人文科学とコンピュータ　30－2
（１９９６．５．１７）

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

富田浩章*　柴山守**　荒木義彦*
(*立命館大学 **大阪市立大学)

ワークステーション(WS)上での古文書のビデオ静止画像における画像処理による文字のセグメンテーションとパターン字書作成の手法を提案する。古文書では続け字が多く、文字毎のパターン字書は作成することが困難である。そこで画像処理、特に2値化処理により文字の特徴、ならびに文字の分割を明確にし、文字毎のパターン字書を得る。また、2値化レベルの変動にともなうセグメンテーションの変化についても検討、考察する。

CHARACTER SEGMENTATION USING BINARY LEVEL CONTROL AND PATTERN DICTIONARY IN THE "KOMONJO"

Hiroaki TOMITA*,Mamoru SIBAYAMA**, Yoshihiko ARAKI*
* Ritsumeikan University　　** Osaka City University

*1916,Noji-cho,Kusatsu,Shiga 525,Japan.
**3-3-138,Sugimoto,Sumiyoshi-ku,Osaka 558,Japan.

We propose a method of character segmentation and pattern dictionary using image processing in a static video image of " Komonjo " on the workstation. In the " Komonjo " there are many characters connecting with others, so it is difficult to construct a pattern dictionary of every character. So using binary level control, characteristics of character and character segmentation become clear and we get the pattern dictionary of every character. We examine segment variation by a change in binary level.

1 はじめに

文字認識の研究は数多くなされ、数々の成果が上げられた。しかしながら、その対象の多くはオンライン入力における口語体の文字であり、古文書を対象とした研究はみうけられない。本研究では、手書き文字による古文書の文字認識、文字検索を考え、その有効な手法としてはパターン辞書の作成が挙げられる。本報告では、古文書の静止画像をもちいた2値化レベル制御による文字セグメンテーション、パターン字書について説明する。

2 古文書

古文書の特徴として次の点が考えられる。まず、墨を用いた毛筆体で書かれており、つぎに、くずし字や続け字であることが多い。さらに、同じ文字でも大きさや形状が異なるといった点である。

本報告では古文書の特徴の一つである墨を用いた毛筆体という点から、筆の筆圧、つまり筆が移動する際の墨の濃度に注目した。画像処理によって濃度のレベル制御をおこない、古文書内の墨の薄い部分は消える。そして、結果として文字間は分割されやすくなると考えられる。

今回の実験に使用した古文書（図1）は、江戸時代に書かれたものであり、これは日本歴史学会の演習古文書選の中でシミの多いものや印のあるものは除外したものから無作為に選んだ一つである。

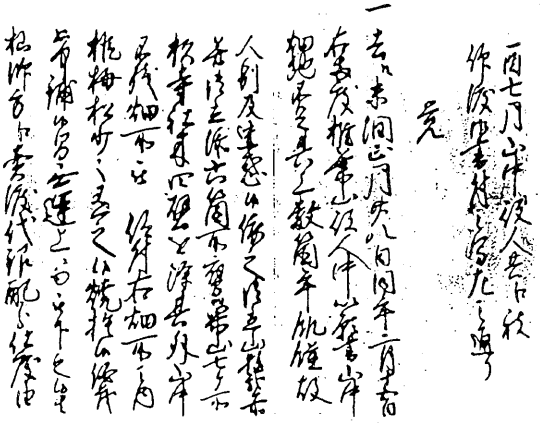


図1:古文書の原画像

3 実験

3.1 システム

本実験では、Sonyのワークステーション NWS-5000を使用し、入力画像は、ビデオカメラで撮影された古文書のカラー画像をxvmap [1]で取り込んだカラー静止画像とした。この画像のサイズは、320 * 240 画素であり、1画素あたりRGB各5bitとなっている。そして、この入力画像に対して以下に述べる画像処理を施す。

[1]NEWS X Window System 上で提供されるビデオマップ制御コマンド

3.2 画像処理

いくつかの画像処理が用いられるが、これはセグメンテーションを容易にするために必要であり、濃淡化処理、2 値化処理、細線化処理、そしてヒストグラム処理といったものが含まれている。

