拡張スプリット検出法による文書構造解析

1K－3

中島 昇　田中 直哉　山田 敬嗣
NEC C&C メディア研究所
e-mail : {noboru | ntanaka | yamada} @ccm.cl.nec.co.jp

1. はじめに

印刷文書の電子化のためには、文書を単に部分領域に分割するのみならず、文書の階層的なレイアウト構造を解析し抽出することが重要である[1]。辻[4]は画素投影パターン上での再帰的な領域分割により、文書のレイアウト構造を抽出するスプリット検出法を提案している。この方法は、高速処理が可能であるが入り組みのある部分領域を分離できない。これに対して、背景部の大きな空白領域をセパレータとして領域分割する方法[2]やテクスチャ特徴を利用して文字領域とその他の領域とを識別する方法[3]は、部分領域間の入り組みがある場合にも領域分割できる。しかし、これらの方法だけでは、文書レイアウトの階層構造を再現する情報を抽出できない。本稿ではスプリット検出を仕切線、画像領域の周囲、大きな空白領域の情報を手がかりに非直線的なセグメンテーションに拡張し、入り組んだ文書構造に対して、領域分割を再帰的に行いながら階層的なレイアウト構造を抽出する方法を提案する。

2. スプリット検出法[4]

スプリット検出法は再帰分割の要領で、文書画像をページ画像全体からブロック・段組、文字列へと領域分割していく。再帰的な分割の過程で得られる階層的な包含関係をレイアウト構造として出力する。部分領域への分離の可否（スプリットの存在）とスプリットの方向・位置は、注目領域内の縦横各方向の画素投影パターンの周期 $\bar{\tau}$、画素投影ヒストグラムの全てのピークの間の分離度 η を入力として、ルールベースで決定される（詳細は[4]参照）。周期は、投影パターンのピーク間距離 τ の平均 $\bar{\tau}$ で定義される。また、分離度は隣り合う 2 つの山の頻度分布を確率分布とみなしたときのフィッシャー比から求める。

分離度 η は、領域間の分離のしやすさを表す。また、図１に示したように水平投影パターン上で周期性が見られる場合に、周期 $\bar{\tau}$ は直交する軸上（垂直投影パターン）での分離可能な空白幅の決定に用いられる。これは、領域間の空白は文字列間の周期より大きいという仮定に基づいている。

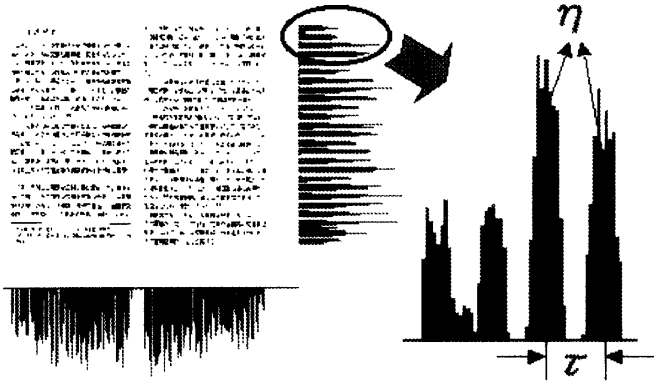


図1 投影パターン上での領域分割

3. 拡張スプリット検出法

本手法は仕切線、画像領域の縁、大きな空白領域等のセパレータ要素を検出し、その配置を利用することで、切り出し境界を非直線的なものに拡張する。

本手法の処理の流れを説明する（図２参照）。

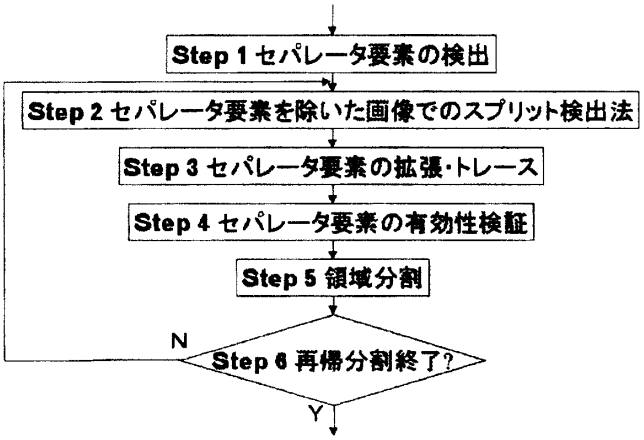


図２ 拡張スプリット検出法

Step 1 (1)仕切線、(2)大きな空白領域、(3)図表・写真領域、(4)分割対象領域の端等をセパレータ要素とする。縮小した画像中の連結成分ごとに、面積、縮小

Document Layout Analysis by Extended Split Detection
Noboru Nakajima, Naoya Tanaka and Keiji Yamada
NEC Corporation, C&C Media Research Labs.
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216, Japan

