2H－4

黒画素塊の上下境界に着目した文書画像の解析方式

天野富夫
日本アイ・ビー・エム株式会社 東京基礎研究所

1　はじめに

文書画像の構造解析は既存印刷文書をコンピュータに自動入力するための重要な技術でありさまざまな手法が研究・開発されてきている。一般に画像の構造解析を効率的に行なうためには元画像に対して適当なぼかし処理を行なった縮退画像から出発して必要に応じて高解像度の画像（または元画像）を処理していく手法が有効である。文書画像では細長い行矩形が主な検出対象であるため、Run Length Smearing(RLS) と呼ばれる水平方向（横書き文書の場合）に黒ランをつなげていく処理が縮退画像生成のために用いられることが多い。このようなぼかし処理によって１）解析初期段階での処理コストの削減、２）大域的な情報（例えば複数の黒画素塊で構成される行矩形）の抽出が可能、などのメリットが得られる。本稿ではこのような縮退画像から出発して文書画像を解析するさいに用いる特徴として黒画素塊の上下境界線分に着目するアプローチを提案する。

2　上下境界線分に着目した特徴抽出手法

提案手法では上下境界線分に着目して二種類の矩形の集合を文書画像解析のための特徴として抽出している。それぞれの矩形は画像中の黒画素の密集した領域と残りの白画素領域に対応している。以後、これらの矩形を黒矩形と白矩形と呼んで区別することにする。

文書画像はRLSによる縮退画像生成のためラスター走査され、各走査線中の白ランのうち長さが閾値S以下のものを黒ランで置き換えた画像が生成される。境界線分のY座標を平滑化するため縦方向にも走査線H本単位でORをとりながら処理を行なっている。上下に連続する２本の走査線中のランデータを比較することにより上下の境界線分のランデータを得ることができる。白ランの下の黒ランは上境界、逆に白ランの上の黒ランは下境界とみなされる。図１は縮退画像および境界線分の検出例である。黒矩形は検出された上境界と下境界に挟まれた領域の座標情報、白矩形は下境界とその下にある黒画素塊の上境界で挟まれた領域の座標情報として計算される。具体的な手順を図２にしめす。ラスター走査の過程である注目走査線とその次の走査線のデータから境界線分をもとめるさい、バッファＡにはそれまでに検出された上境界がX座標でソートされて格納されているものとする。上下のランデータの比較によって境界線分は左から順に検出されてゆく。境界線分が検出されるとまずバッファＡ内のデータのうちこの境界線分より左にあるものがバッファＢにコピーされる。図で下境界３が検出されたときには上境界１が、上境界４が検出されたときには上境界2'がそれぞれバッファＡからＢにコピーされる。下境界が検出された場合には、バッファＡ内に対応する上境界が必ず存在するので上下で挟まれた領域を黒矩形として記録する。上境界データのうち下境界との対応がついた部分はバッファから削除する。例えば上境界２は下境界３と重なる部分を除いた2'に置き換えられる。上境界が検出された場合は、これをバッファＢの最後にセットする。この上境界と以前に検出された黒矩形の 底辺（下境界）で

