デジタル・ドキュメント　8－1
（1997．7．18）

空白領域に着目した文書画像の分割法とその識別法

浦田敏道**　海老名 毅*　猪木 誠二*　井上 彰**

* 郵政省通信総合研究所

** （株）エム研

あらまし

さまざまな形式の文書画像（新聞、論文等）を、コンピュータで解析し、それに含まれる情報を抽出、理解、利用するといった研究が活発に行われている。これらの研究において、文書画像領域分割、領域識別は、重要な前処理的な役割を果たす。我々は、この領域分割、領域識別の手法について提案する。本方式は、領域分割対象画像の白画素に着目し、領域分割に有効な白画素矩形を抽出し、それらによって囲まれる領域を文書構成領域として抽出している為、複雑なレイアウトの文書に対して領域分割可能としている。また、領域識別では、特殊な周辺分布をとり、そのデータをウェーブレット変換に似た変換を施すことにより、識別の難しい表と図の識別を可能としており、文字領域、表領域、写真領域、図領域の分割を可能とした。

The Segmentation and Classification of Document Images Using Blank Spaces

Toshimich Urata**　Tsuyoshi Ebina*　Seiji Igi*　Akira Inoue**

* Communications Research Laboratory, MPT

** M. ken Co., Ltd.

Abstract

We propose a net method of document image segmentation and its classification. Our approach focuses upon white pixels on image data, and we detects the white rectangular areas to separate the document image. The segmented images are classified into figure, table, and picture region by applying wavelet-based transformation.

1 まえがき

さまざまな形式の文書画像（新聞、論文等）を、コンピュータで解析し、それに含まれる情報を抽出、理解、利用するといった研究が活発に行われている。これらの研究において、文書画像領域分割、領域識別は、重要な前処理的な役割を果たす。前者は、文書画像を解析し、それぞれの性質が異なる文字領域、図領域、表領域、写真領域などの領域を分離抽出する。後者は、領域分割されたそれぞれの領域に対し、適切な識別、分類を行う。領域分割、領域識別の後、それぞれ抽出された領域に対し、認識、理解処理を施すことによって、コンピュータにその内容を理解させることができる。領域分割で、近年研究が盛んに行われているものは、マンハッタンレイアウト画像に対する領域分割である。マンハッタンレイアウトとは、構成要素の境界を互いに水平、もしくは、垂直な線分によって表現できるレイアウトである。このマンハッタンレイアウト文書画像に対して、領域分割可能であれば、印刷文書の大部分について、領域分割、識別が可能であると思われる。また、領域識別では文字領域、図領域、表領域、写真領域の識別のうち、図領域と、表領域の判別が難しいとされている。4 画素方形のパターンの出現頻度により、図領域、表領域の判別を行う手法（[1]）もあるが、罫線の無い表に対しては、判別不能としている。

我々は、この領域分割、領域識別の手法について提案する。我々の方式は、領域分割対象画像の白画素に着目し、空白領域から成る白画素矩形を抽出し、抽出された白画素矩形のうち、領域分割に有効な有効白画素矩形によって囲まれる領域を文書構成領域として抽出している為、複雑なレイアウトの文書に対して領域分割可能である。また、水平垂直方向に同等な処理を施すので、領域分割対象画像が縦書き文書、横書き文書にかかわらず、マンハッタンレイアウト文書であれば分割可能である。また、領域識別では、特殊な周辺分布をとり、そのデータをフレンチハット基底関数に似た独自の基底関数を用いてウェーブレット変換（使用している関数についての完全正規直交性は、未確認）することにより、識別の難しい表と図の識別を可能としており、文字領域、表領域、写真領域、図領域の分割を可能とした。罫線のない表についても識別可能で、株式相場、天気予報、相撲の星取表などに対しても表と判別可能にである。

2 領域分割法

幅、高さが、あるしきい値以上の、内部に白画素のみを含む白画素矩形を抽出し、それらを解析して領域分割に有効な白画素矩形を抽出、生成する。その後、それら有効白画素矩形以外の領域を、文書を構成する領域として抽出し、その際それぞれの領域の大きさ及び位置関係に着目し、一まとめにする。

