古文書画像の標題文字セグメンテーション

尾崎 浩司[1] 柴山 守[2] 山田 奨治[3] 荒木 義彦[1]
[1]立命館大学 [2]大阪市立大学 [3]国際日本文化研究センター

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と、標題文字のセグメンテーションについて述べる。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。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。また、標題文字のセグメンテーションでは、文字パターン辞書を用いて、テンプレートマッチングによる切出しと認識を試みた。実験結果では、10 標題、97 文字中約 84.5%の割合でマッチングに成功した。これらの特徴と問題点についても考察する。

Title Character Segmentation for Historical Document Images

Kouji OZAKI[1] Mamoru SHIBAYAMA[2] Shoji YAMADA[3] Yoshihiko ARAKI[1]
[1]Ritsumeikan University [2]Osaka City University
[3]International Research Center for Japanese Studies

As a part of an ongoing research on character segmentation and recognition for historical documents, we have developed a character pattern dictionary focused on title of historical document and an user interface for segmenting characters. This paper describes the generation of outline image, extraction and segmentation 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. For each character segmentation and recognition using the template matching, the ratio of recognition was 84.5% for 97 characters in 10 titles.

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。

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1に示す[1], [2]。

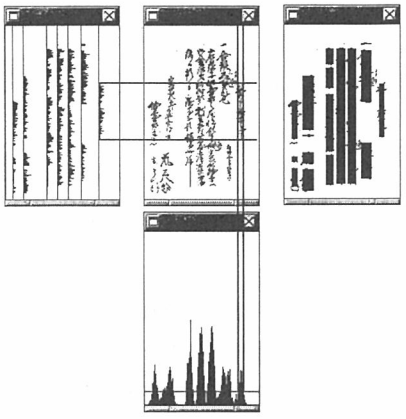
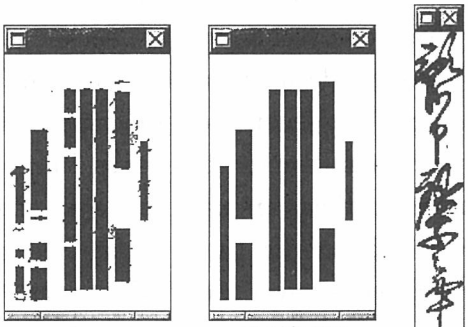


図1 ヒストグラムによる抽出範囲選択

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2（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。この結果を図2（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。図3に抽出結果を示す。



（a）補間前　（b）補間後

図2 文字列の抽出個所

図3 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、文字の削れにより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。

図 4　隣接文字の侵入例

第 3 に、図 4 に示すように隣の文字列からの侵入（図 4 左側：「申」の左側の印影、図 4 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。

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

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［3］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、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。

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。

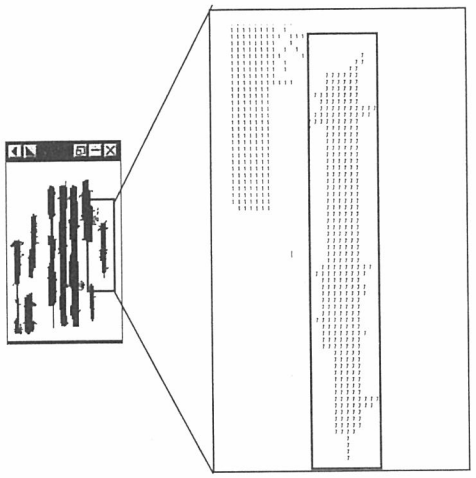


図5　ラベリング処理画像とラベル枠

4-1-3　実験結果

1）標題抽出例

図6（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7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6　標題の抽出例

図7　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7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［4］、［5］。

4-2-1　文字セグメンテーションルール

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8（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

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

ここで、式（4.5）の場合、n1の左端とn2の右端の距離（図8（c）参照）をA、n2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。

