# 形状特徴を用いた非文字領域除去処理による文字列領域抽出の高精度化

松尾 賢一　浦西 友樹*　上田 勝彦　梅田 三千雄†

Elimination of Non-character Regions for
Extraction of Character String Region Using Shape Features

Ken'ichi Matsuo, Yuki Uranishi, Katsuhiko UEDA, and Michio UMEDA

In this paper,we propose the method of improving the accuracy in the character string extraction method using the arrangement feature and shape feature of character candidate regions. Gyouten proposed the method of character string extraction using the arrangement feature of the character candidate in the un-format document. However, Gyouten's method is extracted as a mistaken character string pattern, also when an non-character pattern fulfills the arrangement feature. The proposal method removes an un-character pattern as much as possible,before extracting a character string. As a result, the error of the connectivity of a character string decreases by eliminating the un-character pattern using the proposal method. Therefore, it is expected that the correctness of extraction of a character string region improves. The validity of the proposal method became clear by the experiment of the extraction using the document of ten sheets.

## 1　はじめに

文書画像中の文字をOCR（Optical Character Reader：光学式文字読取装置）を用いて認識するには，あらかじめ背景領域から文字領域を分離し，文字単位で切り出されている必要がある．この文字領域と背景領域の分離には，従来から2値化がよく用いられる．一様な背景に文字だけが存在する文書画像では，画像中の各画素における濃度値を，単一の閾値で2値化するによって，文字領域と背景領域を効果的に分離できる[1]．しかし，文字以外の図，表，写真が混在する文書画像に対して，同様の2値化をすると，文字領域以外の背景領域において数多くの擬似領域が発生する[2][3]．この擬似領域は，認識処理によってなんらかの認識結果を出力してしまう問題がある．したがって，図，表，写真が混在する文書画像中の文字を認識するには，擬似領域を限りなく除去し，認識対象となる文字候補だけが抽出されることが望ましい．この図，表，写真が混在する文書画像中から文字領域だけを抽出する手法には，文書に関する書式情報を用いた手法[4]や，書式を用いずに文字や文字列らしい領域を抽出する手法が提案されている．後者の手法では，文字列らしさの特徴を，文字ストロークの形状特徴[6]や文字の並びに関する配置特徴[5]で捉えている．つまり，形状特徴は，文字の局所的な特徴を，配置特徴は，文字の大局的な特徴をとらえているといえる．したがって，この両者の特徴を組み合わせることで，文字や文字列らしさの特徴量を増加させるとともに，文字列領域の抽出精度の向上が見込まれる．

本研究では，この形状特徴と配置特徴の両者の組み合わせとして，擬似領域を含んだ2値画像内の閉領域に対して，形状特徴を用いて非文字領域を検出し，非文字領域だけを取り除いた2値画像から配置特徴を用いて文字列領域候補を抽出する手法を提案し，文字列領域抽出の高精度化を図る．

## 2　文字配置特徴について

行天らは，文字列中の1文字に着目するとき，その文字の存在が及ぼす影響は，両隣の2文字程度であると仮定し，画像中に存在する文字列に対して，以下の4つの文字配置特徴量を定義した[5]．

