白領域の有効面積に基づく文書画像領域分割

正員 黄瀬 浩 一 (大阪府立大学)

非会員 秋吉 裕 二 (大阪府立大学)

非会員 高松 忍 (大阪産業大学)

Page Segmentation Based on Effective Area of Background
Koichi Kise, Member, Yuji Akiyoshi, Non-Member (Osaka Prefecture University),
Shinobu Takamatsu, Non-Member (Osaka Sangyo University)

Page segmentation is the task of extracting components of a document such as columns, textlines, figures and tables from a page image. This paper presents a method of page segmentation by analyzing background (white areas) of a page image. For page images without skew, it is known that white areas can be represented as white rectangles each of which maximally circumscribes white pixels. In general, however, extracted white rectangles include gaps between characters, words and those in figures which cause erroneous segmentation. Thus white rectangles need to be selected for correct segmentation. The characteristic point of our method is that the selection is based on the simple measure called *effective area* which is to estimate the effectiveness of a white rectangle as a delimiter of components by taking account of proximity with surrounding black areas. Our method correctly segmented 92.7% of components in 154 images of Japanese and English pages with various layout and resolution.

キーワード：文書画像，領域分割，レイアウト解析，文書画像理解

1. まえがき

大量の印刷文書による情報洪水を緩和するために，印刷文書の自動的な電子化が望まれている。ここで，電子化の処理として，単なるデジタル画像への変換ではなく，記号データへの変換を考えると，文字や図表などの文書の構成要素に応じて個別の認識・理解処理を施す必要が出てくる。このためには，構成要素を文書画像から分離・抽出する処理，すなわち文書画像の領域分割が必要となる。

以上の観点から，これまでに多くの領域分割手法が提案されている⁽¹⁾。それらは，黒領域解析^{(2)~(4)}，白領域解析^{(5)~(7)}の2種類に大きく分類できる。

黒領域解析とは，文書の印刷部分を表す黒画素を統合することにより，構成要素を抽出する手法である。この手法は，「同じ構成要素に属する黒画素は近接している」という観点から領域分割を考えるものである。具体的には，近接した黒画素を統合していくことにより，単語，文字列，カラムなどの領域を生成する。この処理では，統合の中心となる黒画素近辺の局所的な情報に基づいて，統合するかどうかを決定する。

一方，白領域解析とは，文書画像の空白部分を解析することにより，構成要素を抽出する手法である。この手法

は，「異なる構成要素は白領域により分割されている」という観点から領域分割を考えるものであり，カラム間，行間などに存在する白領域を抽出し，これに基づいて領域を分割する。ここで，カラム間の白領域は複数の行の並びを反映し，行間の白領域は複数の文字の並びを反映していることから，より大域的なレイアウトの特徴に基づいているということができ，黒領域解析法と比べて効果的に領域を分割することが可能となる。

しかし，ページ内には様々な大きさ，形状を持つ白領域が存在し，これらすべてが領域分割に有効というわけではない。したがって，白領域解析では，領域分割に有効な白領域を抽出することが中心課題となる。

Bairdらは，文書画像で比較的大きな領域を占めるカラム間の白領域を抽出対象とし，矩形で表現した白領域(白矩形)を面積と形状により評価・選択するという簡便な手法を提案している⁽⁵⁾。また，Pavlidisらは，投影処理によりカラム間の白領域を抽出する手法を提案している⁽⁶⁾。以上の手法は，処理が比較的シンプルである反面，カラム領域を抽出するものであるため，後続の処理につなげるには，カラムから文字列を抽出する処理が別途必要となる。これに対してOkamotoらは，白矩形をいくつかの経験的規則により評価・選択することにより，図表などに加

え、文字列を直接抽出する手法を提案している⁽⁷⁾。この手法は文字列の抽出を対象としているが、規則が複雑であり、多くの処理パラメータを必要とするという問題点がある。

そこで本論文では、これらの問題点を考慮した白領域解析の一手法を提案する⁽⁸⁾。本手法では、Okamotoらの手法と同様、文書画像から文字列、図表、写真、罫線の領域を抽出することを考える。本論文では、「領域分割に有効な白矩形は、より多くの黒画素に近接・接触し、それらを分割するものである」と考え、これを具体化した有効面積と呼ぶ特徴量を提案する。本手法の特徴は、有効面積を用いて白矩形を評価・選択することにより、シンプルさを損なわずに文字列を直接抽出可能な点にある。

