古典原本のイメージノイズ除去に関する一考察

原　　正一郎

要　旨　古典原本の電子化の障害となるイメージ雑音除去アルゴリズムの研究を行った。本研究の特色はカラー情報を利用した点にある。研究資料として白色系の和紙に黒色系の墨で書かれた古典原本を用い、イメージ処理として「朱文字」と「裏写り」の除去を試みた。研究成果の概要は以下の通りである。

1）古典原本イメージのカラー情報：古典原本のテキスト・イメージをカラー・イメージスキャナ（100dpi,赤（R）緑（G）青（B）各基本色256諧調）で取り込んだ後、各ピクセルデータをRGB表色系へ展開し、資料のカラー構造を考察した。これより、(a)多くのピクセルは直線R＝G＝Bの周辺に分布する、(b)朱文字のような特別な色彩のピクセルは(a)とは異なった位置に分布する、(c)RGBの各輝度分布は2峰性を示す傾向があること、が分かった。

2）領域の分割：「和紙」領域と「表文字」領域の分割は、上記(c)の性質より、RGBの各輝度分布に対して「判別閾値選定法」を適用することで達成できた。また(a)及び(b)の性質より「朱文字」領域の分割も可能であった。しかしこの方法では、(a)「表文字」の周辺部が脱落してカスレやすい、(b)「和紙」と「裏写り」の分離が不十分である、という問題点があった。そこで、教師情報なし判別法の一法であるクラスタ分析を適用して分離精度の向上を試みた。本法ではある程度の分離精度の改善を得たが、(a)計算コストが高い、(b)画一的な手法やパラメータの適用では多様な古典原本に対処できない、などの問題点も明確になった。

以上の研究から、古典原本のイメージ雑音処理にカラー情報の利用が有効であることが確認された。

Ⅰ．研究の目的

テキスト全体を検索対象とするフルテキスト・データベース（Full Text Data Base：全文データベース）は次世代データベース・アプリケーションの１つとして注目されている。国文学研究資料館においても、古典文学大系などの電子テキスト化を進めるなど、将来のサービス開始を目指して準備を進めつつある。ところで、古典フルテキスト・データベースのソースとなる古典原本は巻物・本などに書かれているため、文字という紙媒体用アナログデータの電子媒体用デジタルデータへの変換が、データ入力のボトルネックとなっている。一方、紙媒体上の文字を電子媒体上のコードへ変換する方法として、OCR(Optical Character Recognition:光学的文字認識）の研究が行なわれている。OCRではテキスト・イメージをイメージスキャナで光学的にコンピュータに取り込んだ後に、イメージ処理をほどこして白地の背景から黒地の文字領域を抽出し、この領域に文字認識アルゴリズムを適用して文字の判定とコード化を行う。OCRは、上質の白紙に上質の黒インクで美しく鮮やかに印刷された文字に対しては高い文字認識率を示す。しかし、国文学研究資料館が対象とする古典原本は万葉から江戸時代の写本・版本が中心であり、当時の製紙技術では、紙の質（つまり、紙面の色や光の反射など）が均一で不純物混入の少ない紙を作ることは困難であった。さらに長年の環境暴露による変色・シミ・欠損の発生、あるいは墨のニジミや裏側に書かれた文字が透けて見えるなど、OCRを利用する上で解決しなければならない問題も多い。これらの問題は、OCR本来の機能である「文字領域の抽出と認識」という点から見ると、「文字領域以外のノイズ（雑音）、をいかに取り除くか」という問題とみなせる。

このような背景から、OCRの持つイメージ処理機能と文字認識機能のうち、イメージ処理によるノイズ除去に限定した研究を進めた。イメージ処理の分野

では数多くのノイズ除去法が提案されている。これらの多くは、輝度分布やパワースペクトルなどの特徴量に注目した方法、注目点の近傍における平均化操作を基本としたものに分類されるが、古典原本に対するノイズ除去法としては有効ではなかった。その原因として、従来のOCRではイメージを白黒の輝度データとして扱い、イメージが持っているカラー情報を考慮しなかった点が挙げられる。例えば、裏写りは光が墨と紙を通過した部分であり、光が墨に反射した部分である文字本体とではカラーの構造が微妙に異なっている。したがって、カラーデータを適切な座標系にマッピングすれば、文字本体と裏写りを区別することが可能であると考えられる。このような発想からイメージノイズの抽出にカラー情報を導入した点が本研究の特色である。

今回は、研究の第一歩ということで、対象を白色系の和紙に黒色系の墨で書かれた古典原本を資料とし、イメージ処理の対象も「朱文字」と「裏写り」の除去に限定した。具体的には、

1）カラー情報の特徴の解析

2）多変量解析の手法を適用したカラー情報パターン判別法

の検討を行った。

Ⅱ．研究の方法

（1）研究資料

研究に用いた資料用テキスト・イメージを図1に示す。本資料は国文学研究資料館蔵の「冥報記」から引用した。この資料は白色系の和紙上に黒色系の墨で文字が書かれているが、部分的に朱文字が加えられている。一方、和紙は比較的保存が良く、破れやシミは殆ど無いものの、裏写りが部分的に見られる。このような特徴は今回の研究材料としては最適であった。

（2）実験装置

今回用いた実験系を図2に示す。実験系は、

1）パーソナルコンピュータ（EPSON PC-386）とカラー・イメージスキャナ（EPSON GT-6000）を中心としたイメージ入力サブシステム、

2）ワークステーション（SUN SPARC station 2）を中心とするイメージ処理サブシステム、

3）Mac（Quarda 950）とカラープリンタ（Cannon Color Laser Copier 300：PIXEL-EPO）を中心とするイメージ出力用サブシステム

から構成される。各サブシステムはEthernetで接続され、FTP（File Transfer Protocol）によりデータ転送を行った。

（3）予備的分析

第Ⅲ章では研究対象の特性を分析するために、テキスト・イメージを構成する全ピクセルについてのR（赤）、G（緑）、B（青）ごとの輝度ヒストグラムを作成した。同様に各ピクセルを輝度ベクトル（R、G、B）として3次元座標上に展開し、その分布特性も観察した。

（4）判別閾値選択法による領域分割

表文字や朱文字の輝度や色度（R、G、Bの混合比）等のカラー特性は作品や頁ごとに異なる。カラー・イメージスキャナで取り込まれたテキスト・イメージごとに、対象領域とそうでない領域のサンプルをコンピュータに指示すれば、それらを教師信号として効率的なノイズ除去が可能である。しかしOCRの自動化という観点からすれば、当然とみなせる知識あるいは目的（この場合では、黒い墨で書かれた文字の部分を抽出するということ）以外の事前情報をシステムに与えることは適切ではない。