### 2.1　2文字に関わる配置特徴量

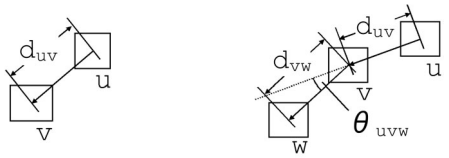
同一の文字列内で隣接する2文字に対して，以下の特徴量を定義する．

＊奈良先端科学技術大学院大学 情報科学研究科
†大阪電気通信大学　総合情報学部

**近接度**　図1（a）のように，矩形$u$と$v$の中心間の距離を$d_{uv}$とし，また，画像中で最も離れている矩形間の距離を$d_{max}$とするとき，近接度$N_{uv}(0\leq N_{uv}\leq 1)$を

$$N_{uv}=\frac{d_{uv}}{d_{max}} \tag{1}$$

と定義する．$N_{uv}$の値が小さいほど，矩形間は近接しているといえる．




（a）2文字に関わる特徴　　（b）3文字に関わる特徴

図1：閉領域の配置特徴

**面積一致度**　図1（a）で，矩形$u$，$v$の面積をそれぞれ$s_u$，$s_v$とするとき，面積一致度$M_{uv}(0\leq M_{uv}\leq 1)$を

$$M_{uv}=\begin{cases}\dfrac{|s_u-s_v|}{s_v} & (|s_u-s_v|\leq s_v)\\ \dfrac{|s_u-s_v|}{s_u} & otherwise\end{cases} \tag{2}$$

と定義する．$M_{uv}$の値が小さいほど，矩形$u$と$v$の面積は類似しているといえる．

### 2.2　3文字に関わる配置特徴量

同一の文字列内で隣接する3文字に対して，以下の特徴量を定義する．

**直線度**　図1（b）で，矩形$u$の中心から$v$の中心へ向かうベクトルを$uv$とするとき，内積$\vec{uv}\cdot\vec{vw}$を用いて，$\theta_{uvw}$は

$$\theta_{uvw}=arccos\frac{\vec{uv}\cdot\vec{vw}}{|\vec{uv}||\vec{vw}|}=arccos\frac{\vec{uv}\cdot\vec{vw}}{d_{uv}d_{vw}} \tag{3}$$

となる．この$\theta_{uvw}$を用いて，直線度$S_{uvw}(0\leq S_{uvw}\leq 1)$を

$$S_{uvw}=\frac{\theta_{uvw}}{2\pi} \tag{4}$$

と定義する．$S_{uvw}$が小さいほど，矩形列$(u,v,w)$は直線的に並んでいるといえる．

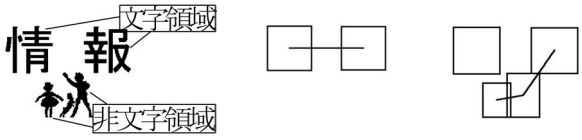
**等間隔度**　図1（b）において，矩形間距離$d_{uv}$，$d_{vw}$を用いて，等間隔度$E_{uvw}(0\leq E_{uvw}\leq 1)$を

$$E_{uvw}=\begin{cases}\dfrac{|d_{uv}-d_{vw}|}{d_{vw}} & (|d_{uv}-d_{vw}|\leq d_{vw})\\ \dfrac{|d_{uv}-d_{vw}|}{d_{uv}} & otherwise\end{cases} \tag{5}$$

と定義する．$E_{uvw}$の値のとり得る範囲は$0\leq E_{uvw}\leq 1$であり，$E_{uvw}$の値が小さいほど，矩形列$(u,v,w)$は等間隔に並んでいることを表す．

### 2.3　配置特徴を用いた抽出

行天らは，画像中の閉領域に外接する矩形の配置を，2.1, 2.2で述べた特徴量をSA（Simulated Annealing）法によって評価し，文字列らしい配列で並ぶ矩形を文字列領域として抽出する手法[5]を提案した．ここで，図2（a）の文字と非文字の領域が混在する画像では，同図(b)のように，「情」と「報」が文字列として連結されることが望ましい．しかし，文字領域の周辺に文字列らしい配列の非文字領域が存在すると，同図（c）のように，非文字領域を誤って連結する問題があった．




（a）入力画像　　（b）抽出結果(正)（c）抽出結果(誤)

図2：行天手法の問題点

## 3　形状特徴について

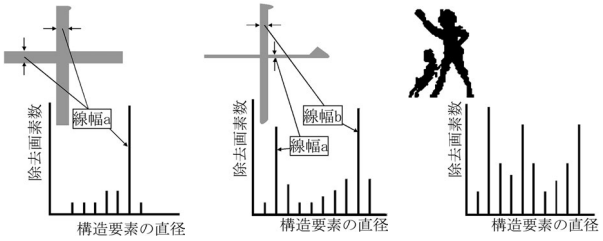
### 3.1　演算とパターンスペクトル

Morphology演算は，2次元空間での対象集合と構造要素との集合演算である．構造要素（本研究では，Circleを用いる）の直径を大きく変化させ，Morphology基本演算であるOpeningを$n$回行いながら，$n-1$回目の演算結果と$n$回目の演算結果との差分をとると，$n$回目の演算で対象図形を$n$倍して除去された画素数が求められる．ここで，横軸に構造要素の直径$d(d=3,5,7,\cdots)$をとり，縦軸に直径$d$の構造要素で除去された対象図形の画素数の頻度値をヒストグラムにする．このときのヒストグラムがパターンスペクトル（Pattern Spectrum）となる[8].