3.2.1 濃淡化処理

入力画像は、カラー画像である。そこで、濃淡化処理によってカラー画像から 32 階調の白黒画像へ変換される。この変換には YIQ 変換を使用し、具体的な変換としては明度を示す Y のみを採用した。この処理は、次におこなう 2 値化処理におけるレベル制御の範囲を限定するためにおこなわれた。

$$\begin{pmatrix} Y \\ I \\ Q \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.274 & -0.322 \\ 0.211 & -0.522 & 0.311 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{1}$$

3.2.2 2 値化処理

2 値化処理では、まず閾値 (l) を設定する。この閾値を 2 値化レベルとし、2 値化レベルと入力画像内の画素のレベルとを比較し、その値を超えるか否かによって 2 値に変換するものである。本報告では、この 2 値化レベルを変動させ、それにともなう文字の分割の変移について検討する。

$$p_l(i,j) = \begin{cases} 0; & p(i,j) \geq l \\ 1; & p(i,j) < l \end{cases} \tag{2}$$

ここで、$p(i,j)$ は入力画像内の画素を表しており、この値が閾値 l より大きければ 1 に、そうでなければ 0 へと変換している。つまり画像内の値は全て 0 か 1 の 2 値で表現される (図 2)。l は先程の濃淡化処理から $0 \leq l \leq 31$ の 32 階調の範囲をとり、今回の実験では、この内の 14〜20 の 2 値化レベルを実データとした。

3.2.3 細線化処理

この処理は横井茂樹氏の 8 連結型の細線化方法を採用しており、画像内の全ての文字を線で表示させることができる。つまり筆の太さをペンの太さにすることができる（図 3)。図からもわかるように文字の特徴だけでなく文字間も明確に示している。

3.2.4 ヒストグラム処理

入力画像はこれまでの処理によって、2 値化された細線化画像となってる。この画像にヒストグラム処理をおこなう。つまり、画像内で文字を示す 1 の累計をとる。

$$f(i) = \sum_{j=0}^{n-1} p(i,j)0 \tag{3}$$

i: x 方向の範囲 $(0 \leq i < 320)$
j: y 方向の範囲 $(0 \leq j < 240)$

ここで、$f(i)$ はヒストグラム処理による累計を示し、$p(i,j)$ は各画素の値 (0,1) を示している。上式では、y 軸方向に対するヒストグラムを考えているが、今回の実験では、x,y 両軸方向のヒストグラムを考えることで 2 次元のヒストグラムを得た。レベル制御で分割できない文字については、このヒストグラム処理によるヒストグラム制御を導入した。

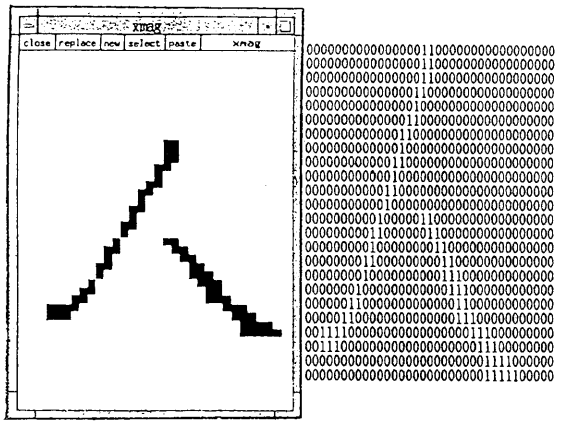
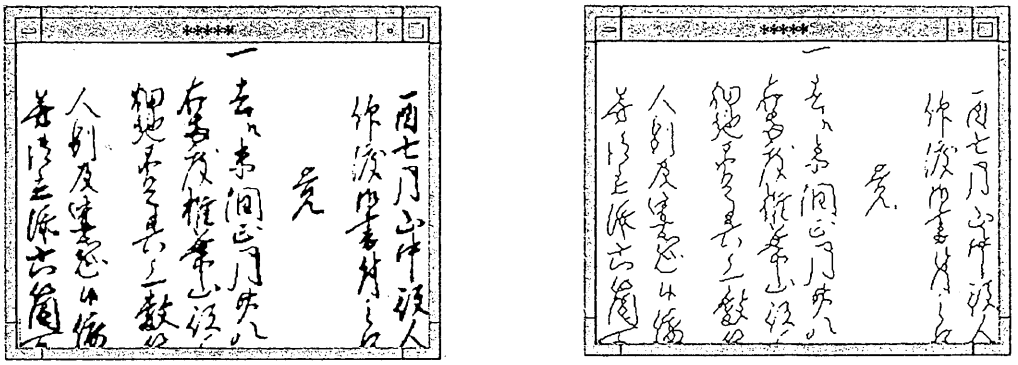


