1P－2

シフト差分法による線幅特徴を用いた文書画像の領域分割

平本　建志　　松内　浩
（株）松下電器情報システム広島研究所

1.はじめに

印刷文書の電子化を実現するためには、文書画像を見出し領域、本文領域、図表等を含む付属領域に分割し、さらに、見出し領域、本文領域を構成する文字列を抽出する必要がある。秋山らの手法[1]は、代表的な領域分割手法の１つであり、基本３特徴を文書画像から適宜抽出することを基本としており、他の多くの手法もこの基本３特徴を基に領域分割を実現している。

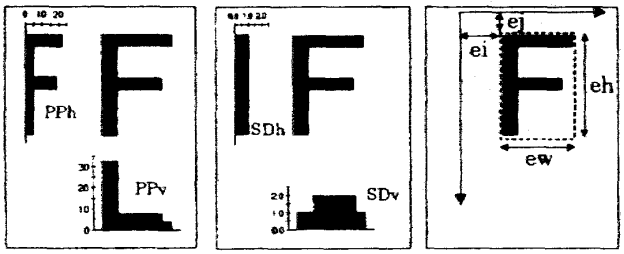
基本３特徴は、文書の大域的な性質を示す周辺分布特徴と線密度特徴、局所的な性質を示す外接矩形特徴からなる。周辺分布特徴と線密度特徴は領域分割過程で繰り返し抽出する必要があり、領域分割処理にかかる時間の大半を占める。

本稿ではこの基本３特徴の代替特徴として、シフト差分法による線幅特徴を用いることにより、領域分割における処理時間を削減することを提案する。

2.領域分割における基本３特徴

周辺分布特徴（垂直方向：PPv、水平方向：PPh）、線密度特徴（垂直方向：SDv、水平方向：SDh）、外接矩形特徴（e_i, e_j, e_h, e_w）の例を図１に示す。

領域分割過程において、外接矩形特徴の取得は、１度だけ行ないメモリ上に保持して使用される。対して、周辺分布特徴と線密度特徴は、罫線や空白領域により区切られる領域に対して再帰的に抽出が行なわれる。



(a)周辺分布特徴　(b)線密度特徴　(c)外接矩形特徴

図１　基本3特徴

A Method of Document-image Segmentation Based on Shift-Difference-Operation
Takeshi Hiramoto, Hiroshi Matsuuchi
Matsushita Information Systems Research Lab Hiroshima Co.,Ltd

図２に領域分割過程における段組部分の判別を行なう例を示す。図２(a)、図２(b)共に周辺分布 PPv から段組部分での分割候補位置として a1、b1 が抽出される。a1、b1 の空間は同じ幅であるが、線密度 SDv の値が大きな図２(a)は領域内に複数の行が存在している段組部分であり、値が小さな図２(b)は段組部分ではないと判別される。

3.線幅特徴を用いた領域分割

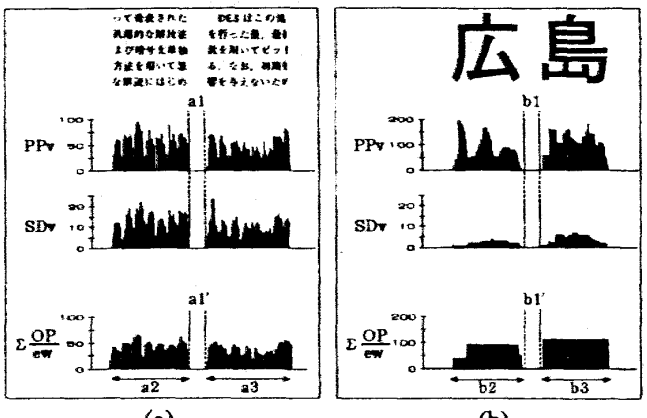
3.1 線幅特徴の概要

本稿で提案するシフト差分法による線幅特徴（以下、線幅特徴）は、外接矩形特徴に原画素数とシフト画素数の２組の値を付加したものである。**①原画素数** OP は、外接矩形特徴のもととなった連結する黒画素要素の画素数を計数することによって求まる値である（図３(a)）。**②シフト画素数**は、外接矩形特徴のもととなった連結する黒画素要素に対して、シフト差分法による空間フィルタを水平／垂直方向それぞれに施した際に得られる画素数を計数することによって求まる値（水平シフト画素数：SPh、垂直シフト画素数：SPv）である（図３(b), (c)）。