画像および原画像での黒画素密度や画素反転数等の特徴量を抽出する。これらの特徴量から予め判別分析により生成した判別関数を用いて連結成分の属性(写真、仕切線、文字等)を判定する。また、大きな空白領域の抽出は文献[3]と同様の方法で行った。

Step 2 既に検出されたセパレータ要素を除いた画像から画素投影パターンを生成し、スプリット検出法により分割可能な空白幅を推定し、従来通りの投影パターン上で空白を検出する。得られた空白はセパレータ要素として登録する。

Step 3 登録された各セパレータ要素を外接矩形の長辺方向に延長する。接した連結成分がセパレータ要素である場合には、さらに短辺方向に他の連結成分に接触するまで拡げ、セパレータ要素として登録する。

Step 4 スプリット検出法で求めた空白幅と各セパレータ要素の幅を比較し、その幅が空白幅より小さいセパレータを棄却する。

Step 5 残ったセパレータ要素で囲まれる閉領域をレイアウト構造の部分領域として抽出し、分割前の部分領域の階層下に登録する。

Step 6 登録された領域分割が文字列レベルでない場合には、***Step 2～5***を再帰的に行う。

4. 実験

多段組で図表・写真と文章や文章どうしが入り組んだ新聞画像 15 ページ分を用いて評価実験を行った。全文字列(2155 行)のうち、1966 行(91.2%)の文字列が上位のレイアウト階層構造を含めて正しく抽出できた。過抽出された文字列は 15 行であった。118 の写真・画像領域のうち、110 領域(93.4%)を正しく判定できた。実験に用いた新聞データの解析結果(画像の一部)を図 3 に示す。

5. まとめ

スプリット検出法にセパレータ要素による領域分割を導入し、効果を新聞画像の構造解析実験により確認した。本方式では、レイアウト構造上の階層に即した適切な空白幅が設定され、望ましい構造の階層化が行える。また、セパレータ要素をトレースして分割位置を検出するため、入り組んだレイアウト構造の解析が可能となる。

謝辞 本研究の機会を与えてくださった NEC C&C メディア研究所津雲部長、システム構築にご協力いただいた NEC 情報システムズ田中主任に感謝致します。

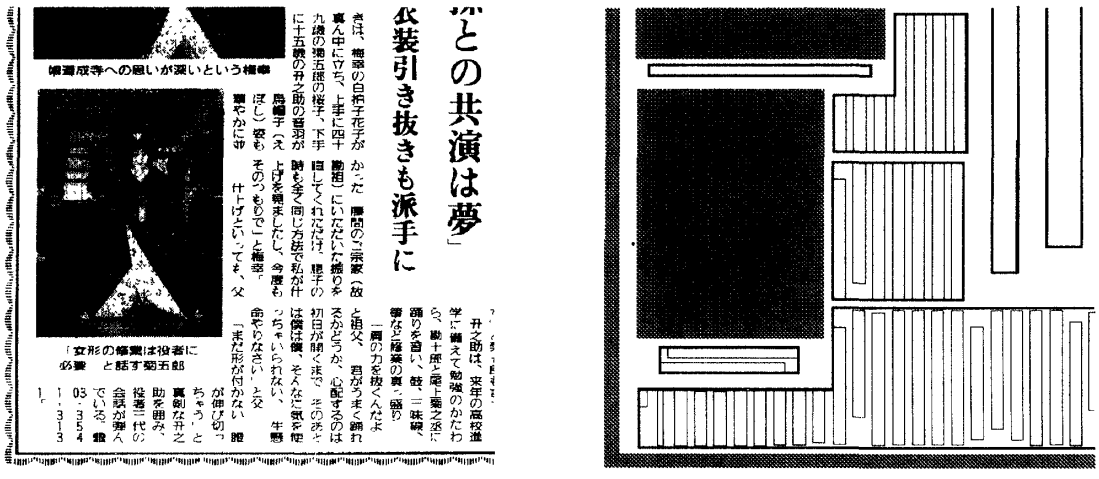
参考文献

[1] Y. Y. Tang, S. W. Lee and C. Y. Suen, "Automatic Document Processing: a Survey", Pattern Recognition, Vol.29, No. 12, pp. 1931-1952, 1996.

[2] A. K. Jain and Y. Zhong, "Page Segmentation Using Texture Analysis", Pattern Recognition, Vol. 29, No. 5, pp. 743-770, 1997.

[3] M. Okamoto and M. Takahashi, "A Hybrid Page Segmentation Method", Proc. 2nd ICDAR, pp. 743-748, 1993.

[4] 辻善丈, "スプリット検出法による文書画像構造解析", 信学論, Vol. J74-D-II, No. 4, pp. 491-499, 1991.



(a)原画像　　(b)解析結果(文字列およびブロックを表示)

図 3 実験結果例