以下本手法について詳しく説明する。

2.1 有効白画素矩形の抽出

有効白画素矩形の抽出では、まず領域分割対象画像の白画素矩形を抽出する。その後、以下に示す白画素矩形のサイズの調整、白画素矩形の統合、白画素矩形の分割を、矩形変

更もしくは除去が行われなくなるまで繰り返す。最終的に得られた白画素矩形を領域分割に有効な白画素矩形として、それらを用いて、領域分割を行う。

2.1.1 白画素抽出の処理手順

以下に記す点Aをラスタ走査上を移動させ、点Aが対象画像の右下隅点まで走査した時点で処理終了とする。

既に入手した白画素矩形内に含まれず、かつ黒画素でない点Aに対して以下の1,2の処理をおこなう。

画面上の点A(ax, ay)としたとき次の手順を行う。

1. 点Aから下方へ黒画素を探しに行き黒画素を発見した点を点B(bx,by)とし、(by-ay)があるしきい値以上であれば、白画素矩形WhiteRect(ax, ay, bx, by) (但し( )内は右からleft,top,right,bottomである)を作る。
2. 点Aは固定したままで、点A'(ax+1, ay)が、黒画素もしくは、既に抽出された、白画素矩形に含まれる点であれば、WhiteRectを白画素矩形と決定し、処理3を行う。またそうでない点A'に対しては、点A'から下方へ黒画素を探しに行き黒画素を発見した点を、B'(ax+1,by')とし、|ay-by'|があるしきい値より小さいか、または、|(by-ay)-(by'-ay)|があるしきい値よりも大きければWhiteRectを白画素矩形と決定し、処理3を行う。そうでなければWhiteRectをWhiteRect(ax, ay, ax+1, min(by, by'))とし、点A'をA' = (ax+2, ay)として、WhiteRectが決定するまで、この処理を繰り返す。
3. 決定された白画素矩形の幅があるしきい値以上であれば、その矩形を白画素矩形グループの要素として、格納する。そうでない矩形に対しては、無視する。

上記の処理を実行した後、格納された白画素矩形データを垂直方向白画素矩形グループとして保存しておき、次に画像を90度回転させたものに対しても、上記と同じ処理を行い、得られた白画素矩形グループを水平方向白画素矩形グループとして保存する。

2.1.2 白画素矩形のサイズの調整

垂直水平白画素矩形グループに含まれる、矩形のサイズを調整する。まず垂直方向白画素矩形グループのある要素矩形を一つ取り出し、その垂直方向白画素矩形と交わる全ての水平方向白画素矩形を入手する。

入手した水平方向白画素矩形のうち、最上部にある矩形の上辺をこの垂直方向白画素矩形グループのある要素矩形の上辺とし、入手した水平方向白画素矩形のうち、最下部にある矩形の下辺をこの垂直方向白画素矩形グループのある要素矩形の下辺とする。その際あるしきい値以下の高さの横方向矩形は最上部最下部矩形とはなりえないものとするが、しきい値以下の高さの白画素矩形が縦方向に連続して配置されている場合それを一つの矩形とみなして処理する(図1)。以上の処理を垂直方向白画素矩形グループ

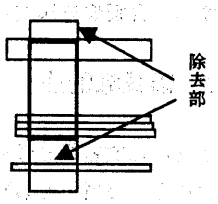

図1 サイズの調整

の全要素に対して行う。また、水平方向白画素矩形グループも同様に、水平方向白画素矩形グループの、ある要素矩形に交わる全ての垂直方向白画素矩形を取り出し、垂直方向白画素矩形と同様な処理を全水平方向白画素矩形に対して行う。

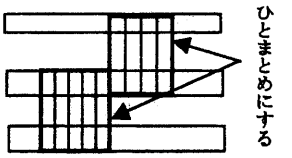

図2 矩形の統合

2.1.3 白画素矩形の統合

任意の白画素矩形が水平方向白画素矩形ならば高さ、垂直方向白画素矩形ならば幅がしきい値以下で同じ方向の白画素が隣接しているならばその白画素を無効矩形として有効白画素矩形グループから解放する。

また、任意の白画素に同じ方向の白画素が隣接し、それらが同一の直交する方向の白画素矩形と交わり部分を持つのであれば、それらを一つの矩形として統合する。(図2)