線幅特徴の取得は、外接矩形特徴の取得と同様に、１度だけ行ないメモリ上に保持して使用される。

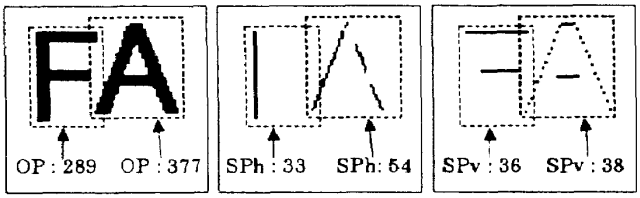
3.2 領域分割への適用

線幅特徴を基本３特徴の代替特徴として使用する場合、外接矩形特徴を除く周辺分布特徴と線密度特徴を利用する領域分割処理を線幅特徴により実現



(a)　(b)

図２　領域分割における段組判別



(a)原画素数　(b)水平ｼﾌﾄ画素数　(c)垂直ｼﾌﾄ画素数

図３ シフト差分法による線幅特徴

する必要がある。

周辺分布特徴に代わる値は、抽出対象領域内の各外接矩形に付属する原画素数 OP を外接矩形の幅 ew で割った値（OP／ew）を、画像の垂直方向画素列毎に累計して求める（周辺分布 PPv の代替値の場合。周辺分布 PPh の代替値では、幅は eh、水平方向画素列毎に累計）。図２の画像について抽出を行なった例では、周辺分布 PPv を用いた場合と同様に、代替値Σ(OP／ew)から分割候補位置 a1'、b1' が抽出される。

線密度特徴については、画像の水平方向、または、垂直方向の画素列毎に算出するが（図２中、SDv 参照）、線幅特徴を用いる場合には、抽出対象領域に対して１つの値を算出する。具体的に垂直方向の線密度特徴 SDv の代替値を求める場合、抽出対象領域内の各外接矩形に付属するシフト画素数 SPv を累計し、水平方向の領域幅で割ることにより求める。図２の画像について、垂直シフト画素数の累計値ΣSPv を、周辺分布特徴の代替値Σ(OP／ew)から求まる領域幅（図２中、a2+a3、または、b2+b3）で割った結果を表１に示す。線密度特徴を用いた場合と同様に、表１に示した代替値が大きな図 2 (a)は領域内に複数の行が存在している段組部分であり、値が小さな図2(b)は段組部分ではないと判別される。

表１ 線密度特徴の代替特徴

画像	線密度特徴の代替特徴
図２(a)	9.40
図２(b)	3.11

3.3 見出し抽出への応用

線幅特徴には、任意の領域内の各外接矩形に付属する原画素数の累計値を、シフト画素数の累計値で割った値が、その領域内における線幅の概算値となるという特性がある。したがって、領域分割の過程において、文字サイズだけでは判別不可能な太字で書かれている見出しの抽出等にも利用可能である。

4.評価

新聞１面の約 10%を解像度 400［DPI］で入力した画像（図４）に対する特徴抽出処理時間の評価実験を行なった。実験に使用した計算機は Pentium Pro 200MHz の CPU を搭載した DOS/V 互換機である。基本３特徴を使用した場合と、線幅特徴から代替値を算出した場合のそれぞれの処理時間を測定した結果を表２に示す。

表２ 処理時間測定結果

特徴抽出手法	処理時間
基本３特徴	1,161[ms]　(171[ms])
代替値使用	257[ms]　(195[ms])

表中の括弧内は、基本３特徴を使用する手法においては、外接矩形特徴の抽出に要した時間であり、代替値を使用する手法においては、線幅特徴の抽出に要した時間である。残りの再帰的に領域分割を行なう過程での処理時間が大幅に削減されていることがわかる。

また、見出し抽出の応用として、太字で書かれた見出しを含む画像と、画像中の各文字行領域における線幅の概算値を、線幅特徴から算出した結果を図５に示す。文字行領域 L1 の線幅概算値は、他の概算値を 75%以上、上回っており、太字による見出しの抽出が可能であることがわかる。

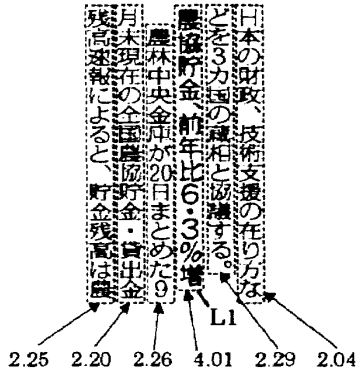


図４ 評価画像　図５ 線幅特徴による見出し抽出

5.おわりに

本稿では、文書画像の領域分割において、基本３特徴の代わりに、線幅特徴を用いることにより、処理時間を大幅に短縮できることを示した。また、線幅特徴は、文字サイズだけでは判別不可能であった太字による見出しの判別に利用可能である。

参考文献

1. 秋山他: "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", 信学論(D), J69-D, 8, pp.1187-1195(1986-08).