汎用を目指した文書画像認識システム
－領域分割処理の改良－

6R－2

牧野　隆雄　米田　政明　長谷　博行　酒井　充　丸山　博
富山大学工学部電子情報工学科

1 はじめに

近年、活字の文字認識に関する研究の進展と計算機の性能の大幅な向上に伴い、様々な文書を計算機に読み取らせる文書画像認識システムの研究が盛んに行われるようになった。このようなシステムにおいては文字認識の前段階として、文書画像を図領域、文字領域、フィールドセパレータなどの領域に分割する処理が不可欠である。また、汎用性のあるシステムの作成を目指す上で、利用対象者を健常者だけに限らず、また、その処理対象についてもできるだけ制限しないものにする必要がある。本稿ではこのような文書画像システムの構築と、その処理対象をより広げるための改良について述べる。

2 汎用システムの機能

2.1 ページ分割処理

書籍を読み込む場合、その入力形態としては次の３つが考えられる。

1.書籍の片ページを読み込む。
2.書籍の両ページを読み込む。
3.一枚の紙を読み込む。

この３種類の形態の判定には、書籍を読み取ったときに見開きの境界の部分がスキャナーの走査面から浮くことにより生じるくぼみを利用する。このくぼみの有無及びその位置によっていずれかを判断し、それに基づいてページ領域の切り出しを行う。

2.2 傾き補正

文書画像中の文字が書かれた部分で黒画素の射影パターンをとると、そのパターンに起伏が生じ、射影パターンをとる際の走査方向が文字列の方向に一致したとき、その起伏が最大になる。この性質を利用して傾きの角度を求め、傾きを補正した画像をつくる。

Document Recognition System for General Use
－Improvement of Segmentation－
Takao Makino , Masaaki Yoneda ,
Hiroyuki Hase , Mitsuru Sakai ,
Hiroshi Maruyama
Faculty of Engineering , Toyama University
3190 Gofuku , Toyama 930 , Japan

3 領域分割処理

3.1 領域分割の手法

3.1.1 外接矩形の生成

黒画素の８連結追跡処理により図形の輪郭を求め、その外接矩形を生成する。領域分割処理は、この外接矩形を処理単位として行う。

3.1.2 平均文字サイズの推定

外接矩形の幅、高さのヒストグラムを取り、二つを合わせたヒストグラムの最頻値を平均文字サイズとする。

3.1.3 点線の抽出

外接矩形を生成した後、同じような矩形が直線上に並んでいる部分において、矩形を統合し点線を抽出する。

3.1.4 属性付け

図、フィールドセパレータは、その外接矩形の幅、高さによりその属性を判断する。

文字属性を持つ矩形に対しては、文字方向属性を付ける。

3.1.5 文字領域の分割

文字領域は、縦、横方向に射影パターンをとり、最も適切な空白領域で二分する。以降、これを再帰的に繰り返すことにより、行単位まで分割する。

3.2 処理上の問題点

この領域分割処理は多くの文書形式に対応しており、文庫本、実用書、教科書類などをはじめ、全般的に非常に高精度な領域分割が可能である。しかしながら、新聞や雑誌類においてはレイアウトや構成要素が多種多様であるため、分割処理が行えない場合が存在する。

図１にその一例を示す。このように写真や図が重なり合っていたり、斜めにレイアウトされているなどして、非矩形の黒画素連結領域が形成される場合、外接矩形処理を行うとその周辺の文字領域まで取り込み、貴重な文字領域を失う原因となる。

こうした場合にも対応するため、以下のようなアルゴリズムを導入した。

図１　レイアウトの複雑な画像

3.3 図領域に含まれた文字の抽出

3.3.1 図領域の選出

3.1.4で述べた判定方法により、図領域と判定された矩形データを選出する。ここで図領域と判定されるものには、文字領域を含む図領域なども含まれている。

3.3.2 メッシュ画像による処理

メッシュ画像の作成

色の薄い写真や線画等を完全な黒画素連結領域にするため、図領域の図形に粗いメッシュをかけた画像を作成する。

メッシュ画像中の８連結追跡処理

このメッシュ画像中の８連結画素の輪郭を辿る。そして、図２のようにその輪郭線がＡ、Ｂ、Ｃ、Ｄ全ての角を通過すれば、図、写真だけの領域か、囲み枠による領域と判断し、矩形データに変更を加えない。逆に、全てを通過しないのであれば、領域内に文字が取り込まれてしまっている可能性がある。この矩形領域については以下の処理を行う。

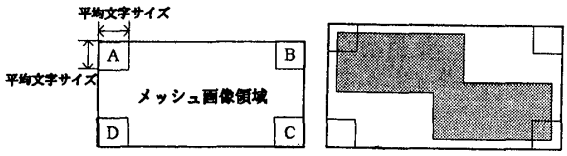

図２　領域判別

3.3.3 図抽出画像の作成

メッシュ画像中の８連結処理で辿った輪郭の中で3.1.4で図属性の判定条件を満たすものは、メッシュ画像と同様のサイズの白い画像に描き写し、その輪郭内を塗りつぶして、図抽出画像を作成する。

3.3.4 領域中の外接矩形データ抽出

得られた図抽出画像により、写真、図、線などから形成される非矩形の黒画素連結領域を知ることができ、残りの領域中から文字の外接矩形データの抽出を行うことが可能になる。

4 実験

改良手法の有効性を確認するため処理実験を行った。サンプルには図の入り組み等を含む３５種類の画像を用いた。どの画像についても目的とする文字領域の抽出が行われ、全体的に良好な結果が得られた。その処理例を図３に示す。

非矩形領域を形成する図形の中で、線画によるものの場合、図を構成する細かい線がメッシュによる処理で取りきれずに残る場合があった。

図３　処理例

5 まとめ

本稿では、非矩形図形により失われる文字領域の抽出を取り上げた。現在、本研究では空白領域を利用して、領域を再帰的に二分割していき、行を求める手法を用いているが、雑誌や新聞などの文書画像では行方向等の複雑なレイアウトにより、現在の二分割処理では対処できない画像が存在することも分かっている。汎用性を目指す上で、こうした点についても引き続き研究が必要であると考えている。

6 参考文献

辻，米田，長谷，酒井:"汎用を目指した文書画像認識システム"，電気関係学会北陸支部連合大会講演論文集，F-41，pp. 386(1994)

岡本，宮澤:"多様なレイアウト、構成要素を持つ文書画像の領域分割"，信学技報，ＰＲＵ89-114，pp. 9-16(1990)