,1,\cdots,m-1)\quad(3.1)$$

で表される。ここでiは概略画像の水平(x)方向位置、jは同画像の垂直(y)方向とする。

行抽出のために式(3.1)に基づく閾値を$t_c(t_c\geq 0)$とし、$v_i\geq t_c$の条件を満たす連続した変数iを縦書き1行と定め、$i_s\leq i\leq i_e$の範囲を抽出する。ここで、i_sは連続したの始点、i_eは終点を表す。ここで、抽出された各々の行をk(ただし、$k=1,2,\cdots,k_{max}$)とする。

次に得られたそれぞれの行kに対して水平射影ヒストグラムh_{kj}をとる。h_{kj}は、

$$h_{kj} = \sum_{i_s}^{i_e} p(i,j) \quad (j = 0,1,\cdots,n-1) \qquad (3.2)$$

で表される。このh_{kj}が得られた個所を文字部分として定め、連続したヒストグラムの上端を行の始端、下端を行の終端とする。

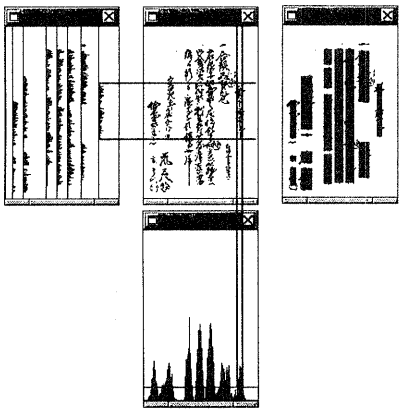


図2　ヒストグラムによる抽出範囲選択

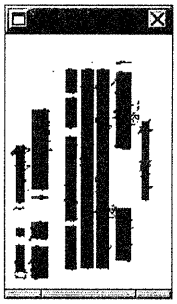
3-2　標題抽出

ヒストグラムによって抽出個所を決定したが、図3 (a) に示すように本来、標題や差出人等の意味のある文字列の一部分で空白が出来ているため、このままでは文字列として抽出できない。そのために必要に応じて補間操作をすることとした。

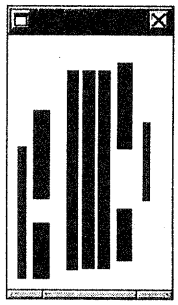
しかし、行内に切れ目が生じる場合の空白の間隔は理論的に解決する手法はない。したがって経験に基づき概略画像上において一定画素以内の間隔の場合、補間を行い連続させる。この結果を図3 (b) に示す。

次に補間した抽出範囲から標題を抽出する際のルールは、①文書の最右端の行を標題と仮定する。

標題の抽出方法は、②最右端の行より抽出個所の矩形4隅の座標を概略画像上で取得する。③その座標を概略画像から原画像用に座標変換を行い、④原画像の標題部分のみを読み取り、抽出する。ただし、抽出した標題は、概略画像と同様に2値化処理を行う。図4に抽出結果を示す。



(a)補間前



(b)補間後

図3　文字列の抽出個所

図4　標題抽出結果

3-3　実験結果

全画像 1987 枚に対して標題が抽出できたのは712 枚、全体に対して抽出できた割合は36%である。しかし抽出できなかった画像数の中には封筒、裏書など元々標題が存在しない画像が993枚含まれている。それらを全体から除き、標題が存在している画像だけで考えると抽出できなかった画像は282 枚である。よって標題が存在する画像だけで考えると、標題が存在する画像994枚に対して、標題が抽出できたのは712 枚であり、72%の割合という結果が得られた。

3-4　射影ヒストグラム法による行抽出の問題点

射影ヒストグラム法による行抽出の問題点は、第1に文字の一部が削れることである。垂直射影ヒストグラムでの閾値により、文字の一部分が欠ける。標題部分 (文字列) としては認識できるが、抽出した文字列に対して文字認識を行う場合、文字の削れで正しい認識ができない。

第2に、文字列が傾いている場合、文字列の始端及び終端部分の垂直射影ヒストグラムの値が低くなり、文字列の始端及び終端の文字の一部が削れる場合がある。また、行間が狭い場合には、垂直射影ヒストグラム上において文字列の終端と隣の文字列の始端部分が重なってしまい、行間で分割する事が困難である。