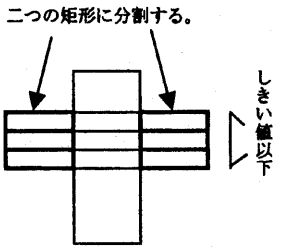

図3 矩形の分割

2.1.4 白画素矩形の分割

任意の水平（垂直）方向白画素矩形に交わる、垂直（水平）方向白画素矩形のうち、接しあっている矩形があればそれらの、矩形の幅（高さ）の総和、接する矩形が無く単独で存在するならばその矩形の幅（高さ）が、あるしきい値以下であるならば、この白画素矩形に交わらないように、また、この矩形が領域分割における役割をなくさないように二つの矩形に分割する。(図3)

2.2 文書を構成する領域の抽出

前節で述べた手法により、有効白画素矩形を抽出した後、領域分割対象画像全体から、全有効白画素矩形を除いた領域を、文字領域、表領域、図領域、写真領域、セパレータ等の何れかに属する文書画像を構成している領域とみなす。

2.3 文字領域識別法

抽出された文書構成領域の最頻出幅、高さを文字ブロック幅、文字ブロック高とし、幅、高さがそれらに相当する領域を文字領域として判別する。

2.4 文書構成領域のグループ化

この段階では、フォントサイズの大きい見出し文字などは、それぞれ一文字ずつ異なる領域として認識されているので、これらの領域をその大きさと並び方に着目して、一つの領域とする。

3 領域識別法

以下に記す領域識別方式では、水平垂直方向においてある特殊な周辺分布をとり、サポートが可変な関数と周辺分布の内積を特徴量とすることにより領域識別を行っている。そのため、表領域については、罫線の無い表領域についても、識別可能としている。

3.1 ブロック周辺分布

最頻出文字間隔よりも大きく、最頻出文字フォントの幅、高さよりも小さい値を定数 length ∈Zとし、また、Z^2上の整数値関数 Image(x, y)を領域分割対象画像の点(x, y)における画素値（黒画素であれば1を白画素であれば0）を返す関数とする時、Z^2上の整数値関数 B_{HORZ}, B_{VERT}を式(2.1),(2.2)で定義し、水平ブロック周辺分布 S_{HORZ}と垂直ブロック周辺分布 S_{VERT}を式(2.3),(2.4)で定義する。

$$B_{HORZ}(x,y) = \min\left(1, \sum_{i=0}^{length-1} Image(x+i,y)\right) \qquad (x,y \in Z) \qquad (2.1)$$

$$B_{VERT}(x,y) = \min\left(1, \sum_{i=0}^{length-1} Image(x,y+i)\right) \qquad (x,y \in Z) \qquad (2.2)$$

$$S_{HORZ}(y) = \frac{1}{width} \sum_{i=0}^{width/\,length-1} B_{HORZ}(i \times length,\ y)\ length \qquad (2.3)$$

$$S_{VERT}(x) = \frac{1}{height} \sum_{i=0}^{height\,/\,length-1} B_{HORZ}(x,\ i \times length)\ length \qquad (2.4)$$

ここで、width,heightは、対象文書画像のドット単位の幅、高さである。

3.2 特徴量の定義

3.1節のブロック周辺分布を用いて特徴量を定義する。本方式で、特徴量として採用した量は、ブロック周辺分布を一次元ウェーブレット変換に似た変換を施して得られた量を特徴量として採用した。基底関数としては、我々独自で作ったものを使用し、その関数についての正規直交性と L^2空間における完全性などの確認をしていないので、「ウェーブレット変換である。」という記述は避けた。しかし、この場合特徴量を抽出するということなので、基底関数が完全正規直交基底である必要は無く、領域識別の特徴量抽出に相応しいハイパスフィルタであればよいのでは、ないかと思われる。今回我々が、作った基底関数というのは、フレンチハット基底に似た関数を採用した。その関数を式(2.5)に示す。この関数を採用した理由としては、ブロック周辺分布のデータから、表のような、一まとまりの黒画素部分がある規則に従って分布しているデータの特徴を上手く抽出できるのではないかという理由で採用した。以下に基底関数Hを記す。

$$H(x) = \begin{cases} -1 & (x \in (-1,-2) \cup x \in [1,2]) \\ 1 & (x \in [-1,1]) \\ 0 & otherwise \end{cases} \qquad (2.5)$$