第Ⅳ章では、事前情報を利用せずにイメージの領域分割を行う方法として判

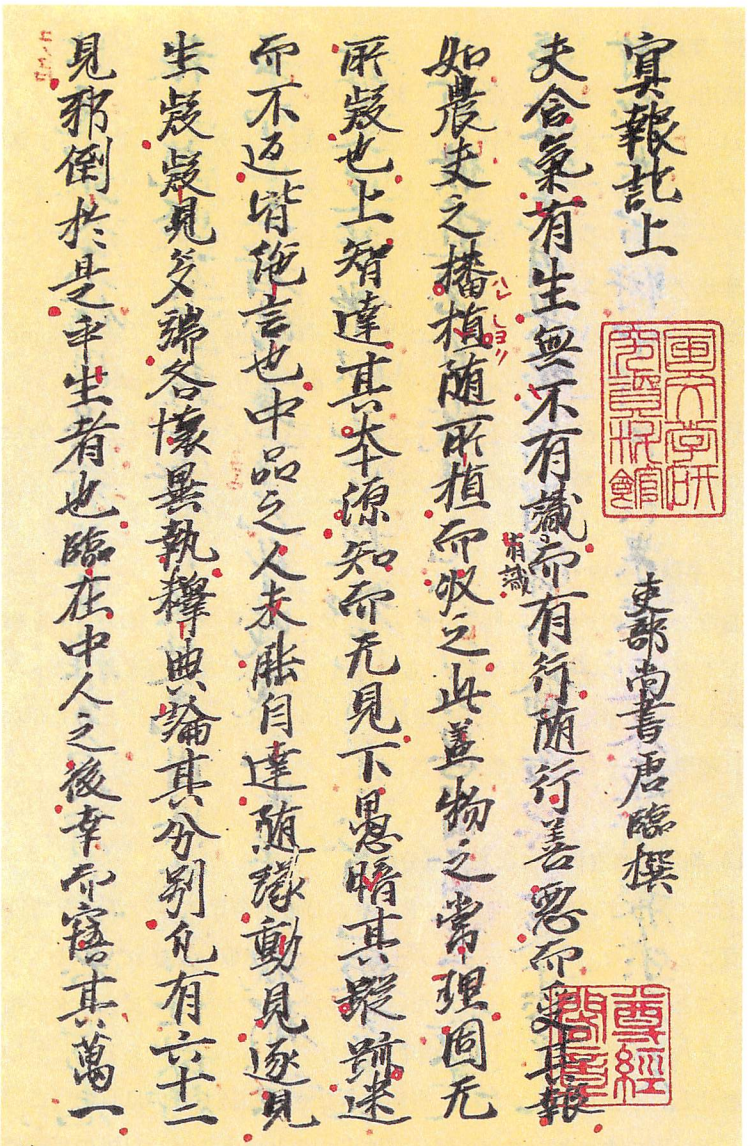


図１．研究資料

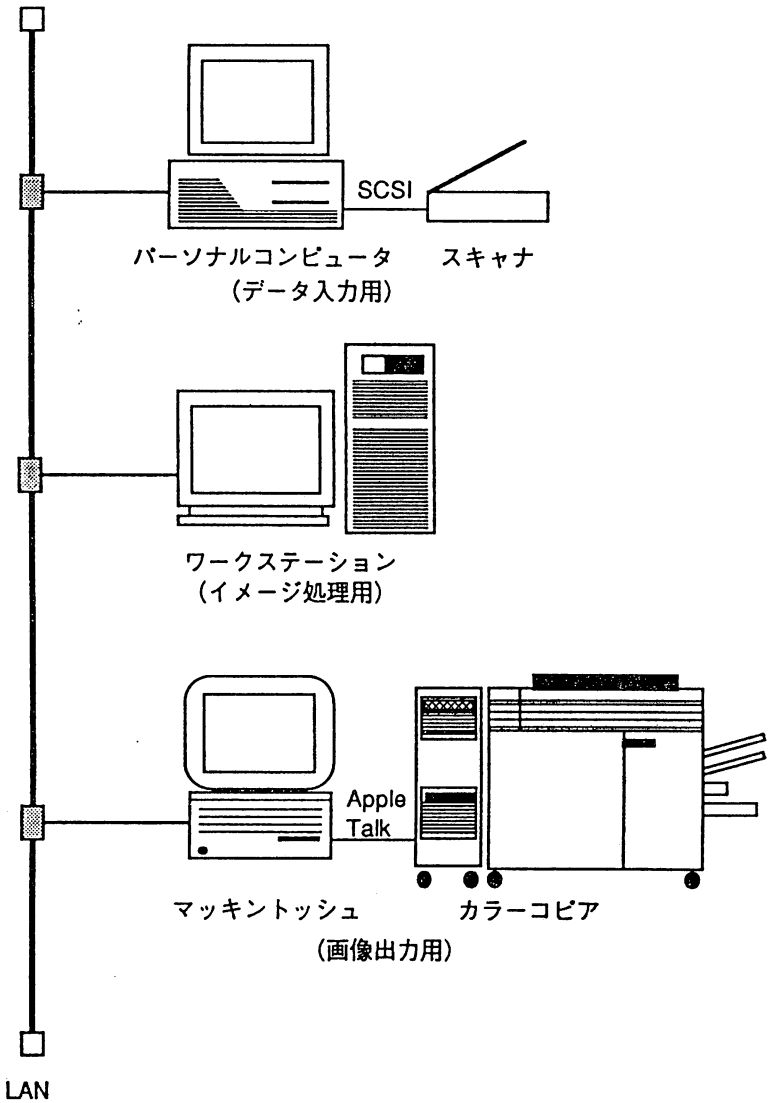


図２．実験システムの概要

別閾値選択法[1]の適用を試みた。

（5）クラスタ分析法による領域分割

第Ⅳ章の判別閾値法では、ピクセルの特性を、イメージ平面上の座標と白黒濃淡との間の関数として捉えたが、本来は、平面上の座標とカラーベクトル（R、G、B）との間の関数として捉える必要がある。さらに、テキスト・イメージは表文字、裏写り、朱文字、和紙など複数の領域から構成されているため、判別閾値選択法を2値化から多値化へ拡張するとともに、クラス数の推定を行う必要がある。これらの拡張は判別閾値選択法においても可能である。しかし、より直接的な手法として、第Ⅴ章ではピクセルの輝度分布を輝度ベクトル（R、G、B）とみなしたクラスタ分析[2]を試みた。

Ⅲ．予備的考察

この論文で利用する資料は、図1のテキスト・イメージを標本点数496（タテ）×877（ヨコ）、分解能100DPI（Dot／Inch）、濃淡の量子化レベル数をR、G、B各基本色成分ごとに8ビット（256諧調）でデジタル化したものである。このように取り込んだイメージ・データを単純な16諧調白黒濃淡イメージとしてワークステーションのウィンドウ上に表示してみると図3のようになる。なお、ウィンドウ用イメージ表示プログラムでは、座標系の原点がウィンドウの左下であるのに対し、カラー・イメージスキャナ用データ入力プログラムではイメージの左上が原点となっている。このため、図3以降におけるウィンドウ・イメージは上下が反対になっている。

図3をみると、白黒濃淡イメージでも輝度の差が著明な表文字領域と和紙領域を分割することが可能であると予想される。反対に、輝度の差が著明でない

和紙領域と裏写り領域を分離することが困難であることも想像に難くない。これらの点を明確にするために、各ピクセルの輝度をR、G、Bの各基本色成分ごとにヒストグラムとして表したものを図4に示す。なお、図の上段はR、中段はG、下段はBを示す。また、各ヒストグラムの横軸は輝度を表しており、0（暗い）から255（明るい）まで256のレベルに量子化されている。縦軸は輝度別のピクセル数である。各ヒストグラムの輝度に関する統計量は以下の通りである。

	R（赤）	G（緑）	B（青）
平　均	76.26	69.39	53.00
分　散	7540.58	6250.31	3578.07

図4の各ヒストグラムを見ると、全ての基本色成分について輝度が30の周辺と150～200の周辺にピークを持つ2峰性を示している。これによりテキスト・イメージを比較的暗い部分と明るい部分の2つの領域に分割可能であることが想像される。この構造を詳細に調べるため、表文字・裏写り・和紙の領域から典型的な部分を目視により100ピクセルづつ抽出し、同様のヒストグラムを作成した（図5）。図4と図5を比較すると、輝度が30の周辺に分布するピクセルは表文字の領域に対応し、150～200の周辺に分布するピクセルは裏写りおよび和紙の領域に対応していることが分かる。したがって、表文字領域とそれ以外の領域を分割する閾値として、直感的には2つのピークの谷間に当たる輝度が適当であると考えられる。一方、和紙と裏写りの領域にはかなりの重複が見られる。図5のヒストグラムは典型的な領域からのサンプルに基づいているので、RとGについてのヒストグラムを見ると、閾値として利用できそうな谷間がおぼろげに見いだせる。しかし、Bについてのヒストグラムではかなり不明

図３．研究資料の白黒濃淡表示例

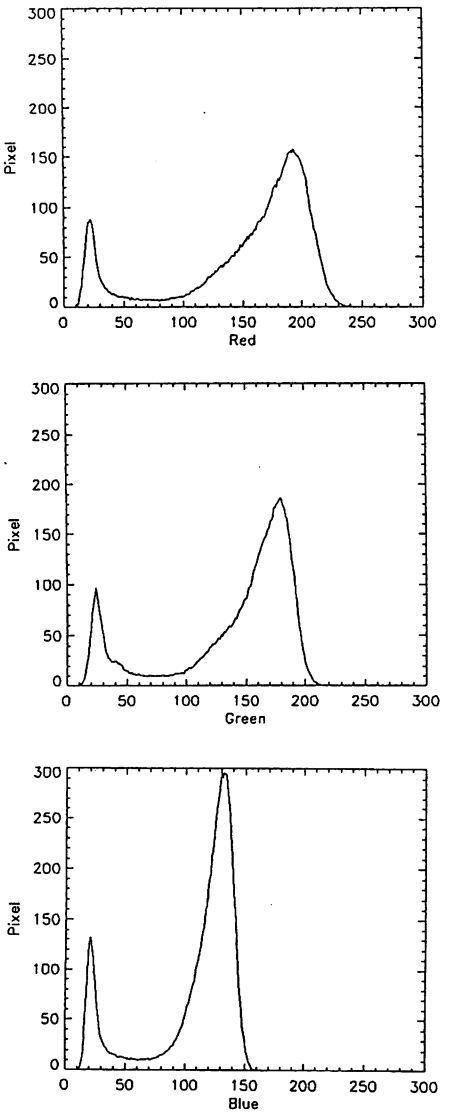


図4．全ピクセルの輝度ヒストグラム

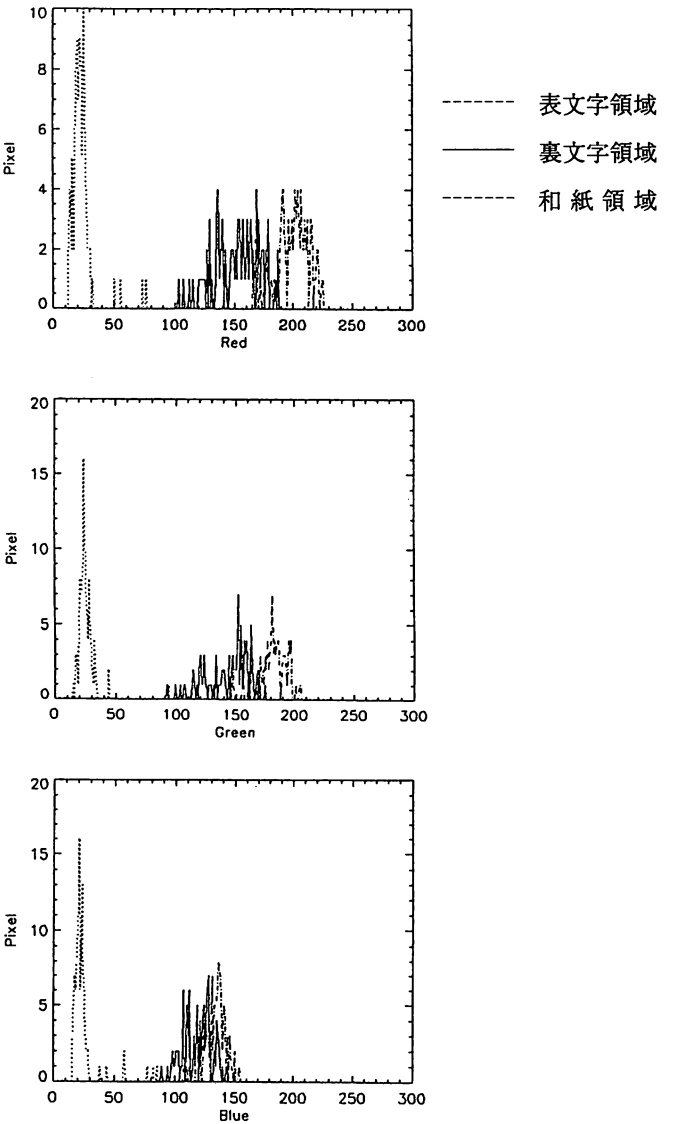


図５．各領域から抽出されたピクセルの輝度ヒストグラム

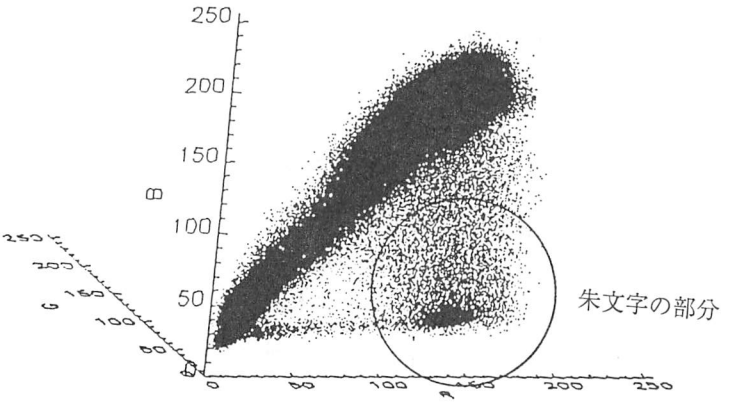
瞭である。さらに図４のように、テキスト・イメージを構成する全ピクセルについて見てみると、裏写りと和紙の２つの領域を区別する閾値を直感的に見つけ出すことはできない。以上の予備的考察から、表文字とそれ以外の領域を分割する閾値を決定することは比較的容易であるように見える。このアプローチについては第Ⅳ章において「判別閾値選択法による領域分割」としてまとめた。