2. 要求と前提

本論文では、抽出対象の構成要素として、文字列、図、表、写真、罫線を考える。ただし、抽出した領域がいずれの構成要素に対応するか(領域分類)については考えない。

一般に、領域分割処理には、誤分割の少ないことが要求される。誤分割は、大きく過分割(over-segmentation)と過統合(under-segmentation)の2種類に分類できる。過分割とは、単一の構成要素を2つ以上の領域に分割するという誤りである。一方、過統合とは、複数の構成要素を1つの領域に統合するという誤りである。実際の誤分割には、過分割と過統合が混在することもある。

領域分割の結果を領域分類、文字切り出し・認識、論理構造化など、文書画像理解を構成する後続の処理へ入力することを考えると、過統合された領域を正しく再分割することは一般に困難を伴うと考えられる。一方、文書のレイアウト構造に関する知識などを用いれば、過分割された領域を再統合することは可能といえる。したがって、2つの領域分割法があり、誤分割が同程度に生じるならば、過統合の少ない方が優れた手法であると考えられる。

本論文では、次の条件を満たす文書画像を対象として領域分割処理を考えていく。

- 1. 2値画像として入力されること。
- 2. 文書画像は傾いていないこと。
- 3. 矩形レイアウトであること。ここで、矩形レイアウトとは、文書のすべての構成要素が、文書の境界に平行または垂直な辺を持つ矩形により囲むことができるものをいう。
- 4. 黒領域の大半が文字列領域であること。

以上の条件のうち、1.は単色刷りの文書を仮定し、2.は前処理として傾き補正処理を仮定していることを意味する。3.、4.は、雑誌の広告部分やチラシなどでは満たされないが、技術論文、書籍、新聞など広範囲かつ実用的な文書で満たされるものである。なお、矩形レイアウトを持つ文書が、傾きのない画像として取り込まれた場合、垂直あるいは水平方向の辺を持つ矩形によりすべての構成要素

図 1 白矩形

Fig. 1. White rectangles

を囲むことができる。以下では、このような辺を持つ矩形を単に矩形と呼ぶ。

3. 白領域解析とその問題点

白領域解析により領域分割を行なうためには、文書画像中の白領域を抽出・表現する必要がある。Bairdらは、その一手法として、極大空白矩形(maximal empty rectangles)を提案している⁽⁵⁾。極大空白矩形とは、白画素のみを含む矩形のうち、それ以上領域を広げることができないものをいう。したがって、極大空白矩形の各辺は、黒画素あるいは画像の端に接することになる。本論文では極大空白矩形を白矩形と呼ぶ。

実際の文書画像を対象に白矩形を抽出すると、非常に多数の白矩形が得られる。この中には、図と本文の間のマージン、カラム間、行間などを表す白領域の他、文字間の白領域などが含まれる。したがって、白矩形に基づく領域分割では、いかに領域分割に必要な白矩形を選択するかが中心的な課題となる。

Bairdらは、カラムを抽出する目的のため、「カラム間の白矩形は面積が大きくかつ細長い」との観点から、白矩形を面積とアスペクト比(白矩形の長辺に対する短辺の比)により評価するという簡便な手法を提案し、実験により有効性を検証している。しかしながら、これを文字列の抽出にそのまま利用することはできない。例として図1に示すA～Eの白矩形を考えてみる。ここでC～Eは文字列、ブロックを分ける白矩形であるため領域分割に必要なものであるが、A、Bは文字列を分断するため排除しなければならない白矩形である。ところが、面積、アスペクト比の観点からは、白矩形A～Dはほぼ等価であり区別がつかない。したがって、白矩形を用いて文字列を抽出するためには、異なる評価基準が必要となる。

4. 白矩形の有効面積

領域分割における白矩形の役割は、文書画像中の黒画素を上下あるいは左右に分割することであると考えられ

図2 有効面積

Fig. 2. Effective area

図3 黒矩形と白矩形のギャップ

Fig. 3. Gap between a black and a white rectangle

る。図1を見ると、白矩形A、BとC、Dの大きな差異として、C、Dは多数の黒画素の間に存在し、それらを上下に分割しているが、A、Bは大部分が白矩形Eと重なるため、ほとんど黒画素を分割していないことが挙げられる。本論文では、このような観点から白矩形を評価する基準として、白矩形の有効面積を提案する。

