多段組の文書における文字行抽出処理の高精度化

塚田 仁志 * 後藤 英昭 † 阿曽 弘具 *
*東北大学大学院工学研究科 †東北大学工学部

1. はじめに

文書認識システムでは,抽出する文章領域が矩形の場合だけでなく,様々な形状の領域である場合でも取扱いが可能であることが望まれている.任意の形状の領域抽出を可能にするため,文章領域の抽出に先立ち,局所的な処理で文字行を抽出する手法-区分直線連結法[1]が提案されている.

しかしこの手法を多段組の文書構造をもつ文書に適用した場合,図1のように異なる領域に属する文字行同士が結合して抽出されることがある.これを**過接合**と呼ぶ.この過接合は,後に行なわれる領域分割の処理に大きな妨げとなる.本研究では多段組の文書における文字行抽出処理の高精度化を目指し,区分直線連結法手法を基礎にして,この過接合を判断し切断する処理法を提案する.

文書を巨視的に処理しようとすると,文書構造に関する知識が必要となり,その知識により対象文書が限定されてしまう.様々な文書に適用できる汎用性をもつ処理のためには,大局的知識を利用せず局所的な観点で処理を進めることが必要と思われる.本手法はこれを実現するものである.

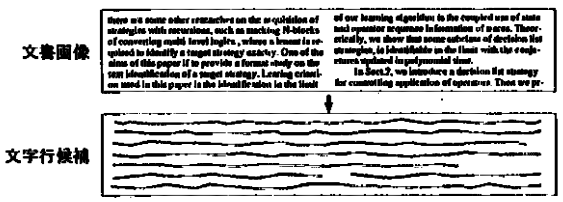

図 1. 過接合の例

2. 過接合切断処理のアルゴリズム

図2に提案する手法である**過接合切断処理**のながれを示す.本手法では一つの文字行に注目したとき,その周辺にある文字行についての情報も考慮することで,過接合を発見する手がかりとする.

まず,抽出された文字行のイメージからその文字行と垂直方向の射影分布を求め,一定以上の長さの空白部分があればその位置を**切断候補点**とする.そして各々の文字行候補について,その**近傍文字行集合**に含まれる文字行の切断候補点を射影する(図3).切断候補点の射影分布を平滑化した後,一定値以上の高さをもつピークが生じればその位置に過接合が生じているものと判断し,その部分で文字行を切断して出力する.

近傍文字行集合は注目する文字行ごとに決定する.θ_i, h_iをそれぞれ文字行iの傾き,高さとし,$d(i,j)$を文字行iと文字行jの芯線間距離として,これらの値を元に文字行iの近傍文字行集合F_iを定義する.Sを抽出された文字行全体の集合とするとF_iは,

$$F_i = \left\{ j \in S \;\middle|\; \begin{array}{lrcl} & d(i,j) & \leq & d_T \cdot h_i \\ and & |\theta_i - \theta_j| & \leq & \theta_T \\ and & \frac{\max(h_i,h_j)}{\min(h_i,h_j)} & \leq & r_T \end{array} \right\}$$

と与られる.しきい値d_T, θ_T, r_Tは予備実験により求める.その際,異なる領域に属する文字行や,異なる性質を持つ文字行などが含まれることを避けるようにする.

3. 実験および結果

一般の雑誌等から任意に選んだ文書画像(A4,200dpi)40ページを用いて本手法の評価実験を行なった.実験結果を表1に示す.

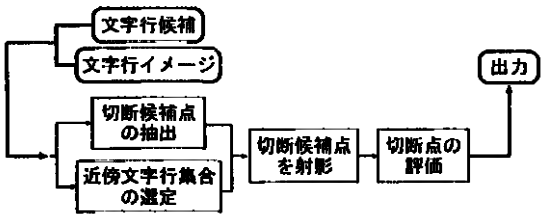

図 2. 過接合切断処理のながれ

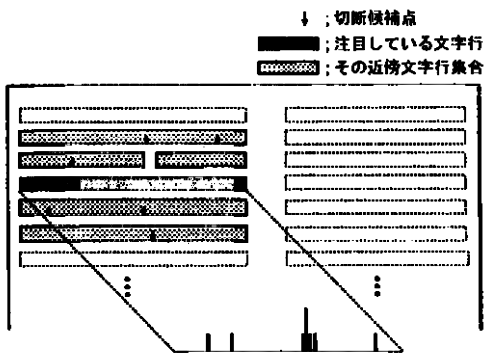

図 3. 切断候補点の射影

過接合を切断することができなかった部分が2箇所あったが,これは注目する文字行の高さを精度良く評価することができず,好ましい近傍文字行集合を形成することができなかったことによるものである.近傍文字行集合の決定法を検討することで,解決の手がかりになると思われる.また,切断する必要のない部分が7箇所誤って切断されたが,これらは日本語文書において句読点が縦方向に重なっているところで生じた.本手法のような局所的な処理では対策が困難と思われるが,このような誤切断は領域分割の処理で救済可能であると考えられる.

表 1. 実験結果

区分直線連結法により生じた過接合	204 箇所
過接合切断処理で正しく切断された箇所	202 箇所
誤って切断された箇所	7 箇所

4. まとめ

既存の文字行抽出処理に新たに提案した過接合切断処理を付加することにより,ほぼ全ての過接合を切断することができた.また,誤って切断された部分も極少数にとどまった.本手法により,多段組の文書構造をもつ文書画像からの文字行抽出の高精度化が達成されたといえる.

今後の課題として,文字行抽出処理のさらなる高精度化と抽出された文字行をもとに文章ブロックを生成,抽出すること等が挙げられる.

参考文献

[1] 後藤英昭,阿曽弘具:"文字行の局所的な直線性を利用した頑健・高速な文字行抽出法",信学論(D-II),**J78-D-II**, 3, pp.465-473 (1995-03).