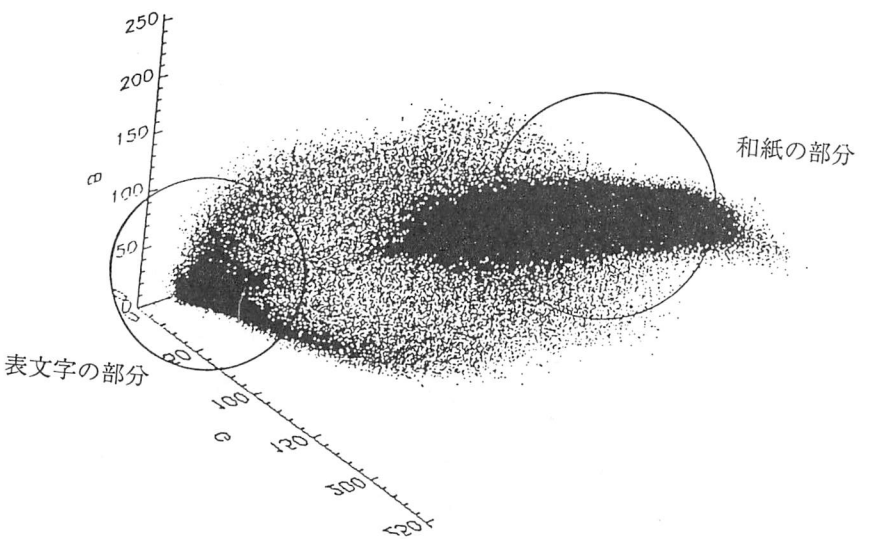
次に、図６は全イメージのピクセルの輝度をRGB座標系に展開したものである。ここで座標軸はＲ、Ｇ、Ｂ各基本色成分の輝度を表しており、０（暗い）から255（明るい）までの256のレベルに量子化されている。

図の上段は各ピクセルの輝度分布を原点およびＢ軸方向から眺めたものである。一方下段は各ピクセルの輝度分布をＧ軸方向から眺めたものである。これより、

１）多くのピクセルは、黒＝（０、０、０）から白＝(255、255、255）を結ぶ直線、つまりＲ＝Ｇ＝Ｂの周辺に分布している、
２）ピクセルの分布が輝度の比較的低い領域（黒っぽいピクセル）と高い領域（白っぽいピクセル）の２つの領域に分かれている、
３）白っぽいピクセルの方が黒っぽいピクセルより多い、

ことが分かる。これはページ・イメージの大部分が白色系の和紙、黒色系の表文字および薄黒色系の裏写りから構成され、かつ和紙の部分が最も広い面積を占めていることを考えれば、当然の結果である。つまり、黒っぽい領域のピクセル群は表文字を、白っぽい領域のピクセル群は和紙の領域を、されに２つの領域の間に分布するピクセル群は裏写りを表しているものと考えられる。これを検証するために、図５で利用した各領域からのサンプルをRGB座標系に展開したものが図７である。図６と図７を比較することにより、表文字領域から選択されたピクセルは原点の周辺に分布し、その他の領域から選択されたピクセルとは明瞭な分布の違いを示すことが確認された。さらに、Ｒ、Ｇ、Ｂごと





6．全ピクセルのRGB輝度座標への展開

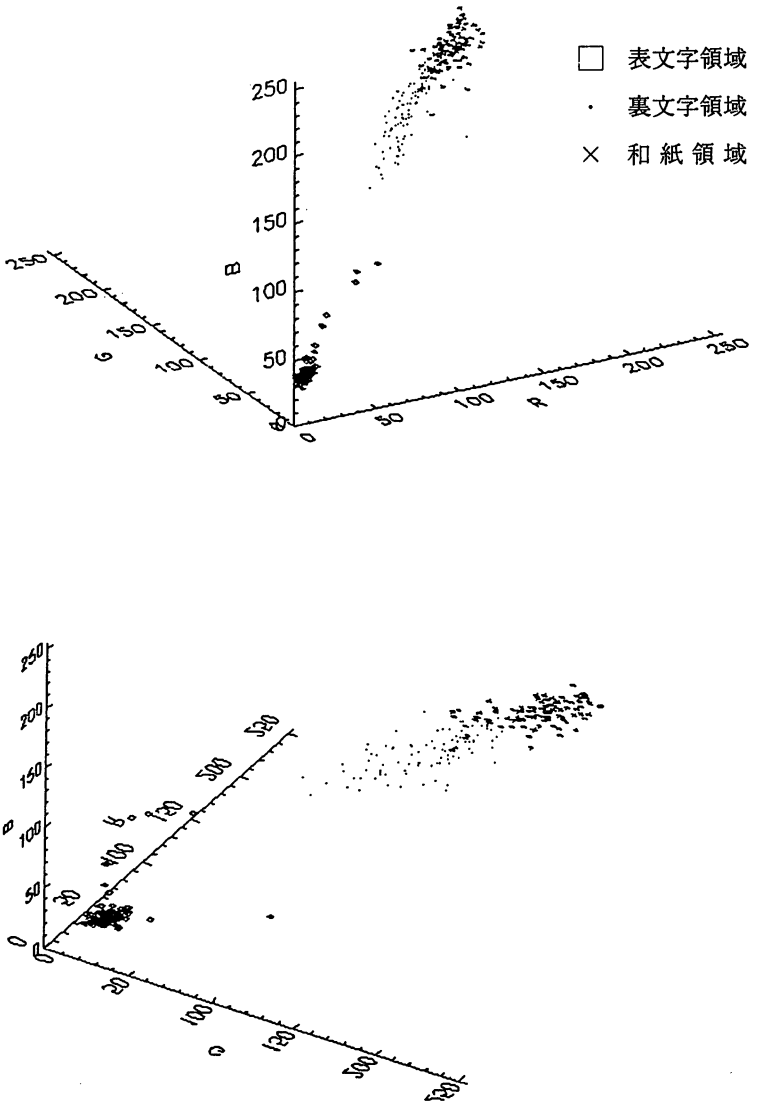


図７．各領域から抽出されたピクセルのRGB輝度座標への展開図

の白黒濃淡ヒストグラムからでは不明瞭であった裏写りと和紙の領域についても、3次元空間上で眺めてみると両者を分割できる明瞭な平面が存在していることがわかる。ところで、図6の上段は全ピクセルの分布を原点付近から眺めたものであるが、R軸に沿った部分にピクセルの集中した領域が見られ、これはほぼ朱文字の部分に相当している。このように、ピクセルの輝度分布を単純な白黒濃淡イメージではなく多次元的なカラー濃淡イメージとして扱うと、より精密な領域分割が可能であることが示唆される。このアプローチは第Ⅴ章において「クラスタ分析法による領域分割」としてまとめた。

Ⅳ. 判別閾値選択法による領域分割

(1) 判別閾値選択法の概要

白黒濃淡イメージを対象領域と背景領域に分割する操作は、イメージ処理の基本的技法の1つである。なお、以下で利用する記号を、

(x、y):ピクセルの座標

d:各ピクセルの輝度

f:座標と輝度の組、つまりf=(x,y,d)

g:分類用の関数

k:閾値

C:分類されるクラス

のように定義する。イメージ処理における領域分割とは、分類すべき対象集合の要素fに対して、関数g(f)と閾値k_iが設定され、

(1) $k_i < g(f) \leqq k_{i+1} \rightarrow f \in C_i$ ただし $i=1,2,\cdots$

のように要素fをあるクラスC_iに決定する手続きである。ここで問題となるのは閾値の設定法である。これについては多くの手法が提案されているが、事

前知識を利用する方法と、事前知識を利用しない方法の2種類に大別できる。前者の典型的手法として、判別分析法を挙げることができる。判別分析法は、p個の変量の組 $f=(x_1、・・・、x_P)$ についてのサンプルがn個のクラスから観測されていたとき、ある未分類の観測値fがこれらn個のクラスのどれに属するかを、すでに分類の明かなサンプルに基づいて判断しようとするものである。判別分析法では、観測されているデータから判別関数と呼ばれる関数 $g(f)$ を作り、この値の大小によって判別を行う。関数gとしては x_i の線形結合を用いることが多く、これを特に線形判別分析法と呼ぶこともある。つまり判別分析法では、判別の根拠となるサンプル、つまり事前知識が必要である。したがって観測対象となる母集団から事前知識を構成するためのサンプル・データの抽出が可能な問題に対しては、判別分析法が有効である。しかし、テキスト・イメージごとに表文字領域（表文字クラス）とイメージノイズ領域（イメージ・ノイズクラス）に属するピクセルを事前知識としてシステムに入力することは、煩雑であり実用的ではない。このことから、今回の研究対象には判別分析法のように、事前知識を必要とする手法が馴染まないことは明白である。

一方、事前知識を利用できない状況下で判別を行う方法として、fのヒストグラムから閾値を求めることが考えられる。図5はその典型例であり、文字領域からのサンプルと、それ以外の領域からのサンプルの輝度分布を示す各ピークの間に、明瞭な谷間が存在している。このような場合は、谷間に対応する輝度を閾値とすればよい（モード法）。しかし、一般的には図5のBヒストグラムにおける裏写り領域と和紙領域からのサンプルの輝度分布のように、明瞭な谷が見られなかったり、あるいは図4のように単峰となってしまう場合もある。判別閾値選択法は、事前知識が無くかつ明瞭な谷間を持たないようなヒストグラムに対しても、判別分析法の立場から自動的に閾値を選択する方法である。以下では文献1に基づいて2値化における判別閾値法について概説する。

いま、テキスト・イメージのデータは、輝度レベル数Lで量子化されている

ものとする。輝度レベルiのピクセル数をn_i、イメージ全体のピクセル数を$N=n_1+\cdots+n_L$とすると、正規化されたヒストグラム $p_i=n_i/N$は輝度の確率分布とみなせるので、テキスト・イメージの1次および2次モーメントは、

(2) $\mu=\sum_{i=1}^{L} iP_i$

(3) $\sigma^2=\sum_{i=1}^{L}(i-\mu)^2P_i$

で与えられる。ここで輝度kを閾値として、輝度レベルを$C_1=\{1、2、\cdots、k\}$と$C_2=\{k+1、k+2、\cdots、L\}$の2つのクラスに分類したものとする（C_1は対象領域、C_2を背景領域、あるいはその逆であるとする）。このときC_1クラスに関する0次および1次モーメントを