以下では図2を用いて有効面積の定義を述べる。

まず、文書画像に対して、黒領域を囲む矩形(黒矩形)を考える。黒矩形は、抽出対象の構成要素の一部あるいは全部を囲むような矩形であり、互いに重複しないものとする。例えば、8連結黒領域を囲む矩形のうち互いに重複するものを統合して得られる矩形を考えればよい。

次に、白矩形に含まれる白画素について考える。縦方向あるいは横方向の白画素の連鎖を白ランと呼ぶ。白ランの少なくとも一方の端が黒矩形と接している場合、その白ランを有効な白ランと呼ぶ。図2の白矩形に含まれる白画素のうち、灰色の部分、縦方向と横方向の有効な白ランが交差する箇所を表している。白矩形の有効面積は、そのような白画素の面積の総和として与えられる。

さて、実際の文書画像では、1つの白矩形の周囲に多数の黒矩形が存在しており、その大部分は白矩形の辺に近接するだけである。このような場合でも、白矩形は黒矩形の分割に寄与すると考えられることから、本手法では、接触に加え近接の場合も考慮して有効面積を求めることとする。具体的には、図3に示すように、白矩形と周囲の黒矩形の間にギャップ g_v あるいは g_h を考え、これが閾値 G 以下のとき、接触と同等に扱って有効面積を求める。

5. 領域分割手続き

本手法は、黒矩形抽出、白矩形抽出、白矩形選択、領域抽出の4ステップの処理からなる。図4(a)の入力画像(一

部分を図示)に対して各ステップの処理を施した結果を、図4(b)～(e)に示す。ここで、処理に用いたパラメータの値は、6章の実験で用いたものと同じである。以下では、各ステップの詳細について述べる。

(1) 黒矩形抽出

入力画像に対してノイズ除去を行なった後、黒矩形を抽出する。具体的には、8連結黒領域を囲む矩形をまず求めておき、次に矩形のうち重複するものを統合する。図4(a)からは図4(b)のような黒矩形が得られる。

この図からもわかるように、抽出される黒矩形の大部分は本文中の文字に対応する。そこで本手法では、本文文字を矩形で囲んだ際の横幅(C_w)、縦幅(C_h)を、黒矩形の横幅、縦幅の最頻値として推定する。得られた C_h 、 C_w は、後の処理でパラメータを設定する際に用いる。

(2) 白矩形抽出

次に、付録に示す手続きにより、黒矩形から白矩形を求める。ところが、この手続きにより得られる白矩形の大半は、明らかに領域分割に不要なものである。そこで、後の処理の手間を軽減するために、面積が非常に小さく明らかに不要と思われるものを排除しておく。3章でも述べたように、面積による白矩形の排除には限界があるため、不要な白矩形が残る。図4(c)の例では、黒丸で囲まれた部分にそのような白矩形が残っている。

(3) 白矩形選択

残りの白矩形に対して有効面積を求め、不要な白矩形を発見・除去する。具体的には、白矩形と黒矩形のギャップを $G = \gamma \cdot C_w$ として有効面積 E_a を求め、以下の条件、

$$E_a < \delta_1 \cdot C_w \cdot C_h ; W_w \geq W_h \text{ のとき、}$$

$$E_a < \delta_2 \cdot C_w \cdot C_h ; \text{その他のとき}$$

を満たす白矩形を除去する。ここで、 W_w, W_h は白矩形の横幅、縦幅であり、 $\gamma, \delta_1, \delta_2$ は処理パラメータである。

ギャップの閾値 G の値を設定する際に、ギャップの方向によらず C_w を基準として用いるのは次の理由による。和文文書の場合には文字を囲む矩形がほぼ正方形であるため、 C_w, C_h のいずれを用いてもほとんど差がでない。また、英文文書の場合には正方形ではないため、 C_w と C_h の値は異なるが(一般に、 $C_w < C_h$)、ギャップの閾値 G は文字の形とは関係がないため、その影響を避ける方がよい。そこで本手法では、 C_w, C_h の一方のみ(ここでは C_w)を用いている。 C_h を用いる場合でも、処理パラメータの値を変更すれば同様の効果は得られるものと考えられる。