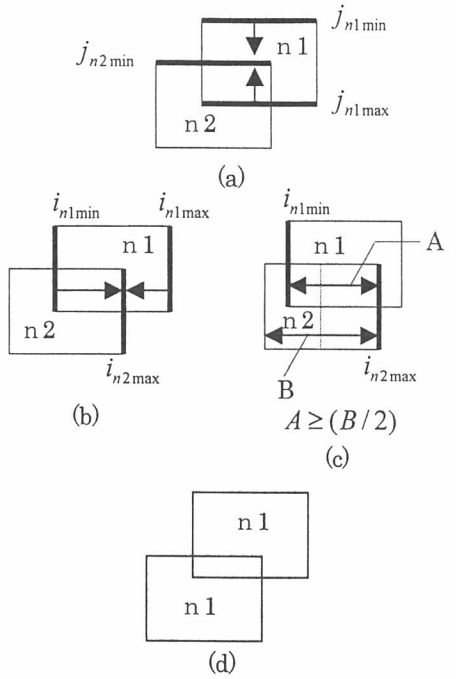


図 8　文字切出しルール

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。

図 9　標題の抽出例

・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 9）。

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 10）。

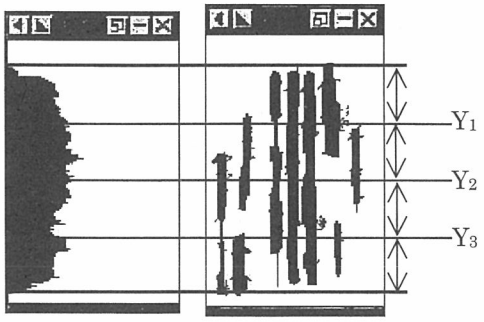


図 10　レイアウト認識基準

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。

以下のルールに基づきレイアウトを決定する（図11）。

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、

$j_{max} \leq Y_1$ のとき、$Q_n=C1$

$Y_3 \leq j_{min}$ 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_{C1min}、i_{C2min}、$Q_n=C1$、$Q_n=C2$、を除く他の行Q_nのi_{min}をi_{Omin}とすると、

$i_{min} \leq i_{C1min}$、または$i_{min} \leq i_{C2min}$

かつ

$i_{min} \geq i_{Omin}$

のとき$Q_n=T$とする。

3）本文（B）

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

とする。

4）日付（D）

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

5）差出人（S）

$Y_2 \leq j_{min} \leq Y_3$ かつ $Y_3 \leq j_{max}$ のとき、$Q_n=S$とする。

6）受取人（R）

$Y_1 \leq j_{min} \leq Y_2$ かつ $Y_2 \leq j_{max} \leq Y_3$ 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。

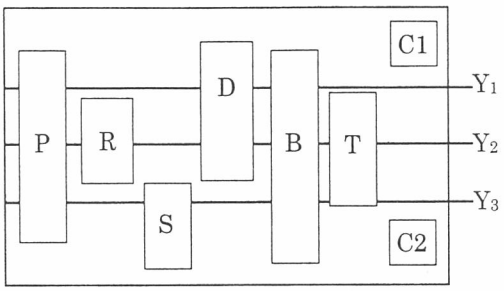


図11　認識基準と要素配置の関係

6.文字パターン辞書による文字セグメンテーションと文字認識

6-1　文字セグメンテーション方法

抽出した標題画像より、文字パターン辞書を用いて文字セグメンテーションを行う。文字パターン辞書とは、標題画像の各文字を一文字ずつに分割し、その文字がどの標題の文字で、どのような文字であるかをまとめたデータベースである。

抽出した標題画像に対して、標題の先頭文字から文字パターン辞書の各文字を用いたテンプレートマッチングを行い、マッチングした場合、その文字を標題画像から切出し、その次の文字に対しても同様にマッチングと文字セグメンテーションを行う（図12）。

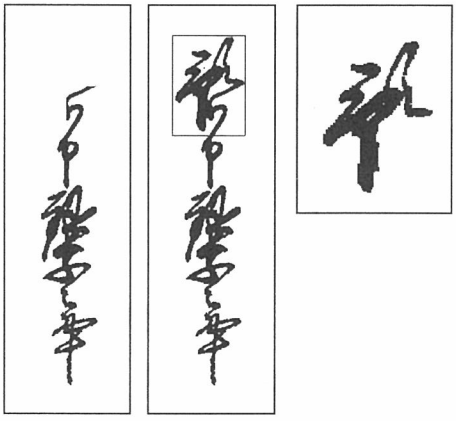
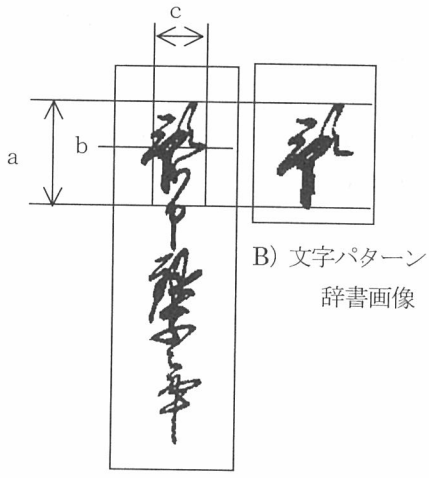