(4) $\omega(k)=\sum_{i=1}^{k} P_i$

(5) $\mu(k)=\sum_{i=1}^{L} ip_i$

と定義する。これより、各クラスの生起確率は、

(6) $\omega_1=P_r(C_1)=\sum_{i\in C_1} p_i=\omega(k)$

(7) $\omega_2=P_r(C_2)=\sum_{i\in C_2} p_i=1-\omega(k)$

であり、同様に各クラスの平均は

(8) $\mu_1=\sum_{i\in C_1} iP_r(i\mid C_1)=\sum_{i\in C_1} ip_i/\omega_1=\mu(k)/\omega(k)$

(9) $\mu_2=\sum_{i\in C_2} iP_r(i\mid C_2)=\sum_{i\in C_2} ip_i/\omega_2$

$=(\mu-\mu(k))/(1-\omega(k))$

となる。明らかに、

(10) $\omega_1+\omega_2=1$、　$\omega_1\mu_1+\omega_2\mu_2=\mu$

が成り立つ。さらに、各クラスの分散は、

(11) $\sigma_1^2=\sum_{i\in C_1}(i-\mu_1)^2P_r(i\mid C_1)=\sum_{i\in C_1}(i-\mu_1)^2p_i/\omega_1$

(12) $\sigma_2^2=\sum_{i\in C_2}(i-\mu_2)^2P_r(i\mid C_2)=\sum_{i\in C_2}(i-\mu_2)^2p_2/\omega_2$

となる。ここで、閾値の性能を評価するために判別関数法で用いられている基準を導入する。これはクラス内分散の和（σ_B^2）が最小で、かつクラス間分散の和（σ_W^2）が最大となることである。この基準は、

(13) $\lambda=\sigma_B^2/\sigma_W^2$

ただし、

(14) $\sigma_B^2=\omega_1\sigma_1^2+\omega_2\sigma_2^2$

(15) $\sigma_W^2=\omega_1(\mu_1-\mu)^2+\omega_2(\mu_2-\mu)^2$

$=\omega_1\omega_2(\mu_1-\mu_2)^2$

のように表される。またこれらの間には

(16) $\sigma_W^2+\sigma_B^2=\sigma_T^2$

が常に成り立つ。これより、最適のkはλを最小とすればよいことがわかる。ところでλ以外の基準として

(17) $\kappa=\sigma_T^2/\sigma_W^2$

(18) $\eta=\sigma_B^2/\sigma_T^2$

があるが、これらは

(19) $\kappa=\lambda+1$

(20) $\eta=\lambda/(\lambda+1)$

となりλと同等の基準となる。これら3つの基準において、全分散σ_T^2はkによらず一定である。一方、σ_W^2とσ_B^2については、σ_W^2が2次の統計量であるのに対して、σ_B^2は1次の統計量ある。したがって閾値の性能評価ではηによる計算コストが最も小さくなる。これより、ηを最大にする、つまりσ_B^2を最大にするkの値を閾値とするれば良いことが分かる。

（２）古典資料への判別閾値選択法の適用

ここでは、判別閾値選択法によるテキスト・イメージの２値化をＲ、Ｇ、Ｂの基本色成分ごとに実行した。その結果を下表に示す。ここでクラス０は輝度の低い領域、クラス１は輝度の高い領域を示す。この研究において抽出すべき領域は、黒色系の墨で書かれた表文字であるから、クラス０が対象となる。そこで、基本色濃淡イメージごとにクラス０に分類された領域を、２値化イメージとして表示したものを図８～図10に示す。

図８の特徴は朱文字の部分が除去されていることである。これは図８が赤色濃淡イメージであり、このイメージにおいて朱色のピクセルは、輝度の高い領域に分布するためである（図６参照）。同様のことは緑色濃度イメージにおける緑系統の領域、青色濃度イメージにおける青系統の領域についても言えるが、今回この資料にはそのような領域がないので、図８のような顕著な領域除去は図９、図10には見られない。

さて、黒色系の墨で書かれた表文字は、RGB輝度座標系の原点に近い領域、つまりＲ、Ｇ、Ｂいづれの基本色成分においてもクラス０の領域であるから、

	赤	緑	青
閾　　値	89	83	63
クラス０の平均	4.4499	5.2603	3.9982
クラス１の平均	174.9671	160.8677	122.5050
クラス０の生起確率	0.5789	0.5878	0.5865
クラス１の生起確率	0.4211	0.4122	0.4135

(20)　｛ｆ＝(Ｒ、Ｇ、Ｂ)｜０≦Ｒ≦89｜０≦Ｇ≦83｜０≦Ｂ≦63｝

を満たすピクセルｆが該当する。その部分を２値化イメージとして表示したものが図11である。図11では朱文字や裏写りは殆ど除去され、表文字のみが抽出されていることが分かる。これより、判別閾値選択法のカラー濃淡イメージへの適用の有効性が例証された。

しかし、図11と図１あるいは図３を詳しく比べてみると、図11の文字はやや細く、カスレ気味であることが分かる。言い換えると、本来は表文字の領域に属しているはずのピクセルが除去されていることになる。この理由の１つとして、領域を分割する平面の精度の悪さを指摘できる。図６、図７を見ると各領域を構成するピクセルを分割できる平面が確かに存在している。ところで、一般にＲ、Ｇ、Ｂの互いに直交する座標軸によって張られた３次元空間上の平面は、

(21) $z = r \cdot R + g \cdot G + b \cdot B$

で表現されるので、RGB輝度空間上に分布するピクセルを分割する平面も(21) 式で記述できる。一方、図11の２値化イメージを作成するためには (20) 式を利用したが、これは各軸を含む矩型領域を表している。つまり、(21) 式の平面によって分割されるべき空間が、(20) 式の矩型領域によって近似的に分割されたことになる。これは、(20) 式を構成する過程で、本来３次元で表現されるべき情報を、各軸上に射影したことにより、情報量を減少させてしまったためである。

これより、領域分割の精度を上げる方法の１つとして、各ピクセルの情報を元の３次元空間上で処理することが考えられる。判別閾値選択法の延長線上でこの問題を処理する方法としては、以下の３つが考えられる。１つは判別分析法の利用であるが、この方法では前述のように教師信号を与える必要があるので、本来の目的にはそぐわない。２つ目はパラメトリックな方法で、(13) 式を評価関数として、(21) 式の各パラメータを繰り返し演算によって求める方法である。しかし、この方法では大量の計算を必要とするうえ、しばしば解が

不安定になることがある。また、本法を多値化へ拡張するとさらに計算量が増加する。3番目の方法はクラスタ分析の適用である。クラスタ分析は教師信号を必要としないことに加えて、多値化の際のクラス数の推定にも便利である。このような理由から、イメージノイズ除去の精度をあげる方法として、クラスタ分析の適用を試みた。

Ⅵ. クラスタ分析法による領域分割

（1）クラスタ分析の概要

クラスタ分析は、観測対象となる個体についての何らかの計測量に基づいて、「近いもの」同士を塊（クラスタ）として集めて行く手法である。イメージ処理におけるクラスタ分析の一般的な利用法は、イメージ・データからランダムに抽出されたピクセルを個体とし、それらをある特徴空間上に展開し、その特徴空間における「近さ」の度合いに応じて、いくつかのクラスタに分類するものである。クラスタリングの手法としては、データを樹系図状に構造化する階層的手法、クラス間のデータ移動によってクラス構成の最適化をはかる再配置的手法、分布関数に相当するモードを疑似的に把握する階層的モード法などがあるが、ここでは最も基本的な階層的手法を適用した[3]。

階層的クラスタリングは、個体間の「近さ」の度合いを距離で評価し、距離の近い個体同士を同じクラスタであると判断して融合させて行く方法である。距離としてはユークリッド距離、重み付きユークリッド距離、マハラノビス距離などがあるが、ここでは最も直感的であるユークリッド距離を利用した。

さてクラスタリングは次の手順で行う。いまN個の個体があったとする。N個の個体同士の全ての対に対して N・(N－1)／2組の距離を計算し、この中から距離が最小となる2つの個体を融合してクラスタを生成する。新しく生