図4(d)では、図4(c)の不要な白矩形がすべて除去されている。

(4) 領域抽出

残りの白矩形により囲まれた黒矩形を統合し、構成要素の領域を抽出する。図4(e)では、すべての構成要素が正しく抽出されている。

(a) 入力画像

(b) 黒矩形

(c) 白矩形

(d) 選択された白矩形

(e) 構成要素

図 4 処理例

Fig. 4. Examples of processing results

表 1 実験サンプル

Table 1. Samples of documents used in the experiment

	解像度	言語	文字列	画像数
1 論文画像	400dpi	和文	横書き	94
2 JEIDA'93 文書 画像データベース	150dpi	和文	横書き	9
			縦書き	20
3 UW Database ⁽⁹⁾	300dpi	英文	横書き	31

表 2 処理パラメータ

Table 2. Values of processing parameters

	γ	δ_1	δ_2
和文横書き	0.25	4.5	2.25
英文	0.75	0.5	12.0

6. 領域分割実験

6.1 実験サンプルと条件

本手法の有効性を検討するため、表 1 に示す 154 ページの文書画像に対して実験を行なった。これらのサンプルは、1 カラム～3 カラムのレイアウトの文書からなる。1 の論文画像は、本実験のために電子情報通信学会、情報処理学会の学会誌・論文誌から 400dpi のスキャナを用いて作成した画像である。2 の JEIDA'93 文書画像データベースからは、小説、実用書、教科書、論文の画像を実験に用いた。このデータベースに収められた文書画像の解像度は 600dpi であるが、本実験ではこれを 150dpi に減じて用いた。ワシントン大学で作成された文書画像データベース (UW Database) からは、英文雑誌、論文誌の初回コピーから作成された画像を用いた。

処理パラメータは、和文横書き、和文縦書き、英文の 3 種類の文書に応じて、表 2 のように設定した。ここで、和文縦書きに対しては、和文横書きの δ_1 、 δ_2 の値を交換して用いた。英文と和文でパラメータの値が異なるのは、

- 英文の場合、アセンダ (ascender) とディセンダ (descender) の影響により、和文に比べて、同一文字列内で黒矩形の並びが乱れる、
- 英文の行間隔は和文に比べてかなり狭いが、カラム間隔は比較的広い、

などの理由による。

また、本実験では次の条件を満たす領域を正解とした。

文字列 本文文字列については、カラム領域内に存在し、かつ文字の並ぶ方向 (横書きであれば左右、縦書きであれば上下) に分割されていないこと。図表のキャプションなど、その他の文字列については、文字列のはじまりから改行までで分割されないこと。

図・表・写真 1 つのキャプションに対して、1 つの図表領域を考え (キャプションは図表に含めない)、図表領域を分割しないこと。図表の中には (a)、(b) などのサ

表 3 実験結果
Table 3. Experimental results

構成要素	構成要素数	抽出率 (C)	
		比較手法	本手法
文字列	9110	78.5%	93.4%
図・表・写真	218	62.8%	61.9%
罫線	79	97.5%	97.5%
全体	9407	78.3%	92.7%

表 4 誤分割の内訳
Table 4. Details of erroneous segmentation

(a) 過分割数			
構成要素	比較手法		本手法
文字列	1893	(20.8%)	493 (5.4%)
図・表・写真	77	(35.3%)	72 (33.0%)
罫線	0	(0%)	0 (0%)
全体	1970	(20.9%)	565 (6.0%)

(b) 過統合を含む誤分割数			
構成要素	比較手法		本手法
文字列	67	(0.7%)	109 (1.2%)
図・表・写真	4	(1.8%)	11 (5.0%)
罫線	2	(2.5%)	2 (2.5%)
全体	73	(0.8%)	122 (1.3%)

(c) 誤分割に対する過分割の割合		
構成要素	比較手法	本手法
文字列	96.6%	82.0%
図・表・写真	95.1%	86.7%
罫線	0%	0%
全体	96.4%	82.2%