図12　文字切出しと文字認識

6-2　正規化

抽出した標題画像の各文字の大きさと、文字パターン辞書の文字の大きさは異なる。そのため、文字パターン辞書の文字を標題の文字の大きさに合うように、正規化を行う。

正規化を行う方法は、文字パターン辞書の文字の高さを基準とし、その範囲内において、標題画像の水平方向の文字幅を検出する。その中で最も大きい幅を最大文字幅とする（図 13）。次に文字パターン辞書に対しても同様に、水平方向の文字幅を検出し、更に最大文字幅を検出する。その標題画像側の最大文字幅と文字パターン辞書側の最大文字幅の長さが同じになるように、文字パターン辞書を拡大、または縮小する。



a ：文字パターン辞書の文字の高さ
b ：a の範囲内での最大文字幅位置
c ：a の範囲内での最大文字幅

図 13　文字パターン辞書の正規化

6-3　n－gram による文字選択

文字パターン辞書から文字を選択する際の知識ベースとして、標題文字の n 文字が隣接して生じる共起関係（n－gram、ここでは 2－gram を採用した）を調べ、これを文字セグメンテーションに用いた［8］。

1 文字のセグメンテーションが終了し、次の文字に対してマッチングを行う際は、2－gram の情報を用いて、次に出現すべき文字の推定を行い、この結果に基づいて、文字パターン辞書とのマッチングを行った。

6-4　実験結果

今回は、抽出した標題画像の中から、10 標題を選択し、切出しと認識の実験を行った。また文字パターン辞書に関しても、この 10 標題に対応したものを使用した。

2－gram により推定された文字郡の中で、それぞれマッチングを行い、残差割合が最も小さいものをマッチングしたものとする。そのマッチングした文字と、マッチングによって切出された標題文字とが同じ文字の場合、マッチングに成功したものとする。

その結果、10 標題 97 文字のうち、82 文字マッチングに成功した。図 14 に実験結果例を示す。

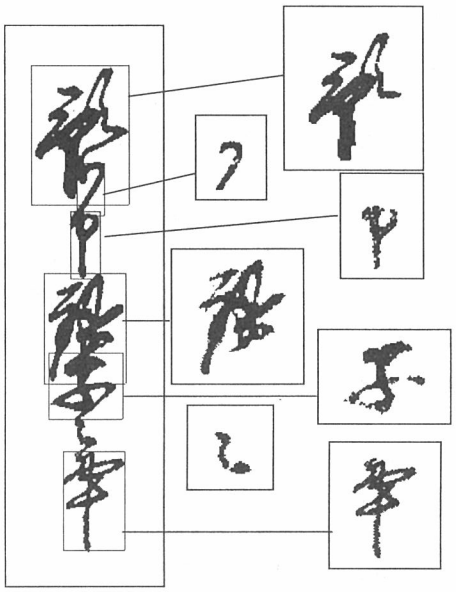


図 14　実験結果例と文字パターン辞書

図 14 に関して、「之」の部分はマッチングできなかった。これは標題画像側の「之」が小さく、また文字幅も狭いため、正規化が正しく出来なかったためである。また、今回の実験では「之」の文字に対してマッチングを行う際、2－gram により「之」の次に出現する語も文字選択の候補に入れマッチングを行った。そのために、「之」でマッチングせずに、次に出現する「事」でマッチングした。

その他に今回の実験での問題点は、標題画像か

らマッチングした個所を削除する際に、ノイズが発生し、その後の文字のマッチングに影響を及ぼしてしまうことである。また、異なった文字でマッチングをした場合は、2−gram による文字選択が有効とならない点である。

7.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。

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。

文字パターン辞書による標題文字セグメンテーションに関しては 10 標題に対して 84.5%の割合でマッチングできた。今後、他の標題画像にも同じ実験を行い、また、ニューロによる文字認識も行う予定である。

参考文献

[1] 尾崎浩司、柴山 守、荒木義彦：古文書レイアウト画像のピラミッド型抽象化と標題の自動抽出、平成 11 年電気関係学会関西支部連合大会講演論文集 G12-6、1999

[2] 尾崎浩司、柴山 守、荒木義彦：古文書画像のレイアウト認識と標題抽出、情報処理学会研究報告「人文科学とコンピュータ研究会」2000-CH-47、Vol.2000、No.67、pp.47-54、2000

[3] 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

[4] 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

[5] 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

[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）

[8] 山田奨治：n−gram による『伏見屋文書』標題翻刻支援の検討、古文書 OCR 全体会議資料（2000.8.7）