### 3.2　パターンスペクトルからみた文字形状特徴

文字のストロークの線幅は，飾り文字を除いて一定であることが多い．図3（a）の全ての方向ストローク幅が同じゴシック体文字のパターンスペクトルは，単峰性を示す．また，図3（b）の明朝体文字は，縦方向と横方向のストローク線幅は異なるものの，方向別でのストローク線幅は一定であるため，パターンスペクトルは双峰性を示す．これとは別に，図3（c）の文字でない図形では，パターンスペクトルは多峰性を示すことが多い．

したがって，このパターンスペクトルの形状によって，文字と非文字領域をある程度の分類が可能といえる．




(a) ゴシック体　(b) 明朝体　(c) 非文字領域

図3：パターンスペクトルによる文字種の分類

ここで，除去画素数の平均値を超えるパターンスペクトルの頂点を峰と定義し，峰が1つであるときは，単峰性を示すとして，閉領域がゴシック体のストローク構造をもつ「文字領域」と判断する．同様に，峰が2つであるときは，双峰性を示すとして，閉領域が明朝体のストローク構造をもつ「文字領域」と判断する．また，峰が3つ以上である多峰性を示すときは，閉領域は図形である「非文字領域」と判断する．

## 4　提案手法

### 4.1　処理手順と前処理

提案手法は，図4に示すように，閉領域の形状特徴に基づく非文字領域除去処理後に，行天らの提案した手法[5]である閉領域の配置特徴に基づく文字列領域候補抽出処理によって構成される．入力画像は，1画素あたり，R（赤），G（緑），B（青）各8[bits]の輝度値で表現されている．このRGB値で濃淡画像を作成し，手動でしきい値を変化させながら，文字領域と背景領域が良好に分離するように2値化する．次に，得られた2値画像をラベリングしながら，画素数が10画素以下の閉領域を雑音とみなして除去する．

### 4.2　シンボリックチェーンコード

文字は，一般的に細長いストロークで構成されているため，この文字領域の輪郭線による構造特徴において，ストロークの端点で輪郭線の方向性の転進部分が数多く見られる．この輪郭線の往復形状の出現頻度を調べるために，チェーンコード（Chain Code）[9]に，改良を加えたシンボリックチェーンコード（Symbolic Chain Code）[10]によって閉領域の輪郭線の方向性を調べる．さらに，シンボリックチェーンコードから往復形状の出現頻度を表す往復度[10]を求め，閉領域の形状特徴量として用いる．

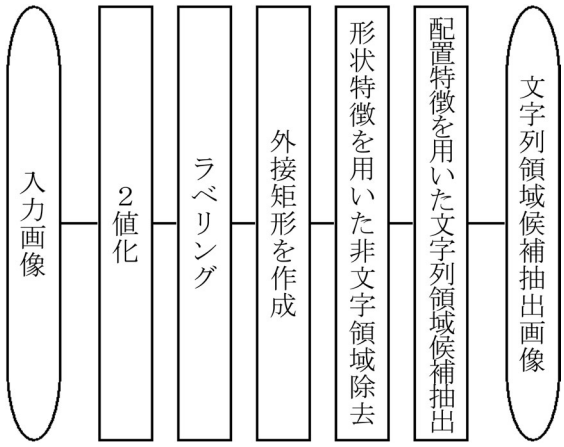



図4：提案手法の流れ

#### 4.2.1　シンボリックチェーンコードの生成

チェーンコードの方向コードを図5に示す．図6（a）の閉領域に対して，“Start”の画素から左回りに，同図（b）の輪郭線追跡によってチェーンコードを生成する．このチェーンコードで，同図(c)にように追跡方向が転進する点を検出し，転進点間のコード列を一つの方向コードで代表させたシンボリックチェーンコードを求める．したがって，同図（d）に示すシンボリックチェーンコードが得られ，コード列は，“284152637486”となる．