ブキャプションを伴う、複数の部分的な図表からなるものもあるが、これらについても分割されないものを正解とした。写真については1枚を1つの領域とした。

罫線 連結した罫線を分割しないこと。

6.2 結果と検討

(1) 比較実験

領域分割に白矩形の有効面積を導入すると過分割を抑制する効果があるが、逆に過統合は増加する可能性がある。したがって、有効面積が領域分割に有効であるためには、過統合をあまり増加させずに過分割を十分抑制する必要がある。この点を評価するため、本手法から白矩形選択のステップ(ステップ3)を除いた手法を比較手法とし、本手法との比較実験を行った。

実験結果を表3に示す。ここで、抽出率(C)とは、正しく抽出された構成要素数を m 、文書に含まれる構成要素数を n としたとき、 $C = m/n$ により定められる値である。また、誤分割の内訳を表4に示す。表4(a)は、過分割数

図5 英文文書の処理例

Fig. 5. Processing results of an English document

独により誤分割となった構成要素数(過分割数)を表す。一方、表4(b)は残りの誤分割数、すなわち、過統合を含む誤分割数(過統合単独、および過統合と過分割の混在による誤分割数)を表す。表4(a)、(b)において括弧内の数字は、構成要素数に対する、過分割数ならびに過統合を含む誤分割数の割合を示す。表4(c)は、誤分割された構成要素数を e 、過分割数を e_s とすると、 e_s/e を表す。

まず、構成要素の種類ごとに見ていく。文字列については次のことがわかる。表4(b)によると、比較手法に対して本手法では過統合を含む誤分割数が増加している。しかしながら、表4(a)に示すとおり、本手法は比較手法の過分割数を1/4程度まで減少させている。その結果、表3に示すように、抽出率が大幅に改善されている。このことから、文字列の抽出には有効面積が効果的に作用していると考えられる。一方、図・表・写真については、比較手法と比べて本手法では過分割数が多少減少しているものの、過統合を含む誤分割数が増加したため、結果として抽出率が多少低下している。したがって、図・表・写真の抽出には有効面積の効果は低いといえる。罫線については、比較手法、本手法ともに過統合により誤分割となった。

全体的には次のようにいえる。有効面積を用いない場合には、表4(c)の比較手法の欄に示すとおり、誤分割の大半が過分割によるものである。これに対して、有効面積を導入することにより、過分割数が全構成要素数の20.9%から6.0%に減少し、抽出率が78.3%から92.7%に改善されている。一方、過統合を含む誤分割数は0.8%から1.3%に増加したが、それほど大きな差とは考えられない。したがって、全体的には有効面積を導入する効果は高いものといえる。

(2) 処理例と考察

以下では、処理例を示しながら本手法について詳しく考察する。

まず対象文書ごとに処理例を示す。和文横書き文書に対する例はすでに図4に示しているの、ここでは英文文書、和文縦書き文書の処理例を、それぞれ図5、図6に示す。図5では、左カラムのグラフで過分割が生じている

図 6 和文縦書き文書の処理例

Fig. 6. Processing results of a vertically written Japanese document

図 7 図の過分割

Fig. 7. Over-segmentation of a figure

が、それ以外の構成要素はすべて正しく抽出されている。図 6 では、ページ下の横書き文字列¹が誤って分割されているものの、その他の構成要素はすべて正しく抽出されている。特に本手法では、文字幅の推定値をもとに処理パラメータを変化させているため、種々の解像度を持つ画像にも対処可能となっている。

次に、構成要素の種類ごとに誤分割について考察する。

図・表・写真に対しては、特に多くの図で過分割が残ったため、結果として抽出率が低下した。実験対象の図の中には、いくつかの部分的な図から構成されるもの、図に文字列が混在したものや、プログラムリスト(すべて文字列からなる)を図としたものがあり、これらの図に対して、本手法では過分割が多く見受けられた。過分割の例を図 7 に示す。

文字列についても誤分割の大半が過分割であった。文字列に対する過分割の例を図 8 に示す。この例は論文の参考文献部分を示しており、(1), (2) などの文献番号が残りの文字列から分離されている。