図 2:2 値化処理の文字例



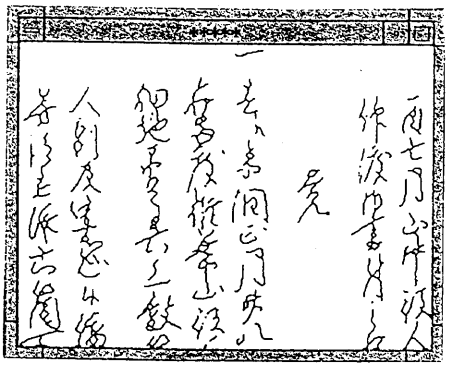
濃淡化画像　　細線化画像

図 3:細線化処理の例

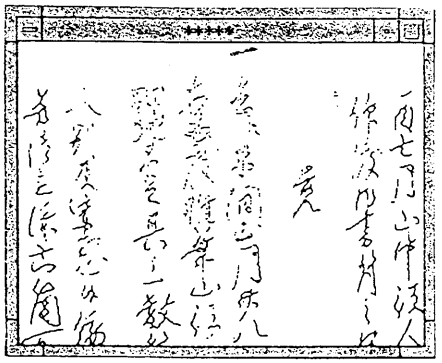
4　実験結果と考察

今回おこなった実験のデータを示す。入力された画像は、これまでに述べてきた画像変換を施され 2 値の細線化画像となっている。図 4 にはレベル 20 とレベル 14 という 2 値化レベルの異なるものを提示する。この 2 つを比べた時、レベル 14 はレベル 20 よりも文字がかすれている部分が多い。これは墨の濃度が薄くなっていることを示しており、これは文字の切り出しと直接つながっている。

図 5 には 2 値化レベル制御とヒストグラム処理による文字例を示す。レベル制御の場合、確かに文字によっては切れている場合もあり、文字例では「足」と「其」が分割されている。だが、文字が同じ筆圧で続けて書かれている場合には、レベル制御をおこなった場合に文字全体がかすむ傾向がみられる。こういった場合にヒストグラム処理を用いたセグメンテーションを考え、文字の抽出をおこなう。これをヒストグラム制御と呼ぶ。2 値化レベル制御の文字例と同じように「其」という文字が抽出できる。しかしながら、文字例中の「之」あるいは「し」、「候」といった、文字そのものが縦の線のみで構成されている場合には、文字全体がヒストグラム処理内の閾値によって消されてしまうことになる。