ここでも、$x \in Z$である。

次に、特徴量I_{HORZ},I_{VERT}を式(2.6),(2.7)で定義する。

$$I_{HORZ}(y_0, j) = \sum_{y=-\infty}^{\infty} H(2^{-j} y - y_0) S_{HORZ}(y) \quad (2.6)$$

$$I_{VERT}(x_0, j) = \sum_{x=-\infty}^{\infty} H(2^{-j} x - x_0) S_{VERT}(x) \quad (2.7)$$

ここでも、$x_0, y_0, j \in Z$で、かつ$x_0, y_0, j \geqq 0$である。

3.3 領域識別処理

図 4.1、図 4.2 、図 4.3 、はそれぞれ、図、表、写真のブロック周辺分布です。図 4.4、図 4.5、図 4.6 は、図領域、表領域、写真領域の特徴量をグラフ化したものです。式(2.6),(2.7)からjを固定した時、x_0　又は、y_0の値が 1 増えると H は、正の方向に2^j移動するということが言えるので、I_{HORZ} 、I_{VERT}は S のサンプル数の $1/2^j$個しか存在しない。グラフには、右半分に j=0 のデータを、残された左半分のうち右半分に j=1 のデータを、また更に、残された左半分のうち右半分に j=2 のデータをといったグラフの書き方をした。写真は、ブロック周辺分布がほとんど 1.00 に近い値を取っており、特徴量は、ほとんど 0.0 で、ところどころピークが現れるが、j の値が大きいところに集まる傾向にある。表についてのブロック周辺分布は、表らしい特徴として、ピークが規則正しく連立するように、現れている。また、図についても同じようなピークの現れ方をするものがあるが、表では、表中に含まれる文字サイズなどの大きさがピーク幅として現れるので、図のピーク幅に比べて狭いという特徴がある。そこで、特徴量のグラフを見比べると、図 4.4, 4.5 中の丸で囲まれているところに違いが見受けられる。つまり、j の特定な範囲において表においては、大きなピークが現れているが、図については現れていない。これは、j の値によって基底関数 H のサポートが決まるので、表の周辺分布の特徴として現れるピークのピーク幅と、H のサポート長が近くなる j において表ではピークが現れる。

