# 文書画像の領域分割に関する一検討

6R－5

糸井 清晃　新井 浩志　小林 幸雄
千葉工業大学

## 1　まえがき

新聞・雑誌など様々な印刷物があるが、これらに含まれる記事のうち自分の必要な物、或いは興味のある物のみを抜き出して経済的に蓄積し、必要に応じて検索出来るようにするためには、まずこれら記事を画像データとして計算機に入力し、更に文字部分、図表・写真部分等性質の異なる領域に分割し、文字部分については文字認識を用いて符号化するという処理を施すのがよい。そして、この様な処理を施す研究が数多く行われ、様々な手法が提案されている。

本報告では、既存の手法からその一部もしくは、その中で用いられている文書中の各領域に関する特徴等の幾つかを組み合わせて利用した手法によって文書画像を性質の異なる領域に分割する処理を行い、その過程で生じる幾つかの問題点とその解決手法について述べる。

## 2　処理の概要

### 2.1　本文文字列の特徴抽出

入力された文書画像を縦方向・横方向に走査し、それぞれの走査方向にある程度以上の長さを持ち、かつそれと垂直方向にある程度以上の幅を持つ白画素矩形領域[3]を検出し、この領域を除いた領域を連結領域とみなして縦・横方向のランレングス分布を求める（図1参照）。得られた両方向分布中の最大頻度のラン値又は、その付近である程度以上の頻度のラン値を持つ領域を本文文字列領域とし、これらのラン値を本文文字列領域の幅とする。このとき、最大頻度は本文が縦組みの場合は横方向分布に、横組みの場合は縦方向分布に現れるので、これによって本文文字列の構成が縦組みか横組みかを決定する。

### 2.2　黒画素に対する拡大縮退処理

拡大縮退処理[1]によって黒画素を融合する。ここで、2.1の段階で本文文字列が縦組みと判断されている場合には縦方向の拡大縮退幅を大きく取り、横組みの場合は、横方向を大きく取る。この処理では、縮退に関して次の２点が問題となる。

1. 文字列と文字列の間に罫線がある場合、拡大によって文字列領域と罫線が融合して一つの連結領域になり正常に縮退しない（図2）.
2. 画像の端付近に黒画素がある場合、拡大したときに端に達してしまうとその先に白画素が無いので正常に縮退しない（図3）.

以下の節では、これらに対する解決策について述べる。

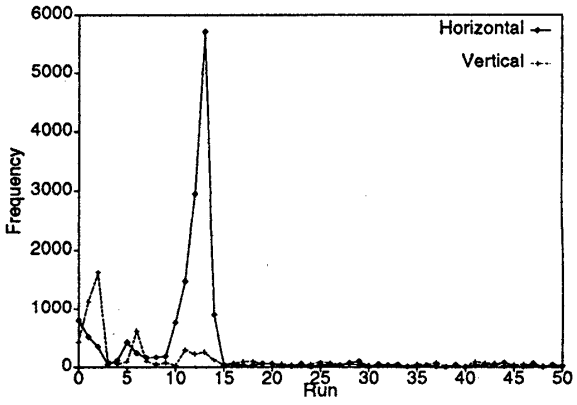


図 1: ランレングス分布（一部）

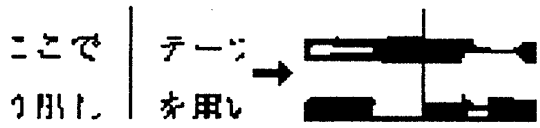

図 2: 罫線と文字列の融合

図 3: 周辺付近の拡大縮退

### 2.3　罫線抽出

まず、問題点1に関しては、拡大縮退処理を行う前に罫線を抽出するという方法を用いる。抽出方法としては、各黒画素連結領域に対し外接矩形と縦・横サイズの大きい方を小さい方で割った縦横比、矩形領域内における長さ方向に垂直な方向の線密度[2]の平均を

---

Printed document segmentation
Kiyoaki ITOI, Hiroshi ARAI, Yukio KOBAYASHI
Chiba Institute of Technology