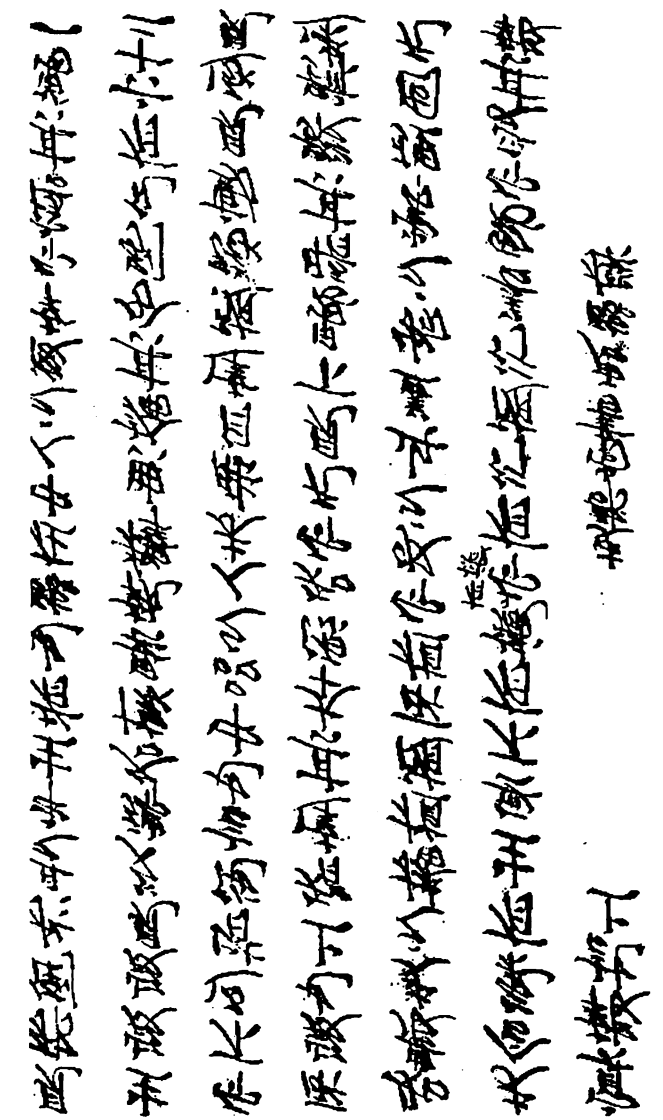


図８．判別閾値選択法による表文字領域の抽出例（赤色濃淡イメージ）

図９．判別閾値選択法による表文字領域の抽出例（緑色濃淡イメージ）

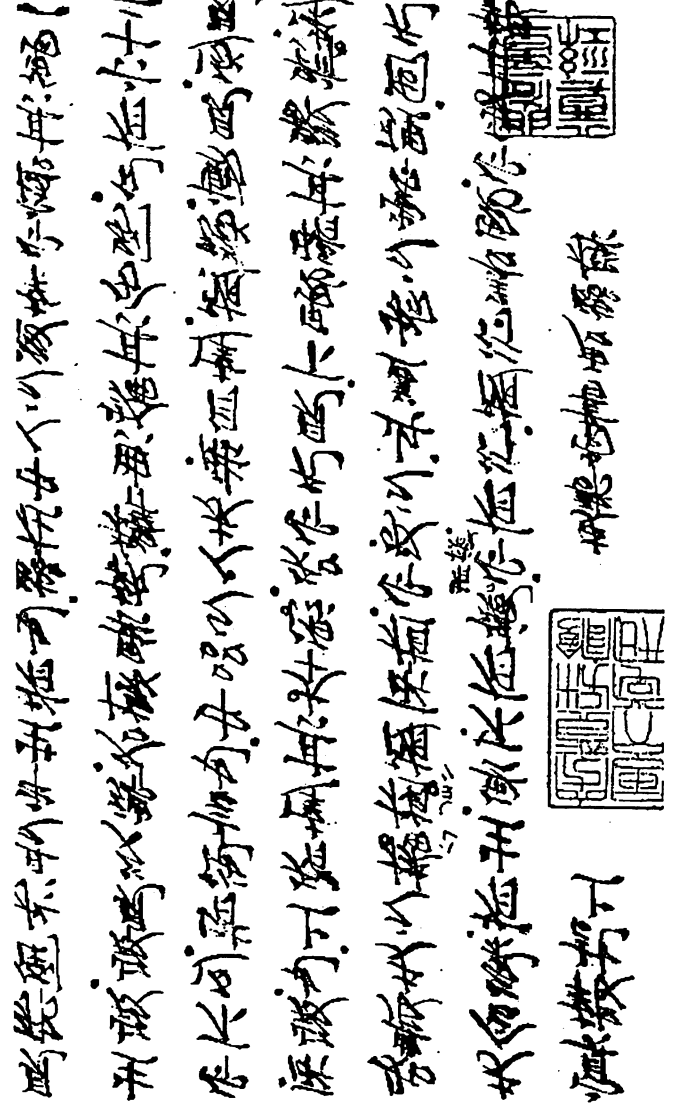


図10. 判別閾値選択法による表文字領域の抽出例（青色濃淡イメージ）

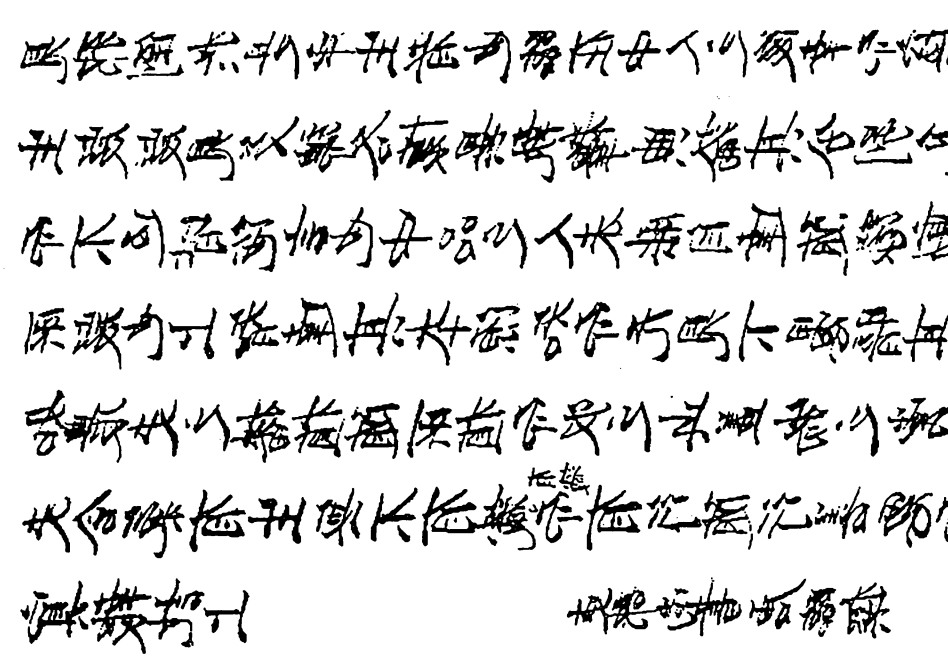


図11. 判別閾値選択法による表文字領域の抽出例（元イメージ）

成されたクラスタは、これ以降の操作において１つの個体と見なされる。次に（Ｎ－１）個の個体について同様の操作を行う。この操作を続けて行くと最終的には１つのクラスタにまとめあげられる。

ここで問題となるのは、新たに生成されたクラスタと、それ以外の個体あるいはクラスタとの間の距離の設定方法である。これには最短距離法、最長距離法、メディアン法、重心法、群平均法、Ward法などがあるが、最長距離法と最短距離法が基本的である。このうち最短距離法はクラスタの融合によって空間が濃縮される（融合したクラスタの周辺の個体が互いに近づく）ので、個体と個体の間に鎖上のクラスタができ、分類感度が低下する。反対に最長距離法は空間が拡散するので分類感度は高くなる。このような理由から最長距離法を採用した。

（２）クラスタ分析による領域分割法の概要

一般に、クラスタ分析は多くの計算時間と記憶容量を必要とするため、イメージ処理においては全イメージ・データを分析の対象とする事は少ない。多くの場合、サンプル用のデータを抽出し、このデータに対して計算を行い、その結果に基づいてイメージ全体の領域分割を行っている。本研究で利用しているイメージ・データも、１ページあたり１、304、976バイトを要するので、全データをクラスタ分析の対象とする事は不可能である。

本研究の特色は、この問題点を解決する方法として、以下に示すような一連の処理を行ったことにある。

１）テキスト・イメージからサンプル用ピクセルを抽出する。
２）サンプル・ピクセルに対してクラスタ分析を行い、サンプルを表文字、裏写り、朱文字、和紙のクラスタに分類する。
３）分類されたクラスタに属するピクセルデータを教師信号として、対象領域とそれ以外の領域を弁別するための線形判別関数を生成する。

4）テキスト・イメージの全データに判別関数を適用して、対象となる領域を抽出する。

つまり、本処理法の特徴はクラスタ分析によって教師信号を生成し、これによって判別関数を構成した点にある。以下に処理の詳細を述べる。

特徴空間としてはテキスト・イメージから抽出されたピクセルのR、G、Bの各基本色成分の輝度をRGB直交座標系に展開したものを利用し、これを

(22) $f=(R、G、B)$

で表した。また、各個体間の距離としてユークリッド距離、つまり個体iとjの間の距離d_{ij}は、

(23) $d_{ij}=((R_i-R_j)^2+(G_i-G_j)^2+(B_i-B_j)^2)^{1/2}$

で定義される。また、クラスタ間の距離は最長距離法、つまりクラスタiとクラスタjを融合して新しいクラスタxを生成したとき、クラスタxと別のクラスタyとの距離d_{xy}は、

(24) $d_{xy}=max(d_{iy}、d_{jy})$

で定義される。このように定義されたクラスタ分析法をテキスト・イメージから抽出されたサンプルに適用し、これらを表文字、裏写り、朱文字、和紙のクラスタに分類する。

クラスタ分析によって分類されたピクセルを、抽出の対象となるクラスとそれ以外のクラスの集まりの2つのクラスに弁別するため、2つのクラスに属するサンプルの輝度ベクトル$f=(R、G、B)$を教師信号として、第Ⅴ章（1）節で述べた判別関数を作成した。具体的には（21）式に示される線形判別関数となる。なお、以降では表文字領域として分類されたピクセルを第1クラス、それ以外の領域として分類されたピクセルの集まりを新たに第2クラスとする。

線形判別分析法では、（21）式に各クラスのサンプルを適用して判別得点zの値、

(25) $z_i^{(k)} = r \cdot R_i^{(k)} + g \cdot G_i^{(k)} + b \cdot B_i^{(k)}$

$(k=1,2;i=1,2,\cdots,n_k)$

を計算する。ただし、kはクラスを、iはピクセルを表す。この$z_i^{(k)}$の変動は、全分散をσ_T^2、クラス内分散をσ_B^2、クラス間分散をσ_W^2とすると、

(26) $\sigma_T^2 = \sum_{k=1}^{2} \sum_{i=1}^{n_k} (z_i^{(k)} - \overline{z})^2$

(28) $\sigma_B^2 = \sum_{k=1}^{2} \sum_{i=1}^{n_k} (z_i^{(k)} - \overline{z}^{(k)})^2$

(29) $\sigma_W^2 = \sum_{k=1}^{2} n_k (z^{(k)} - \overline{z})^2$

となる。これらの間には、

(30) $\sigma_B^2 + \sigma_W^2 = \sigma_T^2$

が成立する。ただし、$\overline{z}^{(k)}$は第kクラスの平均、$\overline{z}$は全体の平均を表す。ここで、クラス間分散 が大きく、同時にクラス内分散が小さければクラス間の弁別が良いことを示す。つまり、

(31) $\lambda = \sigma_W^2 / \sigma_B^2$

を最大にすればよい。これは（13）式の逆数である。したがって、

(32) $\eta = \sigma_W^2 / \sigma_T^2$

を最大にすることと等しい。ηが最大とする係数r、g、bを求めるには、ηをr、g、bについて偏微分してゼロとすることによって求められる。

（3）古典資料へのクラスタ分析法の適用

ここではテキスト・イメージから横50ピクセル×縦30ピクセルの矩型領域を適当に選択し、この領域のピクセルをクラスタ分析用のサンプルとした。この領域のピクセルの基本色成分のRGB座標上の分布を図12に示す。図12と図6を比較すると定性的にほぼ同じ分布であることがわかる。

このサンプルに対して前節のクラスタ分析を行い、ピクセルを７つのクラスタ（クラスタ０～クラスタ６）に分割した。図13～図19に、各クラスタに分類されたピクセルのRGB座標上の分布と、このピクセル群が形成する２値化イメージを示した。これらの分布と図７を比較すると、クラスタ０およびクラスタ１はＲ、Ｇ、Ｂの基本成分色とも輝度が高い領域に分布しており、和紙の領域に相当するクラスタであることが分かる。クラスタ２はサンプル全体の分布の中心付近に位置しており、裏写りの領域に対応したクラスタであることが分かる。クラスタ３はＲ軸に沿って分布しており、朱文字の領域に相当するクラスタであることが分かる。クラスタ４はＲ、Ｇ、Ｂの基本成分色とも輝度が低い領域に分布しており、表文字の領域に相当するクラスタであることが分かる。クラスタ５および６に分類されたピクセルは数が少なく、ノイズと見なすことができよう。

