細線化処理を用いた手書き文書画像からの文字列領域分割法

1D－6

足立雄介　吉川大弘　鶴岡信治
三重大学工学部

1．はじめに

ＯＣＲの開発が進む現在において，未だ手書き文書に対する文字認識では，満足する結果が得られていない。その要因として，個別文字の切り出しや，その前段階である文字列の切り出しが難しい事が挙げられる[1]．本稿では，文書の文字列が白領域で分割されている事に注目し，文書画像の白領域を細線化処理により線図形（芯線）で表し，その芯線から不要な部分を除去することにより，文字列を囲むループを抽出する方法を提案する．

2．文字列領域分割法

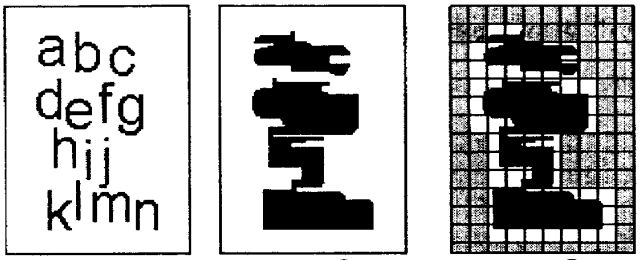
文書画像からの文字列領域分割法は，細線化領域の限定，白領域細線化，不要な芯線の除去から成る．各処理に用いるパラメータは，文書画像の横方向の度数分布をフーリエ変換し，多く含まれる周波数成分の周期で最も小さいものを，文字列の高さh_lと考え，h_lから求めたものを用いている．

2．1　細線化領域の限定

画像全体の白領域を直接細線化しようとすると，処理時間が長くなり，細線化後の芯線も，文字列の領域とはかなり異なる分離線となる．そこで本稿では，明らかに細線化を行う必要のない領域の除去を，次の手順で行う．

①２値画像（図１(a)）に対して，左右方向にh_l画素の膨張収縮処理を行う（図１(b)）．

②膨張収縮処理を行った画像を，大きさ $h_l/4 \times h_l/4$画素の正方領域で分割する．その正方領域の中に黒画素が存在しなければ，その領域は細線化を行わないことにする（図１(c)）．



(a)２値画像　(b)手順①後　(c)手順②後

図１　細線化領域の限定

2．2　白領域の細線化

限定された白領域に対して細線化処理を行う[2]．ここで得られる芯線の画素の連なり$t_1, \cdots, t_n$のうち，t_1，t_nが共に特徴点（端点，分岐点，交差点）であり，他の画素 $t_2, \cdots, t_{n-1}$が全て連結点であるものを“チェイン”と呼ぶ[3]．図２にチェインの例を示す．

文字列の境界を表すループを得るためには，チェインを単位とした取捨選択を考えればよいことになる．

2．3　チェインの評価

横書き文書の場合，文字列を分割するための線は，ほとんどが横線であると考えられる．そこで本稿では，不要なチェインを発見するために，チェインの傾きa及び，高さh_cを特徴量として用いる（図２）．

これらの特徴量により，次の①～⑤の各条件を満たすチェインを段階的に除去していく．すなわち，まず①，②の条件により，傾き高さの大きなチェインを除去し，次に③の条件で，高さの小さなチェインを除去する．最後に④，⑤の条件により，傾きがそれほど大きくないチェインを除去するのである．条件①→⑤の順で行うことで，より不要と思われるようなチェインから順に除去していくことができる．

① $a \geqq 4$，かつ $h_c \geqq h_l/2$

② $a \geqq 2$，かつ $h_c \geqq h_l/2$

③ $a \geqq 2$，かつ $h_c \geqq h_l/8$

Segmentation of a Character String Region Using Thinning Algorithm from Handwritten Documents
Yusuke ADACHI, Tomohiro YOSHIKAWA,
Shinji TSURUOKA
Faculty of Engineering, Mie University