本手法では有効面積を用いて白矩形を評価するため、縦幅あるいは横幅の大きな黒矩形に近接・接触する白矩形を選択する傾向がある。また、個々の幅は小さくても、揃っ

¹ この文字列は、画像作成時に付加されたものであり、オリジナルの文書には含まれない。

- [1] Soto O.G.: "On the Integrated Design and Planning for Advanced Networks", *Computer Networks and ISDN Systems*, 20, 1-5, pp. 89-94 (1990).
- [2] Wang J.L.: "INPLANS-An Integrated Planning and Engineering System for Telecommunication Networks", *ITC 13, Systems Engineering and Traffic Engineering* 分冊, pp. 137-142 (1991).
- [3] Gruszecki M. and Andries R. N.: "Some New Concepts in Demand and Traffic Forecasting and Planning of Future Telecommunication Services", *Computer Networks and ISDN Systems*, 20, 1-5, pp. 65-74 (1990).
- [4] Kapp Y.: "The Use of Dynamic Programming for Planning of Extensions of Conduits and Main Cable Networks in a Local Exchange Area", *Ericsson Rev.*, 46, 3, pp. 102-112 (1969).

図 8 文字列の過分割

Fig. 8. Over-segmentation of textlines

図 9 過統合と過分割の混在

Fig. 9. Mixture of over- and under-segmentation

て並ぶ黒矩形があれば、それらに近接・接触する白矩形を選択する傾向がある。通常、これらの傾向は領域分割に都合良く作用するが、上記のような過分割の原因にもなっていると考えられる。本手法のように対象文書に固有の知識を用いずに処理を行なう場合、上のような過分割、すなわち、部分図を分離したり、図中の文字列を分離するなどの誤りは避けられないものと考えられる。また、本実験では、文字列の正解条件から、図 8 を過分割と判定したが、後続の理解処理などを考えれば、むしろ分離されていた方が都合がよいとの見方も可能であろう。2 章にも述べたように、この種の過分割は、文書のレイアウト構造に関する知識や文字認識結果などを参照すれば、かなりの程度、再統合可能である⁽¹⁰⁾。

誤分割の残りは、過統合、および過分割と過統合の混在によるものである。例を図 9 に示す。この文書では、左上部に式が記述され、右上部にプログラムリストが図として記述されており、式と図の一部が過統合されている。この理由は、式と図の双方で文字列がインデントされているため、左右カラムを分ける白矩形の有効面積が小さくなったことと考えられる。その他にも、フットノートなど、小さい文字からなる文字列は過統合される場合があった。

以上を総合すると、様々な解像度で与えられた和文横書き、和文縦書き、英文の 3 種類の文書に対して、本手法に

より構成要素の92.7%を正しく抽出できた。また誤分割の大半は比較的救済が容易な過分割であり、対処が困難な過統合を含むものは全構成要素の1.3%にとどまった。これから、本手法の有効性が確認できた。

7. むすび

本論文では、多くの黒矩形に近接・接触する白矩形は領域分割に有効であるとの観点から、文字列、図表、写真、罫線を抽出するための白領域解析法を提案した。本手法の特徴は、黒矩形との近接・接触を評価する特徴量として有効面積を導入し、白矩形を有効面積に基づいて取捨選択する点にある。また、様々な解像度で与えられた、和文横書き、縦書き、英文の3種類の文書画像、合計154サンプルを用いた実験の結果、本手法は約93%の精度で領域を正しく分割できることがわかった。今後の課題としては、抽出した領域の分類処理を検討することなどがある。

謝辞 本研究に対して種々御支援をいただいた本学 福永邦雄教授に感謝する。また、実験データの作成・整理に御協力いただいた本学学生 米田昇巧氏(現在、日本電気 勤務)、小山乃り子氏(現在、ジャストシステム 勤務)に感謝する。本研究の一部は文部省科学研究費の補助による。

(平成7年12月25日受付、同8年5月15日再受付)