そこで、表文字に相当するクラスタ４を第１クラス、それ以外のクラスタをまとめて第２クラスとして線形判別関数を生成した。その結果を表１に示す。これから得られた判別関数をテキスト・イメージの全データに適用し、表文字部分に相当する領域を抽出して２値化イメージとして表示したものが図20である。図20と判別閾値選択法により作成された２値化イメージ（図11）と比較してみると、図11に見られる表文字部分の切断やカスレが大幅に改善されていることが分かる。これより、クラスタ分析と判別関数を組み合わせた本法の方が、単純な判別閾値法の組み合わせよりも領域分割を適切に行っていることが分かる。

しかし、クラスタ分析では距離計算に時間がかかるため、このままでは実用的でないことも明らかになった。今回のイメージ処理用プログラムは全てワークステーション（SPARC station2）上でＣ言語により作成したが、50×30ピクセルの矩型領域に対するクラスタ分析には約30分を要している。

さらに、ここではクラスタとイメージ領域との対応付けをヒトの手によって

行っているが、自動化を実現するためには、抽出対象領域に対応するクラスタを自動的に判定できなければならない。つまり、画一的な手法やパラメータの適用では、多様な古典原本に対処することは困難であり、何らかの人工知能的アプローチが必要であることを示唆している。例えば今回の例では、「表文字は白地の紙面上にかなりの面積を占めている黒い領域」であるから、

規則：「クラスタの重心座標が最も原点に近く、かつ、そのクラスタに属するピクセルの比率が全体の20％を越えるクラスタに属するピクセルは表文字領域に対応する」

というような規則を適用すれば、表文字に対応したクラスタを判別することは困難ではないと思われる。この方法の長所は、規則を変更するだけで朱文字を抽出することも可能になるなど、同じイメージ処理法をベースにして、融通のある領域分割が可能な点であるが、これらは検討は今後の課題である。

V．まとめ

古典原本へOCRを適用する際に問題となるイメージノイズ除去に関する研究を行った。

イメージノイズ除去としては、領域分割法的なアプローチを採用した。具体的には、抽出対象である表文字と、それ以外の領域を分割することにより、ノイズの除去を行った。これはノイズの対象である裏写りや朱文字が、いわゆる白色ノイズとは異なって構造を持っているため、従来の周波数解析的な方法では除去できないことが明かであったためである。