第3に、図5に示すように隣の文字列からの侵入（図5左側：「申」の左側の印影、図5右側：「舞」の左側の印影）の影響がある。垂直射影ヒストグラムによって文字列と定めた範囲に、隣の文字列の文字の一部が侵入している場合、その侵入している文字の一部も抽出される。

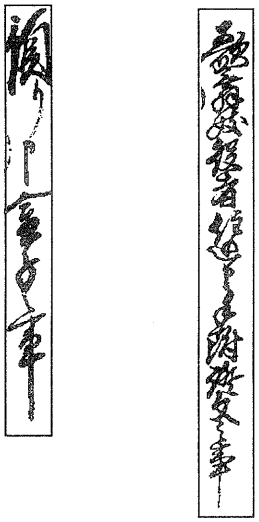


図5　隣接文字の侵入例

4.射影ヒストグラム法とラベリング法による標題抽出

4-1 ラベリング法による標題抽出

次に射影ヒストグラムの問題点を改善するためにラベリング法の併用を考える。

概略画像よりラベリング法を用いて標題を抽出する。ラベリング法の利用は黒色、つまり文字部分を1つの塊としてみることができるため、前章示した射影ヒストグラム法による行抽出の問題点が解決できる。

4-1-1　前処理

1）結合処理

概略画像に対してそのままラベリングを行うと偏と旁、文字と文字がそれぞれ離れたいた場合や文字にかすれがある場合に、1つの文字、行として抽出することが難しい。この手法は柴山［1］が行った実験でも示されている。したがって、偏や旁、文字と文字など抽出した意味のある文字列を1つの塊として把握するために、以下のルールに基づく塗りつぶしによって文字間の接続を行う処理（以下、結合処理という）を行う。

しかし、x 軸及び y 軸方向すべての方向に結合処理を行うとある行に位置する文字の一部が隣接する行への侵入が存在する場合に、行間の結合処理が行われることになり、隣接する2つの行が1つの行として認識される。これを防ぐために、垂直射影ヒストグラムによる一定の閾値以上の範囲を目安として結合処理を施すことにした。

2）ノイズ除去

結合処理を行った後に、ノイズ除去を行う。ノイズ除去は窓サイズ 3×3 のフィルタ窓の中央の値が 1（黒）であり、その他の窓の値が 0（白）である場合、中央の値を0に変換する。この操作を画像全体について行う。

4-1-2　ラベリング法

前処理を行った画像に対してラベリング処理を行う。ラベリング法とは連結している全ての画素に対して同じラベル（番号）を付け、異なった連結成分には異なったラベルを付ける処理である。ラベル付けを行うと同時に各ラベル（連結成分）のラベル枠

$$q_n = (i_{\min}, j_{\min}, i_{\max}, j_{\max}) \qquad (4.1)$$

n ：ラベル番号$(n = 1,2,\cdots m)$

も求める。

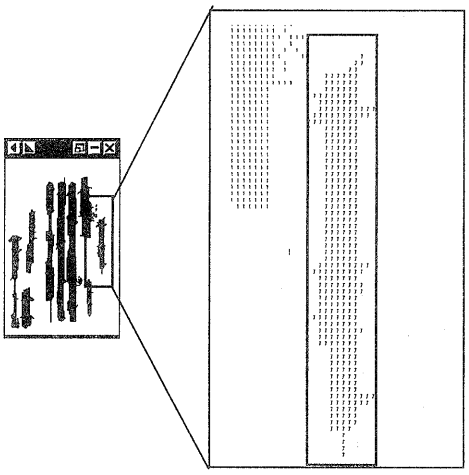


図6　ラベリング処理画像とラベル枠

4-1-3　分類

今回は標題のみを対象としているが、今後日付、差出人、受取人などに関しても同様の抽出を前提にしている。それを考慮するために、本文とその他の部分（標題、日付、差出人、受取人）との分類を行う。

1）水平射影ヒストグラム

結合処理を行った画像に対して水平射影ヒストグラムh_jをとる。

本文部分抽出のためにh_jの閾値を$k_c(k_c \geq 0)$とし、$j=0$から順にh_jを求め、初めて$h_j \geq k_c$の条件を満たしたときのjの値をj_hとする。このとき前節「4-1-2　ラベリング法」で求めた各ラベルq_nのラベル枠$j_{\min}$に対して、$j_{\min} \geq j_h$を満たすとき、そのラベルは本文とみなす。