④ $a \geqq 4/3$，かつ $h_c \geqq h_l/4$

⑤ $a \geqq 1$，かつ $h_c \geqq h_l/4$

ただし，各条件によるチェイン除去終了後，芯線の中でループを構成していないチェインは除去する．また，“１番外側に存在する線は枠なので消さない”という条件も加える．

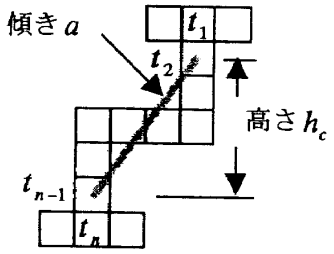


図２ チェインの例

２．４ 過分割領域の結合

以上の処理では，どうしても過分割された領域が残ってしまう．そこで次の処理により，高さの小さい領域（高さが $h_l/2$ 未満の領域）を過分割領域とみなして，その領域を，本来含まれるはずの文字列の領域に結合させる．

過分割領域の左右を別々に見て，他の領域と接している部分の高さを求める．過分割領域は，その高さが最も大きな領域に含まれるものとする．図３では，h_2 が最も大きいので領域②は領域③に含まれることになる．

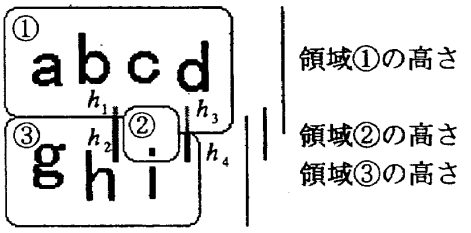
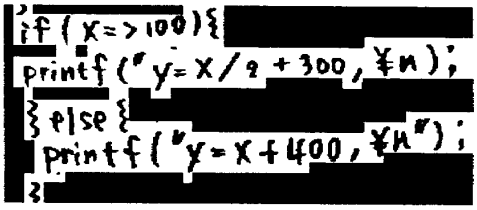


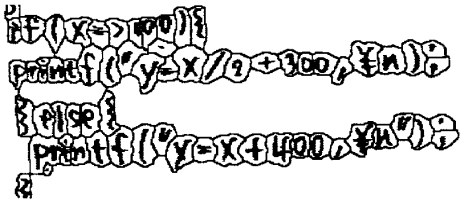
図３ 過分割領域の結合

３．実験結果

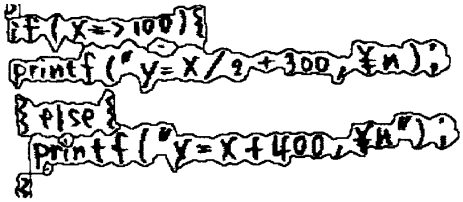
図４(a)，(b)，(c)及び(d)に，各処理後の画像の例を示す．白紙の紙に書かれた筆記者の異なる 11 画像に対し本手法を適用したところ，全 215 行の文字列のうち 191 行（89%）を適切に分割する事ができた．過分割は 13 行（６%），過統合は 15 行（７%）に存在した．過統合は，主に文字列間の接触が原因であった．



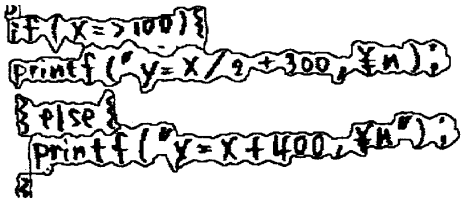
(a) 細線化領域の限定後の画像



(b) 白領域細線化後の画像



(c) チェイン除去後の画像



(d) 文字列領域分割結果

図４ 各処理の結果例

４．おわりに

本稿では，領域を限定した細線化と，特徴量に傾きと高さを用いた，チェインの段階的な除去により，手書きの文書領域から文字列領域を分割する手法を提案した．しかし，文字列間に接触がある場合，本手法ではそれらの文字列を分割することができないため，今後，それらに対する検討が必要である．

参考文献

[1]上杉，鶴岡，三宅，石田，“手書き文字認識のための文字列切り出し方法”，電子情報通信学会パターン認識・メディア理解研究会，PRMU96-152，pp.85-90，Jun.1997.

[2]鶴岡，木村，吉村，横井，三宅，“ディジタル図形の一細線化法と手書き文字認識システムへの応用”，電子通信学会論文誌(D)，Vol.J66-D,No.5，pp.525-532，1983.5.

[3]黄瀬浩一，柳田修，“白領域の細線化を用いた文書画像の領域分割”，電子情報通信学会論文誌(D-Ⅱ)，Vol.J80-D-Ⅱ,No.6，pp1608-1616，1997.6.