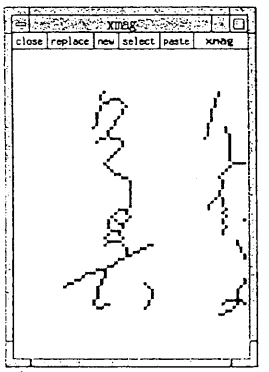


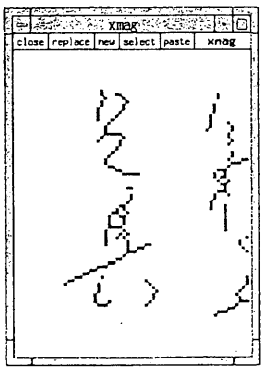
レベル 20



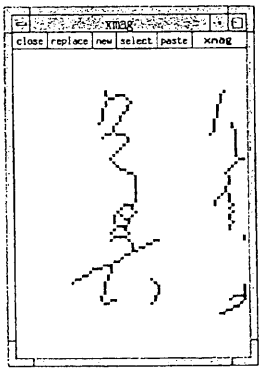
レベル 14

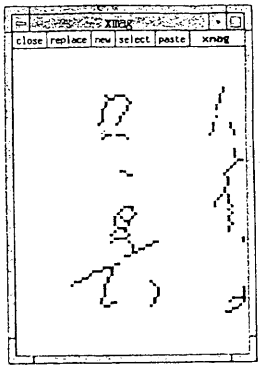
図 4:2 値化レベルの異なる画像例



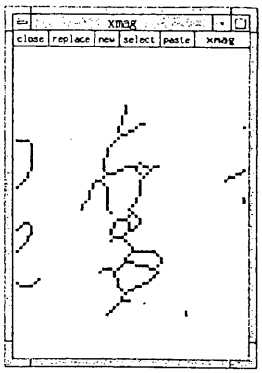


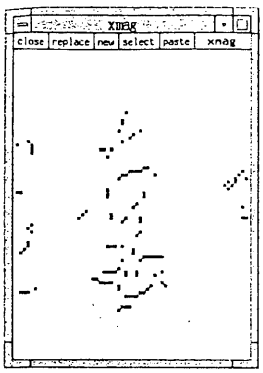
成功例





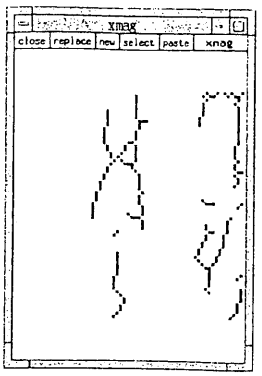
成功例

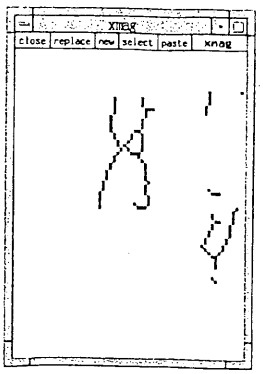




失敗例

レベル制御による文字例





失敗例

ヒストグラム制御による文字例

図 5:制御別の文字例

表 1: 文字数の比較

制 御	原画像	細線化後	セグメンテーション後 (レベル別)			
			20	18	16	14
2 値化レベル	49	29 59.2(%)	29 59.2(%)	32 65.3(%)	41 83.7(%)	44 89.8(%)
2 値化レベルとヒストグラム	49	29 59.2(%)	39 79.6(%)	43 87.8(%)	44 89.8(%)	44 89.8(%)

これまでの経緯から、実際にシステム上においてどのように文字が分割されているかを図 6 に、またパターン字書の文字例を図 7 に示す。図 6 では図 3 における左端の 1 行についてレベル 20 のセグメンテーションを表しており、文字は確かに分割されている。図 7 では文字の 2 値化の値をファイルに書き込んでいるので、結果だけを見れば 2 値化処理の文字例となんらかわりはない。しかし、タブレットなどによる手書き文字とのマッチングには有効性を発揮するものと考える。以上の結果を表 1 にまとめる。レベルが高い場合にはヒストグラム制御を加えた場合の文字数は、2 値化のそれと比べ抽出度は高い。しかしレベルを下げてみると同じ結果となり、必ずしもヒストグラム制御を行う必要はないであろう。

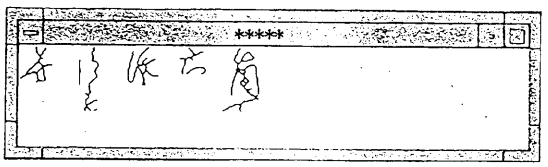


図 6:セグメンテーション例

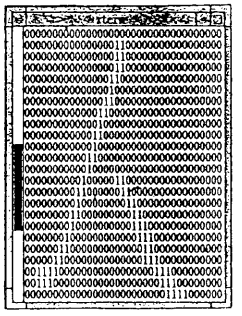


図 7:パターン字書例

5 おわりに

本報告では、2 値化レベル制御を用いた古文書の静止画像における文字セグメンテーションを行い、分割された文字についてはパターン字書を作成した。これは続け字の分割に有効な方法であると考えられる。しかし、レベルによっては分割する必要のない部分でも分割してしまう傾向もある。この点を改良し、他文書によるセグメンテーション、手書き文字による文字のマッチングについても今後検討していく。

参考文献

[1] 柴山 守: X11 による画像処理, 技術評論社,1994

[2] 長谷川, 興水, 中山, 横井: 画像処理の基本技法, 技術評論社,1985

[3] 日本歴史学会編: 演習古文書選 (続近世編), 吉川弘文館,1979

[4] 富田 浩章, 西門 秀人, 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について", 平成 7 年電気関係学会関西連合大会講演論文集 G12-47,G337,Nov.1995

[5] 柴山 守, 富田 浩章, 荒木 義彦: "ビデオによる古文書画像の入力と文字抽出について", 京都大学大型計算機センター第 52 回研究セミナー,1996