文 献

- (1) Haralick, R. M.: "Document image understanding: geometric and logical layout," *Proc. CVPR'94*, pp.385-390(1994).
- (2) Wong, K., Y., Casey, R. G. and Wahl, F. M.: "Document analysis system," *IBM J. Res. Develop.*, Vol.26, No.6, pp.647-656(1982).
- (3) 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」, 電子情報通信学会論文誌(D), Vol.J69-D, No.8, pp.1187-1196(1986).
- (4) O'Gorman, L.: "The document spectrum for page layout analysis," *IEEE Trans. PAMI*, Vol.15, No.11, pp.1162-1173(1993).
- (5) Baird, H. S., Jones, S., E. and Fortune, S. J.: "Image segmentation by shape-directed covers," *Proc. of 10th ICFR*, pp.820-825(1990).
- (6) Pavlidis, T. and Zhou, J.: "Page Segmentation and Classification," *CVGIP: Graphical Models and Image Processing*, Vol.54, No.6, pp.484-494(1992).
- (7) Okamoto, M. and Takahashi, M.: "A hybrid page segmentation method," *Proc. of 2nd Int'l Conf. on Document Analysis and Recognition*, pp.743-748(1993).
- (8) 秋吉裕二, 黄瀬浩一, 高松忍, 福永邦雄: 「空白の構造に基づく文書画像の領域分割」, 電子情報通信学会技術研究報告, PRU94-101(1995).
- (9) Phillips, I. T., Chen, S. and Haralick, R. M.: "CD-ROM Document Database Standard," *Proc. of 2nd Int'l Conf. on Document Analysis and Recognition*, pp.478-483(1993).
- (10) 小山乃り子, 黄瀬浩一, 高松忍: 「適用性を高めた印刷文書の論理構造化の一手法」, 1995年電子情報通信学会総合大会, D-581(1995).

付録 (白矩形抽出手続き)

本手続きは、基本的には文献(5)によるが、説明の簡便化のためにオリジナルとは異なる形で示す。効率的なインプ

(a) support

(b) open と close

図10 白ランと黒矩形の関係

Fig. 10. Relation between a white run and black rectangles

リメントや計算量については、文献(5)を参照されたい。

まず、手続きの説明に必要な諸概念を定義する。画像の左上を原点とし、横方向に x 軸(原点右側が正)、縦方向に y 軸(原点下側が正)をとる座標系を考える。矩形 r を左上の点 (x_s, y_s) 、右下の点 (x_e, y_e) の座標により、

$$r = (x_s, x_e, y_s, y_e)$$

のように表す。特に画像全体を囲む矩形を、

$$(0, x_{max}, 0, y_{max})$$

とする。また、画像から抽出された黒矩形の集合を、

$$B = \{b_1, \dots, b_n\} \text{ ただし, } b_i = (x_{si}, x_{ei}, y_{si}, y_{ei})$$

で表す。

x 軸に平行で、左端点 (x_l, y) 、右端点 (x_r, y) を持つ白ラン w を、

$$w = \langle [x_l, x_r], y \rangle$$

と表す。図10(a)に示すように、白ラン $w = \langle [x_l, x_r], y \rangle$ が与えられたとき、 w の上側と接する黒矩形を考える。一般に、白ラン w の区間 $[x_l, x_r]$ には、黒矩形と接する部分とそうでない部分がある²。このうち、黒矩形と接する部分を区間 $s_1, \dots, s_k$ で表し、その集合を $\text{support}(w)$ とする。なお、白ラン w の上側に接する黒矩形がない場合には、 $\text{support}(w) = \emptyset$ とする。同様に、図10(b)に示すように、白ラン w の下側に接する黒矩形を考え、黒矩形と接する区間の集合 $c_1, \dots, c_k$ を $\text{close}(w)$ と表す。また、 $\text{close}(w) \neq \emptyset$ のとき、白ラン w のうち黒矩形と接触しない区間 $p_1, \dots, p_k$ の集合を、 $\text{open}(w)$ と表す。

次に白矩形の求め方について述べる。 $S = \text{support}(w) \neq \emptyset$ であり、かつ両端点が黒矩形あるいは画像の端に接するような白ラン $w = \langle [x_l, x_r], y_u \rangle$ が与えられたとき、その白ランの一部または全部を上辺に含む白矩形は、図11の手

²白ラン w の上側に黒矩形 b_i が接するとは、 $y = y_{ei} + 1$ かつ $\max(x_l, x_{si}) \leq \min(x_r, x_{ei})$ が成り立つことである。