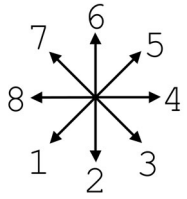


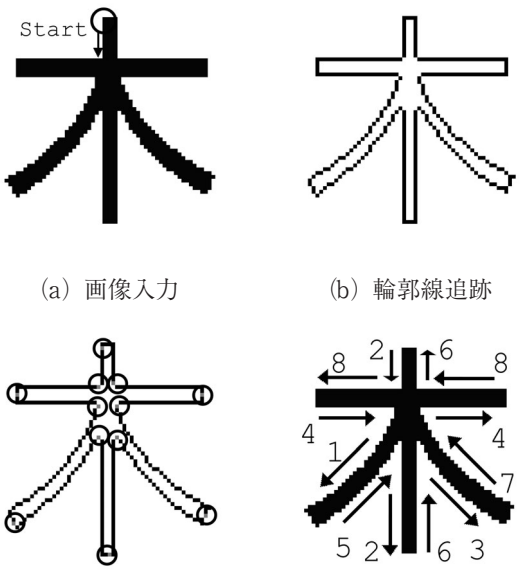

図5：方向コード

#### 4.2.2　往復度

文字領域がもつ輪郭線特徴として，文字ストロークの端点部分で往復する方向コード部分が数多く出現する．この往復コードの頻度を調べるために往復度を定義する．

シンボリックチェーンコード“$s_1 s_2 \cdots s_i \cdots s_m$”において，追跡方向$s_{i-1}$と追跡方向$s_i$がなす角度を$\theta_i$とすると，$\theta_i$のとりうる値は，0，$\frac{\pi}{4}$，$\frac{\pi}{2}$，$\frac{3}{4}\pi$，$\pi$[$rad$]の5つとなる．そこで，角度変化の評価値$A_i$を

$$A_i = \begin{cases} 0\ (\theta_i = 0) \\ 1\ (\theta_i = \frac{\pi}{4}) \\ 2\ (\theta_i = \frac{\pi}{2}) \\ 3\ (\theta_i = \frac{3}{4}\pi) \\ 4\ (\theta_i = \pi) \end{cases} \tag{6}$$




（a）画像入力　（b）輪郭線追跡

（c）転進点検出　（d）コード生成

図6：シンボリックチェーンコードの生成例

と定義したうえで，往復度$R$を

$$R=\sum_{i=1}^{m} A_i \tag{7}$$

で算出する．

### 4.3　形状特徴を用いた非文字領域除去

パターンスペクトルからは，平均値を超える要素が3個以上存在するか，シンボリックチェーンコードからは，往復度が定められた閾値$t_R$（予備実験により$t_R=10$）を満たさないか，を2値画像内の全ての閉領域に対して判定する．どちらかの条件を満たした閉領域を「非文字領域」として2値画像から除去する．

### 4.4　配置特徴を用いた文字列領域候補抽出

行天らは，閉領域の配置特徴量をパラメータとする評価関数$F_c$を定義し，画像中に存在する全ての閉領域の組合せに対して$F_c$を計算し，$F_c$が最小値をとり得る組合せをSimulated Annealing（SA）法で探索した．

しかし，この手法は，画像中に存在する全ての矩形の組合せを総当たりで探索するため，矩形数の増加とともに，計算量が爆発的に増大する問題があった．そこで，SA法に代わる新たな探索手法を提案し，文字列領域候補抽出における計算量の削減をはかる[10].