これは標題、日付、差出人、受取人の部分の書き始めの位置が本文より下であり、また本文の書き始めの位置は各行ともほぼ同じであるという仮定から定める。

本文を抽出した後、その他のラベルより標題を抽出する際のルールとして、最右端に存在するラベルを標題とする。

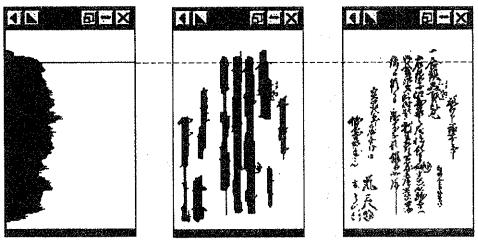


図7　水平射影ヒストグラムを用いた分類

つぎに標題の抽出方法は、最右端のラベル番号よりその連結成分の座標を概略画像上で読む。ラベル枠内には他のラベル番号が侵入してきている場合があるが、最右端のラベル番号のみを読み取るため、他の連結成分を抽出する事はない。その座標を原画像に合うように変換を行い、原画像の標題部分のみを読み取り、抽出する。ただし、抽出した標題は概略画像と同様に2値化処理を行う。

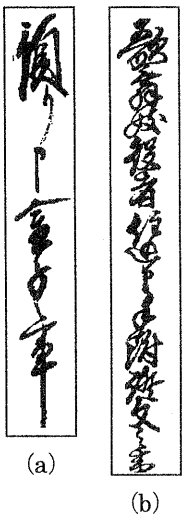
4-1-4　実験結果

1）標題抽出例

図8（a）（b）は前節「3-4　射影ヒストグラム法による行抽出の問題点」で示した隣接文字の侵入に関する問題に対して、隣接文字の侵入を抽出することなく標題のみを抽出できた。

2）標題抽出不可例

図9は結合処理の際、垂直ヒストグラムの閾値が固定値によるために、標題文字が閾値以下になり結合処理が実行されなかったため、文字の一部しか抽出できていない。



(a)　(b)

図8　標題の抽出例

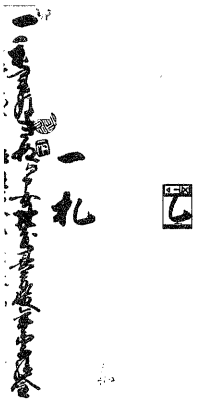


図9　標題抽出不可例

4-2　ラベル枠を用いた抽出

ラベリング法のみでは前節「4-1-4 実験結果」図9のように標題の一部分のみ抽出されてしまう場合が生じる。そのためにラベリング法による抽出方法にラベル枠による抽出方法を合わせて標題の抽出を行う方法が考えられる。これは例えば左右に分離した文字の一部分に外接する矩形を描き、その矩形が一部重なるような場合、同一ラベルを与え1つの文字と見直す手法である［2］、［3］。