研究は、古典原本のテキスト・イメージをカラーイメージ・スキャナで取り込んだ後に、各ピクセルデータをRGB表色系へ展開したものを基本的なデー

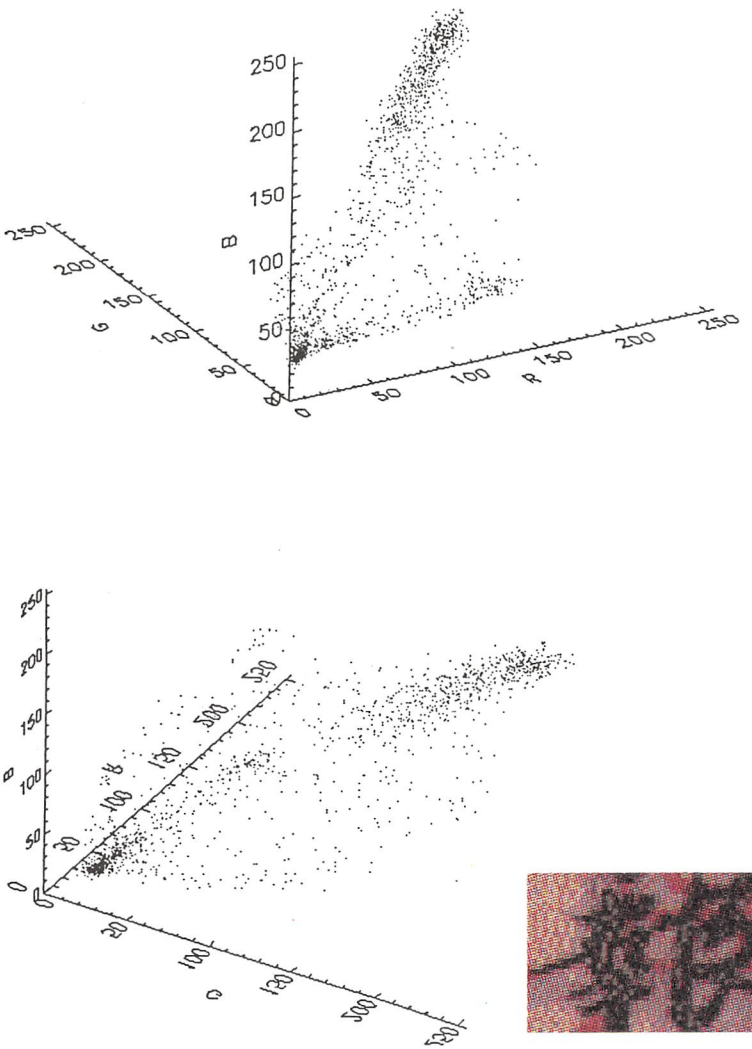


図12．クラスタ分析用サンプル・ピクセルのRGB輝度座標上における分布

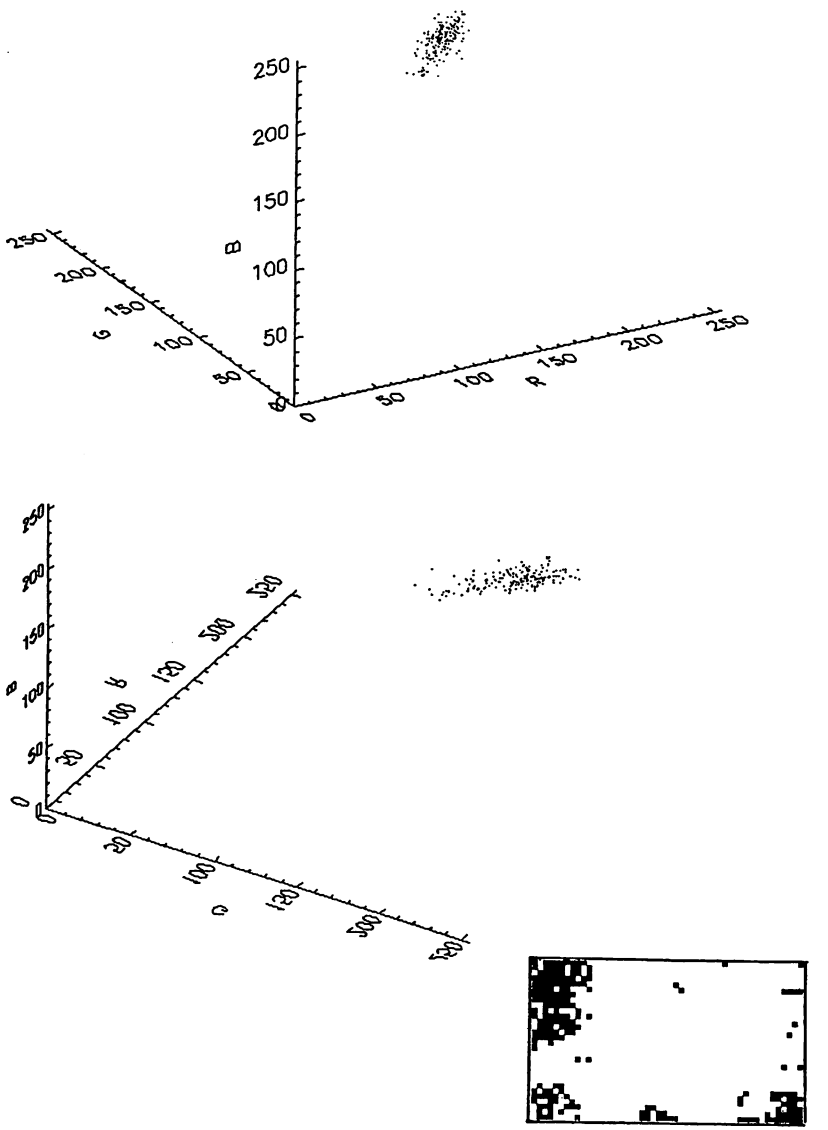


図13. クラスタ分析例（クラスタ０のピクセル分布とその２値化イメージ）

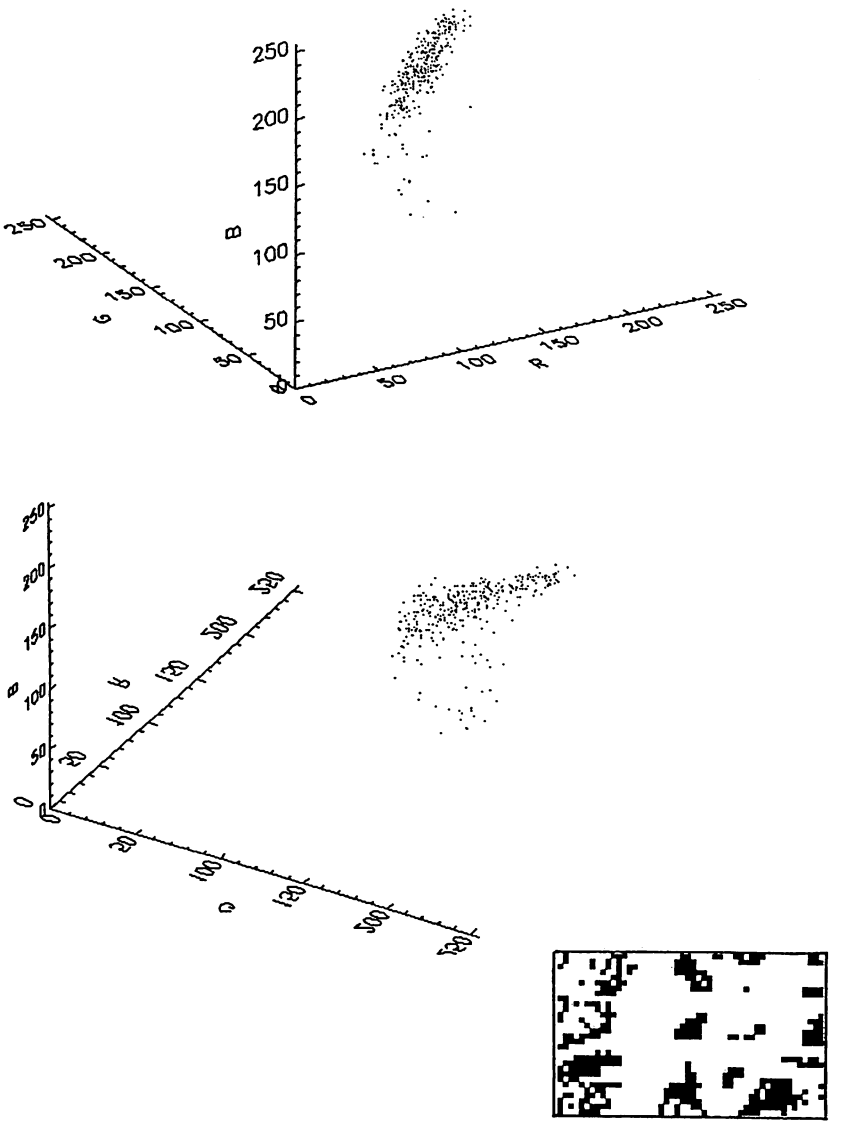


図14. クラスタ分析例（クラスタ１のピクセル分布とその２値化イメージ）

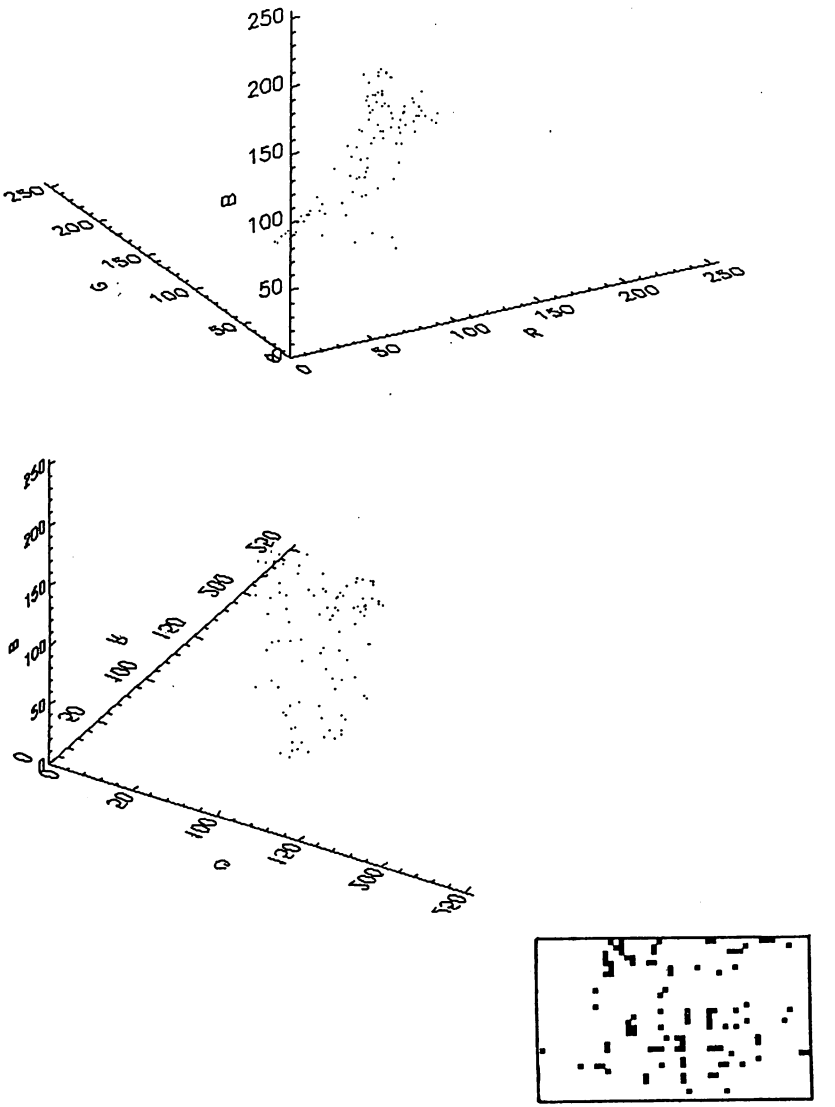


図15. クラスタ分析例（クラスタ2のピクセル分布とその2値化イメージ）

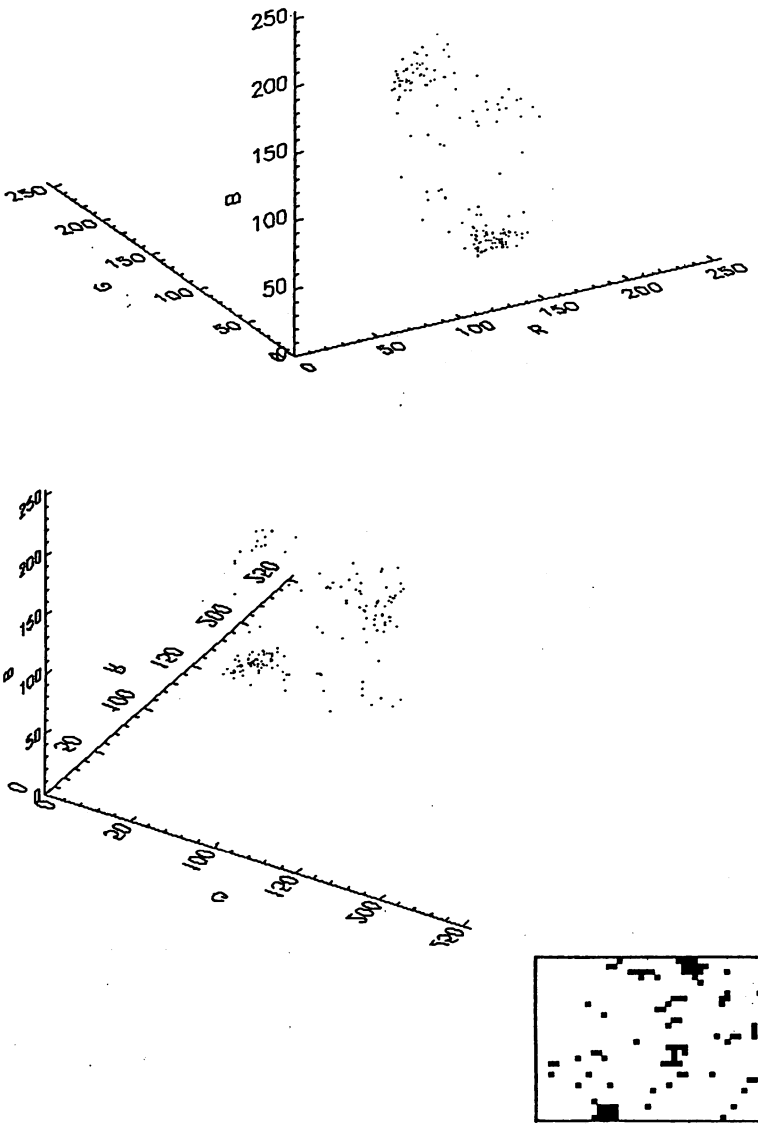


図16. クラスタ分析例（クラスタ3のピクセル分布とその2値化イメージ）

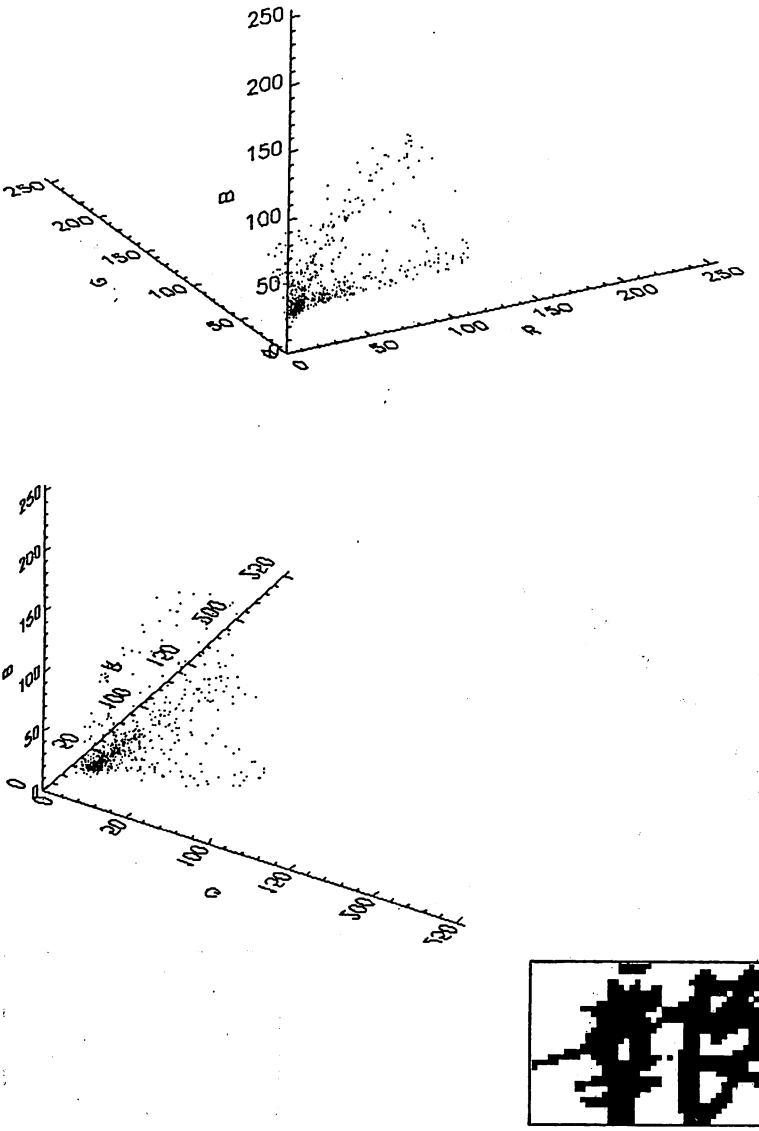


図17. クラスタ分析例（クラスタ4のピクセル分布とその2値化イメージ）

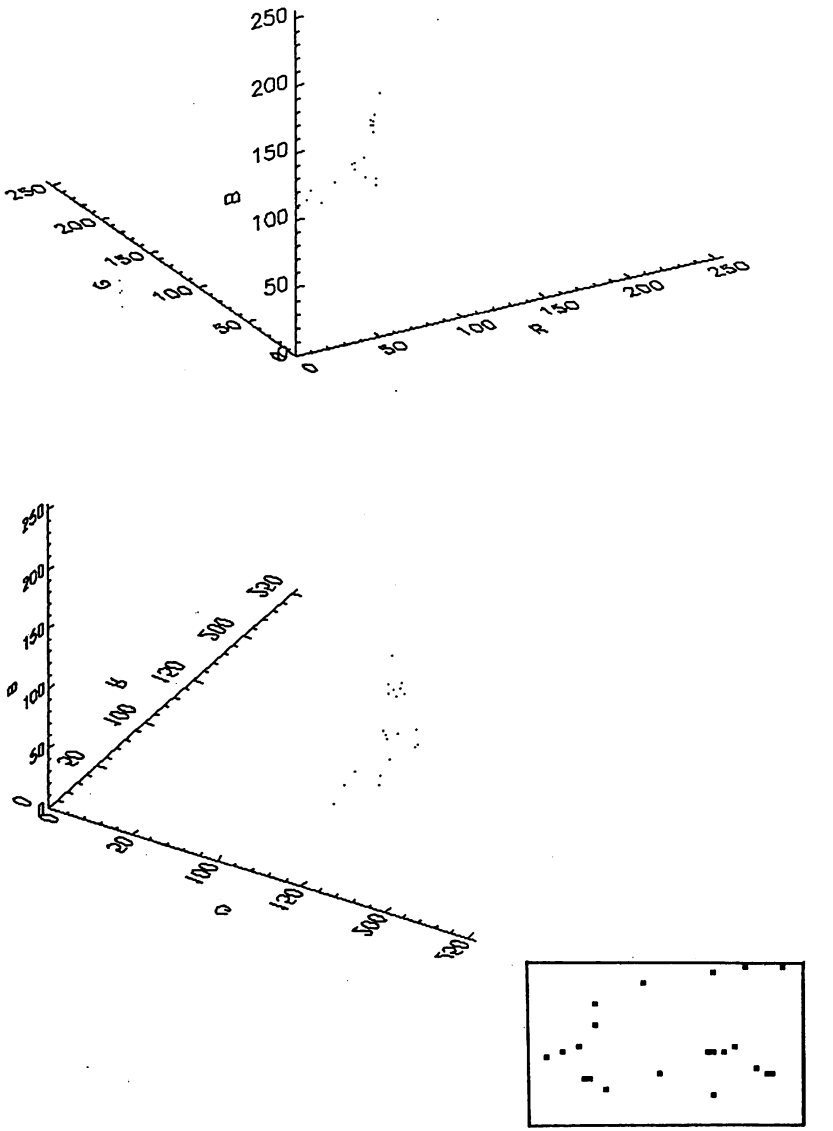


図18. クラスタ分析例（クラスタ５のピクセル分布とその２値化イメージ）

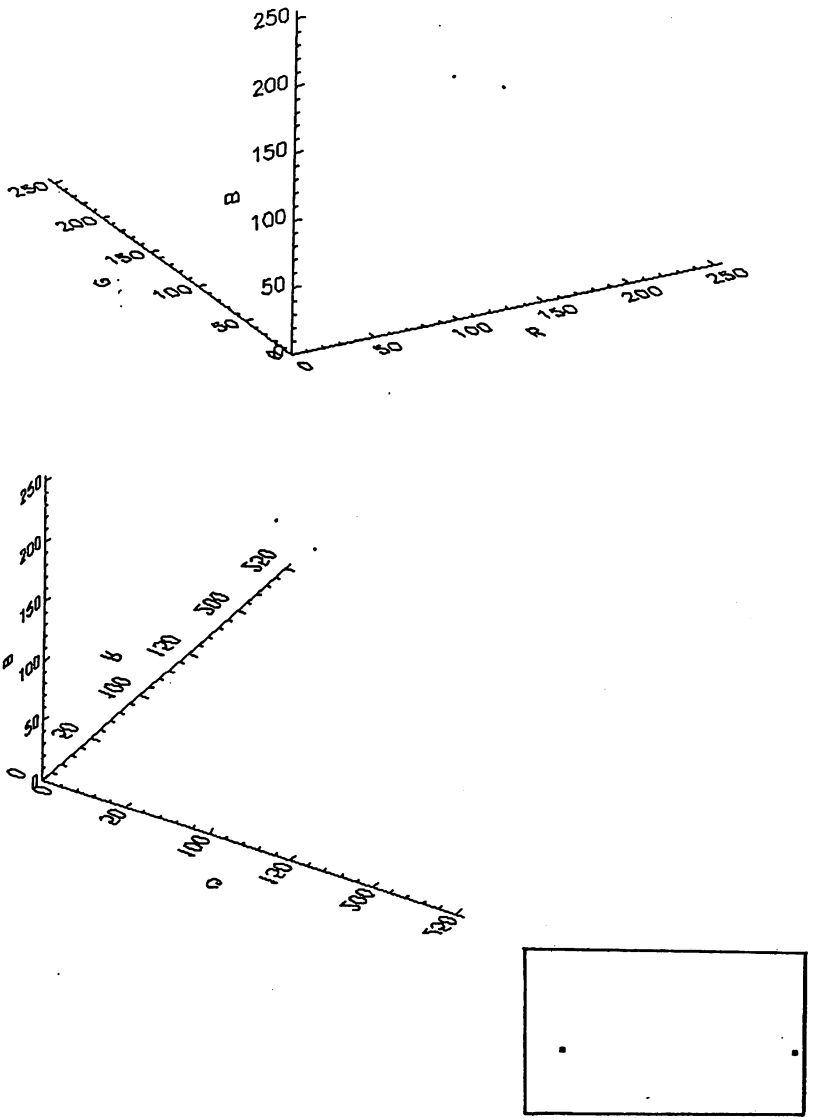


図19. クラスタ分析例（クラスタ6のピクセル分布とその2値化イメージ）

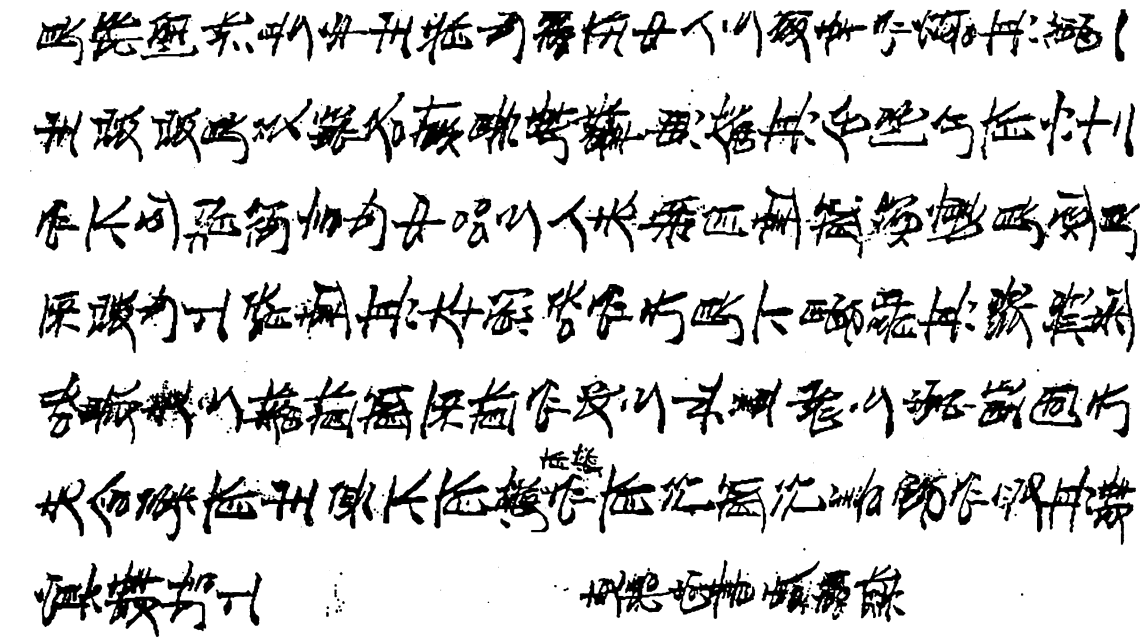


図20．クラスタ分析による表文字領域の抽出例

＊＊＊＊＊＊＊Linear Discriminant Analysys of Two Groups＊＊＊＊＊＊

＊＊＊＊＊＊＊＊＊＊　Basic Statistics of the First Group　＊＊＊＊＊＊＊＊＊＊

Number of Cases＝651

Variable	Average	Variance	S.D
(1)	25.05837	708.11658	26.61046
(2)	20.80952	356.72363	18.88713
(3)	20.89708	538.10480	23.19709

Covariance (Upper) and Corelation (lower) Matrix

	(1)	(2)	(3)
(1)	708.11658	342.41727	422.73831
(2)	0.68130	356.72363	218.20958
(3)	0.68483	0.49805	538.10480

Determination of Covariance Matrix＝38540156.000000