#### 4.4.1　評価関数の定義

入力画像中に存在する閉領域に外接する矩形の集合を$C$とする．ある矩形$u$を定めたとき，矩形$v,(u,v\in C,u\neq v)$が，$u$と同一の文字列に属する文字らしいかを評価する関数$f_{uv}$を定義する．評価関数$f_{uv}$は，矩形列$(u,v)$が文字列らしい配置であるとき，値が小さくなる．

$f_{uv}$の値が最小になる矩形$v_{min}$を探索し，$v_{min}$が後に述べる探索停止条件を満たさなければ，矩形$u$と$v_{min}$を同一文字列として連結し，$v_{min}$を新たなuとして，処理を繰り返す．また，ここで探索開始矩形を$b,(b\in C)$と定義し，画像中に存在するすべての矩形が$b$となるように処理を繰り返し，1回以上連結された矩形間を「連結された」と判定する．

**$u=b$のとき**　配置特徴量である近接度$N$を用いて，評価関数$f_{uv}$を

$$f_{uv}=1.0N_{uv} \tag{8}$$

と定義する．$u=b$である最初の探索では，3文字に関わる特徴量を計算できないので，2文字に関わる特徴量のみで，矩形列$(u,v)$の文字列らしさを評価する．

**$u\neq b$のとき**　4つの配置特徴量である，近接度$N$，面積一致度$M$，直線度$S$，および等間隔度$E$を用いて，評価関数$f_{uv}$を

$$f_{uv}=0.5N_{uv}+0.5M_{uv}+1.0S_{tuv}+2.0E_{tuv} \tag{9}$$

と定義する．$u\neq b$のとき，少なくとも一回以上移動するため，移動元の矩形$t$と移動前の矩形$u$，移動候補となる矩形$v$の配置特徴を調べることで，矩形列$(t,u,v)$の文字列らしさを評価する．ここで，評価関数$f_{uv}$における各特徴量の重みは，予備実験により決定した．矩形$b$，$u$に対する矩形の探索例を図7に示す．探索開始矩形b，現在の矩形$u$，および，移動候補となる矩形$v_1$，$v_2$，$v_3$によって，それぞれ評価関数$f_{buv_1}$，$f_{buv_2}$，$f_{buv_3}$を計算する．

結果として，関数$f_{buv_1}$，$f_{buv_2}$，$f_{buv_3}$の中で，関数$f_{buv_2}$の値が最も小さくなるため，矩形$b$，$u$と同一の文字列に属する矩形として，矩形$v_2$を連結する．

#### 4.4.2　探索停止条件

評価関数$f_{uv}$の値が最小になる矩形$v_{min}$は，以下の探索停止条件を満たすとき，$u$と$v_{min}$は同一の文字列には属しないと判断し，探索を終了する．

**面積に関わる停止条件**　図8（a）に示す矩形$u$の面積$s_u$に対し，矩形$v$の面積$s_v$が$\frac{1}{2}$倍未満あるいは2倍を超えるとき，矩形$v$は矩形$u$と同一文字列ではないと判定する．

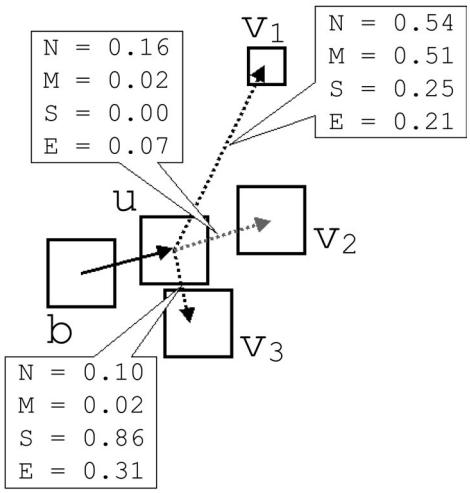
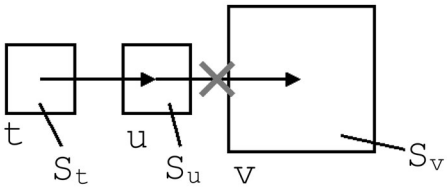
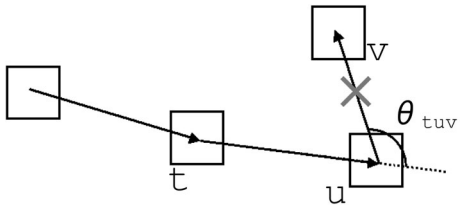