4-2-1　文字セグメントルール

ラベル番号が$n1, n2$となる連結成分が存在するとき、そのラベル枠をそれぞれ前節「4-1-2 ラベリング法」の式（4.1）より

$$q_{n1} = (i_{n1\min}, j_{n1\min}, i_{n1\max}, j_{n1\max}) \quad (4.2)$$

$$q_{n2} = (i_{n2\min}, j_{n2\min}, i_{n2\max}, j_{n2\max})$$

とする。ただし$n1 < n2$とする。

このとき、q_{n1}に対してq_{n2}が以下の 3 つの条件を満たすとき、$n2$のラベルを$n1$に変換する。

$$j_{n1\min} \le j_{n2\min} \le j_{n1\max} \quad (4.3)$$

且つ

$$i_{n1\min} \le i_{n2\max} \le i_{n1\max} \quad (4.4)$$

且つ

$$i_{n2\max} - i_{n1\min} \ge (i_{n2\max} - i_{n2\min})/2 \quad (4.5)$$

上記の各々は、図 10（a）（b）（c）に対応する。

または

$$j_{n1\min} \le j_{n2\min} \le j_{n1\max} \quad (4.3)$$

且つ

$$i_{n1\min} \le i_{n2\min} \le i_{n1\max} \quad (4.6)$$

且つ

$$i_{n1\max} - i_{n2\min} \ge (i_{n2\max} - i_{n2\min})/2 \quad (4.7)$$

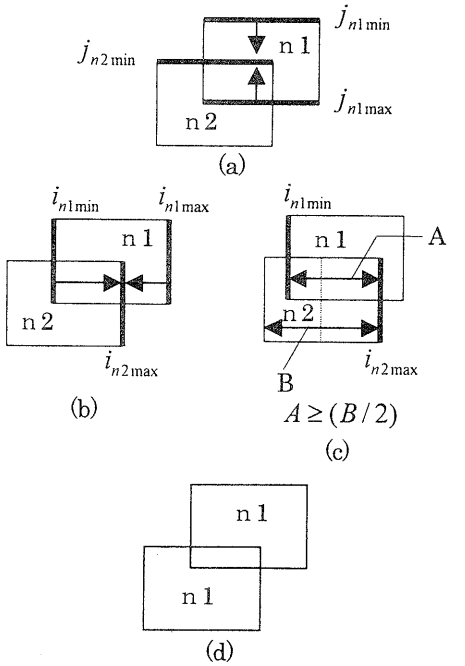


図 10　文字切出しルール

ここで、式（4.5）の場合、n 1 の左端と n 2 の右端の距離（図 10（c）参照）をA、n 2 の左端と右端の距離を B とする。条件$A \ge (B/2)$を満たすとき、つまりx方向に対して$n2$の領域が 1/2 以上$n1$に含まれるとき、ラベル変換を行う。

以上の手法を全ラベルに対して行う。

4-2-2　実験結果

標題が存在する画像 994 枚のうち、前章で述べた射影ヒストグラムによる標題抽出によって標題が抽出できなかった 282 枚を対象に、ラベリング法による標題抽出を行った。

その結果 282 枚のうち 64 枚に関して標題を抽出する事ができた。

1）標題抽出例

前節「4-1-4 実験結果　2 標題抽出不可例」で示した古文書に対して、ラベリング法による標題抽出では標題の一部だけ抽出されていたのに対し、ラベル枠との併用による標題抽出では、正しく抽出できた。（図 10）

図 11　標題の抽出例

5. レイアウト認識

古文書画像において、標題、本文、日付、差出人、受取人等を認識するルール、及びその実現す

る手法について提案する。

5-1 行の定義

前章「4-2-1 文字切出しルール」で求めたラベル枠を用いてそのラベル枠の左上の座標を $(i_{\min}, j_{\min})$、右下の座標を $(i_{\max}, j_{\max})$ とし、それによって求められる行 Q_n を

$$Q_n = (i_{\min}, j_{\min}, i_{\max}, j_{\max}) \qquad (6.1)$$

n ：ラベル番号

と定める。

5-2 認識ルール

各々のレイアウトを

注釈1（標題より右側上部にある行）：C1

注釈2（標題より右側下部にある行）：C2

標題：T、　本文：B、　日付：D、

差出人：S、　受取人：R、　追記：P

とする。これを要素という。

また、概略画像の水平射影ヒストグラムをとり、その上端と下端の中心を Y_2、上端と Y_2 の中心を Y_1、Y_2 と下端の中心を Y_3 とする（図12）。

ここで、$Y_1=h/4$、$Y_2=Y_1+h/4$、$Y_3=Y_2+h/4$ である。ただし、水平射影ヒストグラムの上端と下端の距離をh、原点は左上隅とする。

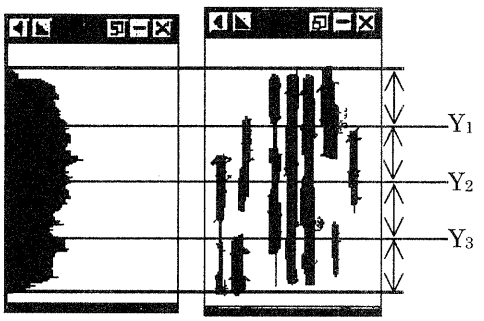


図12　レイアウト認識基準

以下のルールに基づきレイアウトを決定する（図13）。

1）注釈1（C1）、注釈2（C2）

ラベル枠の座標とレイアウト分割ルールにおいて、

$j_{\max} \le Y_1$ のとき、$Q_n = C1$

$Y_3 \le j_{\min}$ のとき、$Q_n = C2$

とする。

2）標題（T）

$Q_n = C1$、$Q_n = C2$、の $i_{\min}$ をそれぞれ $i_{C1\min}$、$i_{C2\min}$、$Q_n = C1$、$Q_n = C2$、を除く他の行 Q_n の $i_{\min}$ を $i_{O\min}$ とすると、