図 1: RLS による縮退画像生成と境界線分検出

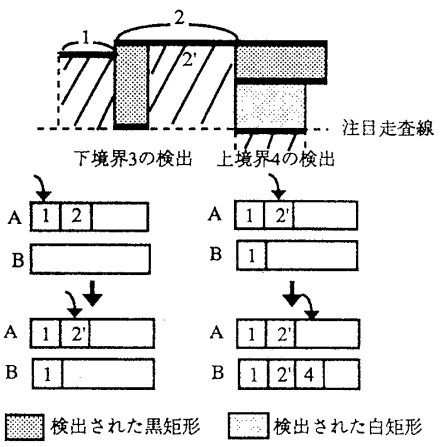


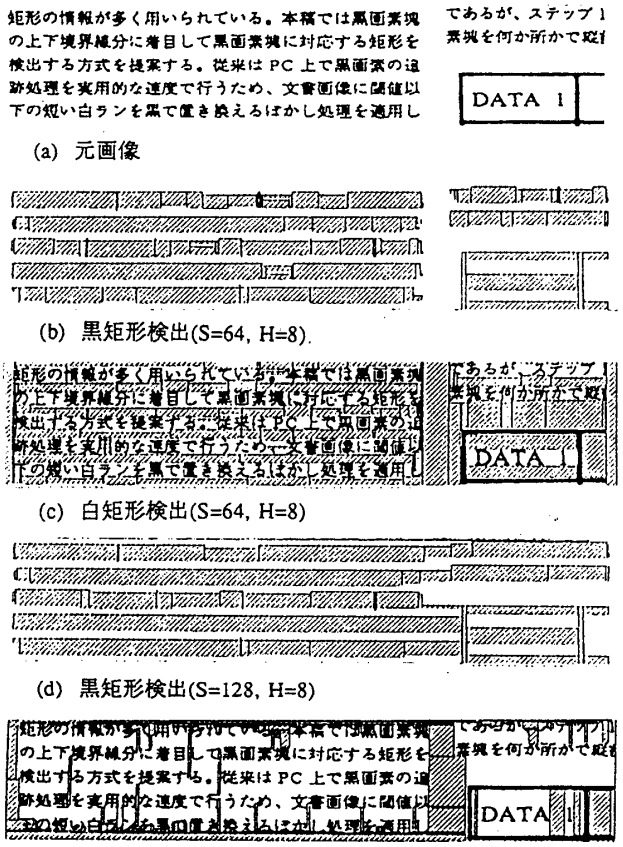
図 2: ダブルバッファを用いた白/黒矩形の検出

Document image analysis by using horizontal top/bottom boundaries
Tomio Amano
IBM Research, Tokyo Research Laboratory

挟まれた領域を白矩形として記録する。バッファ内の境界線分データはX座標でソートされた状態に保たれており、上下の境界の対応関係を効率的に調べることができる。注目走査線からの境界線分の検出が終了したら、バッファAとBを交換し次の走査線を処理してゆく。

3 上下境界線分特徴による文書画像解析

図3に黒矩形/白矩形の検出例をしめす。処理対象としたのは情処全国大会予稿集のページを400dpiの解像度でスキャンしたものの一部である。主走査を縦方向に行なって白矩形を検出する例（図3 (e)）は事前に画像を90度回転しておくことによって作成した。前回に報



(a) 元画像

(b) 黒矩形検出(S=64, H=8)

(c) 白矩形検出(S=64, H=8)

(d) 黒矩形検出(S=128, H=8)

(e) 縦方向スキャンによる白矩形検出(S=128, H=8)

図 3: 黒矩形/白矩形の検出例

告[1]したように検出された黒矩形の高さから文字列の候補を絞りこみ元画像からの情報を利用して文字列を同定する、といった利用のしかたが可能である。白矩形は黒矩形と補完関係にある特徴として今回新たに検出対象としたものである。文書画像解析処理においてはフィールドセパレータと呼ばれる大きな白画素の領域に注目してブロックの分離を行なったり[2]、行間の空白領域の情報から文字行が等ピッチで並んでいる領域を同定する[3]等の手法が用いられる。白矩形特徴はこの種の情報の収集に役立つものである。

RLSの閾値によっては異なるカラムに属するはずの文字列の間、あるいは文字列と非文字成分の間で連結が生じ矩形情報が正しく検出できない場合がある（図3 (d)）。このような場合でも主走査の方向を90度変えて得られる白矩形特徴には縦長の白ギャップの存在をしめす情報が保持されており過剰連結の有無をチェックすることができる。連結している黒画素に対する外接矩形をもとめる方法と異なり、本手法では過剰連結が生じても文字列や表を構成する水平/垂直線分ごとに複数の矩形が検出されるのでこれらの座標情報に基づいて元画像にアクセスすることにより比較的容易に過剰連結の影響を補正することが可能である。

4 戸籍画像処理への応用

筆者らはこれまでに本手法を図面内の文字読み取り[1]や論文フロントページの解析[2]に利用しているが、その他に戸籍入力用システム（法改正により戸籍事務業務が電算化されることになり膨大な既存戸籍をデータベースに入力する必要が生じている）においても白黒二値の戸籍謄本画像から垂直/水平線分を検出するために本手法を用いている。戸籍の場合、線分の太さはほぼ一定なので黒矩形集合の中から閾値処理によって縦長/横長の矩形を選びだしそれらを統合して線分としている。検出された線分データはOCRによる文字認識の前段階として1）スキューの検出、2）線分の位置関係から生年月日や身分事項などの領域を同定する、3）文字切り出しの前に線分を削除する、等の処理に利用されている。

5 まとめ

文書画像の解析を縮退表現の上で行なううえで黒画素塊の上下境界線分に着目して特徴を抽出する手法を提案した。本手法は一回のラスター走査で黒画素塊の密集した領域をしめす矩形と空白領域をしめす矩形の両方を検出することができる。黒矩形/白矩形と検出時の主走査方向（水平/垂直）の組み合わせによってRLSによる過剰連結に関して相補的な特徴を得ることができる。したがって縮退画像を使うメリットを活かしつつほかし処理の悪影響に対して頑健なアルゴリズムを構成することが期待できる。今後は、より広範囲の文書画像にこの手法を適用し汎用性を高めていきたいと考えている。

参考文献

[1] 天野：“上下境界線分に着目した文書画像からの黒画素塊検出方式”, 情処第47回全大 2-145 (1993).

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

[3] Hirayama Y.: “A Block Segmentation Method for Document Images with Complicated Column Structures”,Proc. of 2nd ICDAR , pp.91-94, (1993).