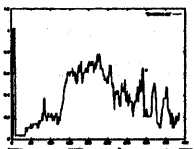

図 4.1 図のブロック周辺分布

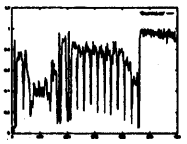

図 4.2 表のブロック周辺分布

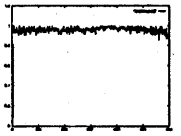

図 4.3 写真のブロック周辺分布

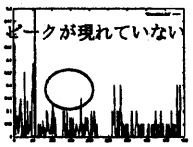

図 4.4 図の特徴量

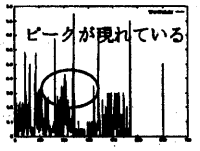

図 4.5 表の特徴量

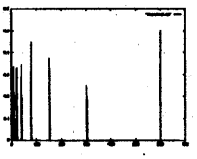

図 4.6 写真の特徴量

4 実験とその考察

図 5、はある A4 サイズ 200dpi の新聞画像を領域分割、領域識別した実験結果である。図領域を  、表領域を  、写真領域を  で表示した。図 5 の新聞画像には、表、写真、図がそれぞれ、一つずつ含まれており、それらについて正しく領域分割され、また識別されていることが分かる。この実験は、SUN の SPARC station 4 で行ったところ領域識別までで、40 秒かかった。本手法では、黒画素の集まりを一つの矩形として抽出する黒画素矩形抽出処理を省いて、直接、文書の空白領域に着目するので、処理時間はかなり削減されるかと思われたが、白画素矩形抽出後の有効白画素矩形抽出処理に時間がかかっているようである。領域識別では、写真領域の識別は、ほぼ 100%に近い確率で識別できているが、図と表の区別については、100%とはいかない。図領域の中には、ディザのかかった帯が規則正しく配置されているものや（図 6 ）、フローチャートのような図形でその中に含まれる、フォントが太字で領域全体に比べ大きい図領域（図 7）などは、表に似たブロック周辺分布をとる。この場合でも、特徴量は、やや図領域よりの特徴量をとるが、表との判別の、ぎりぎりの値をとるので、その識別については、なかなか 100%とはいかない。3.3 節で説明したように式(2.6),(2.7)より、j の値が大きくなれば、基底関数 H のサポートは、x0, y0 の値が 1 増える毎に、正の方向に 2^j 移動するが、これでは表のブロック周辺分布で、互いに接近しあうピークに対して上手く特徴量が抽出できないので、式(2.6),(2.7)の H の括弧内を $H(2^{-j}(y-y0))$、$H(2^{-j}(x-x0))$にすることによって、つまり x0, y0 が 1 増える毎に、H のサポートが正の方向に 1 増えるようにすれば、精度は幾分上がると思われる。又、文字領域の識別法において有効白画素矩形の最頻出幅、高さから文字領域を判別しているので、縦書き横書き混合文書については、その文書中に縦書きを多ければ横書きが、横書きが多ければ縦書きが、文字領域と判別されないので、その領域を領域識別処理することにより表領域として判別される。また、図 5 中の右上部に少々大き目のフォントサイズ 2 行の文字列が一まとまりの領域として抽出され、識別結果が“表領域”となっている。これも同様に、この領域が文字領域として判別されなかった為、領域識別処理により、表と判定されている

5 まとめ

本研究では、マンハッタンレイアウトの文書画像に対しての領域分割法と領域識別法について提案した。新聞画像について実験を試みた結果、図領域、表領域、写真領域についてある程度判別できた。しかし、新聞には、枠や、セパレータが複雑に存在している文書（例えばある写真領域がある枠に接していて、その枠は別の文書構成要素領域とも接している）、が多々存在する。そのような文書に対しては、上手く領域を抽出することができない。今後は、枠やセパレータについても何らかの解析処理を施し、その結果、複雑な枠、セパレータに対しても、対応できるよう検討する予定である。本研究は、郵政省の、情報通信基盤プロジェクトの研究の一部である。

図 5　領域分割識別の抽出結果

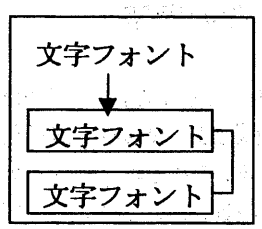

図 6 表とご認識されやすい図領域

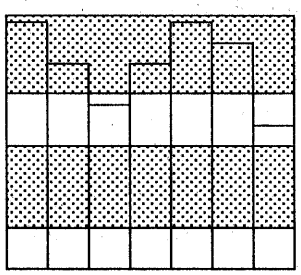

図 7　表とご認識されやすい図領域

参考文献

[1] 園田浩一郎、米田政明、長谷博幸、酒井充、丸山博：“統計的手法による文書画像中の領域解釈”,情報処理学会第 51 回 全国大会, 2-201

[2] 朴 栄碩、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”,信学論 D-II Vol.J75-D-II ,No.2 ,pp.246-256, (1992-02)

[3] 山下 晶夫、天野 富雄：“モデルに基づいた文書画像のレイアウト理解”,信学論 D-II Vol.J75-D-II, No.10, pp.1673-1681, (1992-10)

[4] 秋吉 裕二、黄瀬 浩一、高松 忍、福永 邦夫：“空白の構造に基づく文書画像の領域分割”, 信学技報, PRU, pp.94-101,(1995-01)

[5] 長谷博幸、辻正博、園田浩一郎、米田政明、酒井充：“汎用を目指した自動文書認識システム” ,信学技報, PRU, pp.94-33,(1994-09)

[6] John J.Benedetto 著：“ウェーブレット 理論と応用”, シュプリンガー・フェアラーク東京, (1995)

[7] 榊原 進 著：“ウェーブレット ビギナーズ ガイド”, 電機大出版局 ,(1995)