$i_{\min} \le i_{C1\min}$、または $i_{\min} \le i_{C2\min}$

かつ

$i_{\min} \ge i_{O\min}$

のとき $Q_n = T$ とする。

3）本文（B）

$j_{\min} \le Y_1$ かつ $Y_3 \le j_{\max}$ のとき、$Q_n = B$

とする。

4）日付（D）

$j_{\min} \le Y_1$ かつ $Y_1 \le j_{\max} \le Y_2$ のとき、$Q_n = D$

とする。

5）差出人（S）

$Y_2 \le j_{\min} \le Y_3$ かつ $Y_3 \le j_{\max}$ のとき、$Q_n =$ S

とする。

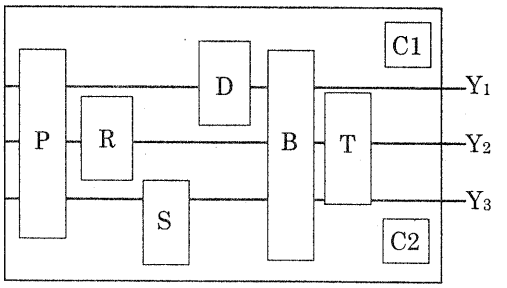


図13　認識基準と要素配置の関係

6）受取人（R）

$Y_1 \le j_{\min} \le Y_2$ かつ $Y_2 \le j_{\max} \le Y_3$ のとき、$Q_n = R$ とする。

7）追記（P）

$Q_n=D$、$Q_n=S$、$Q_n=R$のi_{min}をそれぞれi_{Dmin}、i_{Smin}、i_{Rmin}とすると、

$i_{min} \leq i_{Dmin}$、または$i_{min} \leq i_{Smin}$、

または$i_{min} \leq i_{Rmin}$

かつ

$j_{min} \leq Y_1$ かつ $Y_3 \leq j_{max}$ のとき、$Q_n=P$

とする。

6.おわりに

古文書画像のピラミッド型によるレイアウト抽出を行い、その結果を判断し、標題の抽出を射影ヒストグラム法とラベリング法の2つの手法を用いて行った。その結果、78.1%の割合で標題抽出を行え、形式が未知である文書の分類が会話型で短時間に行えるユーサインターフェースを開発した。

しかし、印影や裏写りの影響を受けたものに対しては、本実験では解決されず、また誤って文字の一部分のみ抽出されたものもある。文字の一部分のみ抽出された文書に対する改善は、今後各閾値を一定値から各画像の画素値の分布に対して変化させた実験を行いたいと考えている。

また、古文書画像において、レイアウトを認識するルール、及びその実現する手法について考察した。現在、このレイアウト認識の実験を進めている。

なお、本研究は科学研究費基盤研究(B)「古文書OCRの試論的研究」、及び基盤研究(B)「古文書解読プロセスの知能情報学的解明」による。

参考文献

[1] 柴山 守：古文書画像の文字切出しを考える、人文学と情報処理 第18号 特集 挑戦古文書OCR、勉誠出版、1998

[2] 馬場口登、塚本正義、相原恒博：手書き日本文字列からの文字切り出しの基本的考察、電子通信学会論文誌、Vol.J68-D、No.12、1985

[3] 馬場口登、塚本正義、相原恒博：認識処理の導入による手書き文字切出しの一改良、電子通信学会論文誌、Vol.J68-D、No.11、1986

[4] 富田浩章、柴山 守、荒木義彦：2 値化レベル制御による古文書画像の文字セグメンテーションとパターン字書について、情報処理学会研究報告、96-CH-30、Vol96、No.42、pp.7-12、1996

[5] 富田浩章、柴山 守、荒木義彦：古文書画像の2 値化レベル制御による対話型文字分割とその評価、電気学会論文誌C、Vol118-C、No.4、pp.503-509、1998

[6] 井野英文、猿田和樹、加藤 寧、根本義章：ストローク情報に基づく手書き郵便宛名の切出しに関する一手法、情報処理学会論文誌、Vol.38、No.2、pp.280-288、1997

[7] 富田浩章、柴山 守、荒木義彦：古文書画像の文字セグメンテーションとツール開発、京都大学大型計算機センター第 57 回研究セミナー（1997.3.26）