＊＊＊＊＊＊＊＊＊＊　Basic Statistics of the Second Group　＊＊＊＊＊＊＊＊＊＊

Number of Cases＝849

Variable	Average	Variance	S.D
(1)	157.59836	912.50531	30.20770
(2)	135.63840	1887.38904	43.44409
(3)	100.96113	1460.65759	38.21855

表１．クラス４とそれ以外のクラスを分割する線形判別関数の出力例

Covariance (Upper) and Corelation (lower) Matrix

	(1)	(2)	(3)
(1)	912.50531	846.27350	734.74274
(2)	0.64486	1887.38904	1130.20508
(3)	0.63642	0.68069	1460.65759

Determination of Covariance Matrix=690534912.000000

********** Coeficient of LDF **********

Coefficient1	:	-0.16506
Coefficient2	:	-0.03401
Coefficient3	:	0.04135
Constant	:	15.21539

表1（続き）

タとした。

予備的な観察から、

1）多くのピクセルは直線R＝G＝Bの周辺に分布する、

2）朱文字のようなピクセルは1)とは異なった位置に分布する、

3）RGBの各輝度分布は2峰性を示す、

ことが分かった。

これより、和紙と文字の分離は上記3)の性質より、RGBの各輝度分布に対して判別閾値選定法を適用することにより達成できた。また1)及び2)の性質より同手法で「朱文字」部分の分離も可能である。しかし本法では、

1）「表文字」の周辺部が脱落してカスレたようになりやすい、

2）和紙と「裏写り」の分離が不十分、

という問題点があった。そこで、教師情報なし判別法の一法であるクラスタ分析を適用して、分離精度の向上を試みた。本法ではある程度の分離精度の改善を得たが、

1）計算コストが高い、

2）画一的な手法やパラメータの適用では古典原本に対処できない、

などの問題点も明確になった。

以上より、古典原本のイメージノイズ除去にカラー情報の利用が有効であることが確認された。今後の研究課題としては、より低コストな大域的領域分割法の開発と、分割された領域周辺の局地的情報を利用した分割精度の向上が挙げられる。

参考文献

[1] 大津展之：判別および最小2乗法に基づく自動閾値選択法、電子通信学会論文誌、Vol. J63-D、No. 4、pp349～356、1980.

[2] 奥野忠一他：クラスタ分析、続多変量解析法、pp. 207～237、日科技連、1987.

[3] 高木幹雄、下田陽久監修：分類、画像解析ハンドブック、機能編、第Ⅱ部、pp. 6

43～688、東京大学出版会、1992.