求め、縦横比が2.1 で求めた文字列幅・行間幅より求めたある閾値より大きくさらに線密度の平均が同様にして求められた閾値より小さい場合にこれを罫線であると判断・削除して別途保存した後、前節の要領で拡大縮退を行う。その際、罫線のあった領域の上下に存在する文字列などの領域が融合して一つの連結領域になってしまい本来縮退すべきところが縮退しない場合があるので、保存しておいた罫線を用いてこれら不正な融合を切る事によって領域を分ける。しかし、こうして分けられた領域はまだ正しく縮退した訳ではないので、これらに対して次節で述べる拡大縮退補正を施す事によって正しい文字列連結領域を得る。

### 2.4 拡大縮退補正

次に、問題点2と前節の処理によって生成された連結領域に関する補正処理について述べる。これら二つの状態の連結領域は、画像の周辺にあるか比較的内側にあるかの違いがあるだけで、いずれも本来の縮退位置、つまり拡大前にその連結領域内で一番外側にある黒画素の位置まで戻っていない（図4：網掛け部分は拡大縮退後の連結領域）。

そこで、同図の点線で示される外接矩形を考えて、この4辺のうち原画像上の黒画素に接していない辺（図では右の辺）があれば、それを矩形領域の面積が減少する方向へ移動しながら連結領域を削って行くという処理を辺が黒画素に接触するまで処理を施す（図5）。

### 2.5 領域の判別

以上の過程によって、文書中の連結領域が生成されたので、各連結領域に対する特徴量として線密度の平均と分散・ランレングス分布を水平垂直方向それぞれ求め、以下の要領で性質の異なる領域に分割する。ただし、領域の判別に用いる値の範囲は、 2.1 で求められている文字列幅・行間幅より求められる物と各領域のみの画像を用い実験的に求めた物を用いている。

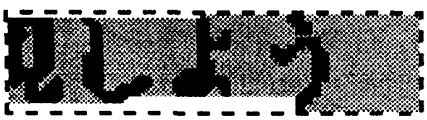

図 4: 縮退の不完全な領域

図 5: 補正処理後の領域

1. 本文文字列：水平・垂直方向ランレングス分布の平均のうちどちらかが 2.1 で求めた文字列幅の範囲にあり、線密度の平均・分散がそれぞれある範囲の値を持つ領域
2. 見出し領域：ランレングス分布の平均が文字幅の範囲にないが、線密度の平均・分散については上記と同様の性質を持っている領域
3. 写真・図：以上の条件を満たさず縦・横のサイズともに文字列幅よりも大きい領域

## 3 結果と考察

数種類の文書に対して本手法による実験を行った結果を以下に示す。ただし、抽出率については、全連結領域に対する人間が行う判別と一致した領域の割合(%)である。

実験試料と抽出率

| 種類 | 特徴 | 抽出率 |
|---|---|---|
| 新聞記事 | 縦組み、グラフ有 | 100 |
| 新聞記事 | 縦組み、写真・グラフ有 | 96 |
| 新聞記事 | 英字、横組み、写真有 | 100 |
| 論文 | 横組み、写真・数式有 | 87 |
| 論文 | 横組み、表・グラフ有 | 93 |

実験の結果比較的高い抽出率が得られた。上記表中抽出率87%と他より低い結果となった物があるが、これは、用いた論文中に数式がありこれらの幅は本文文字列より広いので見出し文字列として判別されてしまったためである。

この実験により、本手法のように単純で、しかも文書全体の構造上の性質などを用いず、領域の幾つかの特徴量を用いた方法でも比較的正確に文書の領域分割を行えることが分かる。

## 参考文献

[1] 中村納, 岡本教佳, 庭田剛, 南敏, “欧文テキスト画像における文字領域の抽出アルゴリズム” 電子通信学会, 1983.

[2] 秋山照雄, 増田功, “周辺分布、線密度、外接矩形特徴を併用した文書画像の領域分割” 電子通信学会, 1986.

[3] Young Seak PARK, 海老名毅, 伊藤昭, “汎用的な文書画像の階層的領域分割と識別法” 電子情報通信学会, 1992.