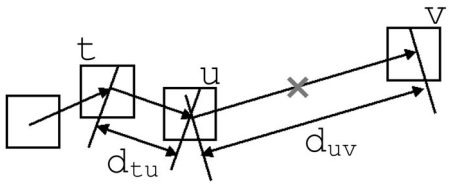

図7：文字列領域抽出処理の一例




(a) 面積に関わる条件




(b) 角度に関わる条件




(c) 距離に関わる条件

図8：探索停止条件

**角度に関わる停止条件**　図8（b）に示す矩形$t$と$u$の中心を結ぶ線分$tu$の延長線と，矩形$u$と$v$の中心を結ぶ線分$uv$のなす角が$\frac{\pi}{2}$以上のとき，矩形$v$は矩形$t$，$u$と同一文字列ではないと判定する．

**距離に関わる停止条件**　図8（c）に示す矩形$t$と$u$の距離$dtu$に対し，矩形$u$と$v$の距離$d_{uv}$が$\frac{1}{2}$倍未満あるいは2倍を超えるとき，矩形$v$は矩形$t$，$u$と同一文字列ではないと判定する．

## 5　文字列領域候補抽出実験

### 5.1　入力画像および前処理

イメージスキャナにより解像度150[dpi]で取り込んだ雑誌表紙内の不規則な配置で存在する文字列を抽出対象とした．画像ごとに，文字が良好に視認できる閾値を手動で定め2値化する．そして，2値画像をラベリングすることで各閉領域の外接する矩形が得られる．

### 5.2 連結誤りの定義

図9のように，本来は同一文字列ではない文字間の連結誤りを「誤連結」，本来は同一文字列である文字間を連結しない誤りを「未連結」と定義する．「誤連結」は，同一文字列に属さない2つの文字領域間を連結する誤りである「C–C（Character to Character）誤り」，非文字領域除去処理で除去できない非文字領域と文字領域の間を連結する誤りである「C–N（Character to Non-character）誤り」，非文字領域間を連結する誤りである「N–N（Non-character to Non-character）誤り」の3種類とする．

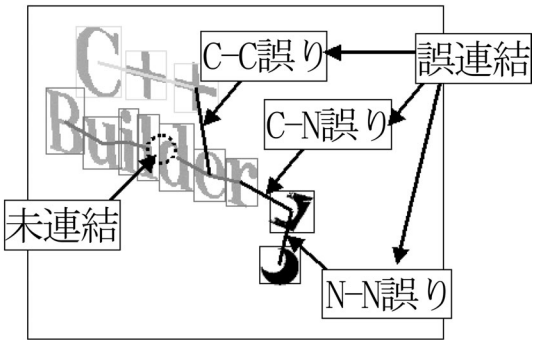



図9：連結誤りの種類

### 5.3　非文字領域による連結誤りの調査

ゴシック体で書かれた3文字から構成される文字列領域と文字領域と同程度のサイズをもち，ランダムに配置された0個から20個の非文字領域が混在する画像21枚を用いて，非文字領域の除去を用いずに，配置特徴だけで抽出処理を行ったときの連結誤りの影響の調査を行なった．結果として，配置特徴による抽出では，図10のように雑音の増加とともに，C–N誤り，C–C誤りは一定であるが，N–N誤り数だけが増加した．