Procedure search(w, S)
 step1 $\text{close}(\langle [x_r, x_l], y \rangle) \neq \emptyset$ かつ $y \geq y_u$ を満たす y のうち、最小のものを y_d とする。
 step2 そのような y_d が存在しない場合、
 ($x_l, x_r, y_u, y_{\max}$) を白矩形とする。
 step3 一方、 y_d が存在する場合:
 (a) (x_l, x_r, y_u, y_d) を白矩形とする。
 (b) $Q = \{q_i | q_i \in \text{open}(\langle [x_r, x_l], y_d \rangle) \text{ かつ } \text{support}(\langle q_i, y_u \rangle) \neq \emptyset\}$
 (c) $Q \neq \emptyset$ ならば、各 $q_i \in Q$ について、
 search($\langle q_i, y_u \rangle, \text{support}(\langle q_i, y_u \rangle)$) を実行する。

図 11 白矩形抽出手続き “search”
 Fig. 11. Procedure “search”

続き search によりすべて求めることができる。この手続きは、まず、 w の全体を上辺とする白矩形を抽出し (step2 および step3 の (a))、次に w の区間のうち白矩形の上辺となる部分を求め (step3 の (b))、その部分を上辺とする白矩形を再帰的に抽出する (step3 の (c)) ものである。

図 12 の例を用いて具体的に説明する。いま、画像中には図 12 に示す黒矩形 $b_1, \dots, b_5$ が存在し、 w_1 が同図のように与えられているとする。ここで、

$$S = \text{support}(w_1) = \{s_1, s_2\}$$

である。search(w_1, S) の step1 において、まず同図に示す y_d を求める。次に、step3 の (a) において、 w_1 全体を上辺とする白矩形 r_{w1} を抽出する。step3 の (b) では、

$$\text{open}(\langle [x_l, x_r], y_d \rangle) = \{p_1, p_2\},$$

$$\text{support}(\langle p_1, y_u \rangle) \neq \emptyset,$$

$$\text{support}(\langle p_2, y_u \rangle) = \emptyset$$

なので $Q = \{p_1\}$ となる。したがって、白ラン $\langle p_1, y_u \rangle (= w_2)$ は白矩形の上辺となり得る。そこで、step3 の (c) で search($w_2, \text{support}(w_2)$) により、そのような白矩形を求める。search($w_2, \text{support}(w_2)$) では、 w_2 に対する y_d が存在しないので、step2 で白矩形 r_{w2} を抽出し、処理を終了する。

一般に、画像に存在するすべての白矩形は、次のように求めることができる。ただし、黒矩形は画像の周囲に接しないものとする。

1. search($\langle [0, x_{\max}], 0 \rangle, \{[0, x_{\max}]\}$) により、画像の上端に接するすべての白矩形を求める。
2. 黒矩形 $b_i \in B$ の座標から、 b_i の下辺に接し、かつ両端点が他の黒矩形または画像の端に接する白ラン w_i を求める。
3. 各白ラン w_i について、 w_i の一部あるいは全部を上辺に持つ白矩形を、search($w_i, \text{support}(w_i)$) により求める。ただし、同一の白ランに対しては、search は一度だけ適用する (複数の黒矩形から同じ白ランが得られることもあるため)。

図 12 白矩形抽出の例

Fig. 12. Example of extraction of white rectangles

黄瀬 浩一

(正員) 1986 年大阪大学工学部通信工学科卒業。1988 年同大学大学院工学研究科博士前期課程修了。1990 年大阪府立大学工学部助手。現在、同講師。パターン認識、画像理解の研究に従事。工学博士。電子情報通信学会、情報処理学会、人工知能学会各会員。

秋吉 裕二

(非会員) 1993 年大阪府立大学工学部電気工学科卒業。1995 年同大学大学院修士課程修了。現在、日立電子サービス (株) 勤務。在学中、文書画像理解の研究に従事。電子情報通信学会、人工知能学会各会員。

高松 忍

(非会員) 1971 年大阪府立大学工学部電気工学科卒業。1973 年同大学大学院修士課程修了。同大学助手、講師、助教授を経て、1996 年 4 月より、大阪産業大学教授。工学博士。自然言語理解、知識情報処理、仕様記述言語の研究に従事。電子情報通信学会、情報処理学会、人工知能学会各会員。