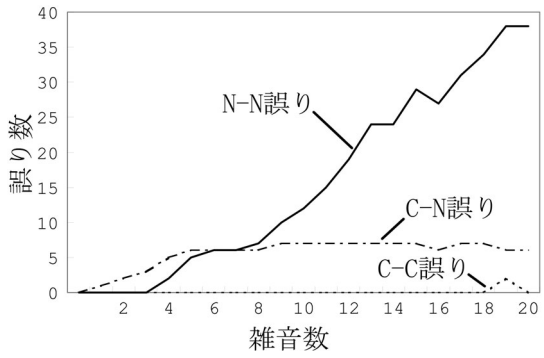



図10：雑音数と誤り数の関係

### 5.4　抽出実験および実験結果

文字領域598文字，非文字領域333文字が存在する入力画像10枚に対して，非文字領域除去処理の有無による連結別の文字列領域候補の抽出結果を表1に示す．

表1：連結別の文字列領域候補抽出結果

| 非文字領域除去 | 正連結 | 未連結 | 誤連結 | | |
|---|---|---|---|---|---|
| | | | C–C | C–N | N–N |
| なし | 290 | 71 | 31 | 9 | 9 |
| あり | 291 | 70 | 43 | 5 | 1 |

表1より，非文字領域除去処理を導入することで，誤連結におけるC–C誤りは増加し，C–N誤りとN–N誤りが減少したことがわかる．

C–C誤りの増加は，現段階で異なる文字列間での文字候補の連結性を高めたにすぎないといえる．しかしながら，全対象領域内の文字領域と非文字領域との比を最大にし，文字認識過程を導入したときの非文字領域の認識結果である無意味な認識結果の発生の防止に寄与したといえる．

また，C–N誤りやN–N誤りは，本来非文字領域である閉領域を文字領域候補として誤抽出する誤りであり，C–N誤りとN–N誤りが減少したことで，文字列領域候補の抽出精度は向上したといえる．

## 6　おわりに

本研究では，図，表，写真と文字が混在する画像から，文字領域候補を抽出するために，閉領域の配置特徴に基づく文字列領域候補を抽出する前に，閉領域の形状特徴を用いて，非文字領域とみられる領域をあらかじめ除去する手法を提案し，文字列領域候補抽出の高精度化を図った．

提案手法は，文字列領域抽出の精度を低下させる非文字領域を，パターンスペクトルの峰の数とシンボリックチェーンコードの往復度を用いて除去する手法を導入することで，抽出精度の高精度化を実現することができた．

さらなる高精度化を実現するためには，線幅が一定であり，なおかつシンボリックチェーンコードに往復形状が多数出現する非文字領域と，本来の文字領域を判別できる新たな特徴量の導入が必要であろう．

## 参考文献

[1] 後藤英昭, 阿曽弘具, “様々な画像に適用できる文字パターン抽出手法について～サーベイおよび一構成例～”, 電子情報通信学会技術報告, PRMU99-234, pp.23–30 (2000)

[2] 美濃導彦, “文書画像処理の現状と動向",電子情報通信学会誌, Vol.76, No.5, pp.502–509 (1993)

[3] J. Ohya, A. Shio, S. Aakamatsu：“Recognizing Character in Scene Images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.16, No2, pp.214–220, 1994

[4]中野康明, “文字認識・文書理解の最新動向[VI・完]–文字切り出しと文書理解–”,電子情報通信学会誌, Vol.83, No.7, pp.576–580 (2000)

[5] 行天啓二, 馬場口登, “制約充足型文字領域抽出の基礎検討”, 電子情報通信学会技術報告, PRU92-119, pp.9–16 (1993)

[6] 顧力栩, 田中直樹, 金子豊久, R.M. Haralick, “表紙画像からの文字領域抽出方式”,電子情報通信学会論文誌D-II, Vol.J80-D-II, No.10, pp.2696–2704 (1997)

[7] 中野康明, “文字認識・文書理解の最新動向[I]–文字認識とは–”,電子情報通信学会誌, Vol.83, No.1, pp.64–68 (2000)

[8] 小畑秀文, “モルフォロジー”, コロナ社, pp.136-148 (1996)

[9] 安居院猛, 長尾智晴, “画像の処理と認識”, 昭晃堂, pp.73–74 (1992)

[10] 浦西友樹, 松尾賢一, 上田勝彦, “形状特徴を用いた非文字領域除去処理による文字列領域抽出の高精度化”, 2004年電子情報通信学会総合大会 講演論文集, D–12–38 (2004)