

9 K-9

文章領域抽出による新聞記事の 自動区分システム

加藤誠巳 渡辺 熊
(上智大学理工学部)

1 まえがき

新聞記事、特に電算化される以前の新聞記事の紙面情報をデータベースに落とす際、記事の紙面分割（見出し領域、文章領域、写真領域、図領域、空間領域、黒の縦ライン領域、広告領域、等）を自動的に行い、1つ1つの記事ごとに領域をまとめ、紙面の記事をそれぞれの内容ごとに区分し、また文字認識によりキーワード検出をすることによって記事内容を分類することは、データを管理保存する際に極めて有用であると考えられる[1]-[4]。本稿では、紙面をイメージスキャナで読み取った結果の白黒2値情報の縦または横方向の黒ドット数のヒストグラム情報を用いて新聞紙面を記事内容ごとに自動区分する手法について述べている。

2 記事自動区分の手順

白黒256階調の解像度68.2(dot/inch)で新聞記事の紙面データをイメージスキャナで読み取ったデータをファイルとして保存する。この256階調データに対し、異なる閾値処理をすることにより2種類の2階調データを作り、その2つのデータを面積投影し、更なる閾値処理をすることによって領域を抽出していく。領域抽出の後、領域に適当な番号を付けを行うことにより新聞の記事内容ごとに区分していく。尚、現在は新聞1ページの1/4の領域を対象としている。

2.1 白黒256階調画像データの2値化

256階調画像データを低閾値($\theta_1 = 185$)によって2値化したデータと、高閾値($\theta_2 = 248$)によって2値化したデータをつくる。 θ_1 、 θ_2 の値は現在上記の様に定数としているが、新聞記事の状態により0を変えなければならない場合がある。

2.2 低閾値画像データによる領域抽出

低閾値画像データを縦および横方向に面積投影して閾値処理することによって、見出し領域、写真領域の濃度値が高い部分、広告領域、図領域と推定される領域を抽出する。

これらの領域抽出を先に行うことにより、次に実行される高閾値画像データによる領域抽出をより正確なものにすることが可能となる。（図1参照）

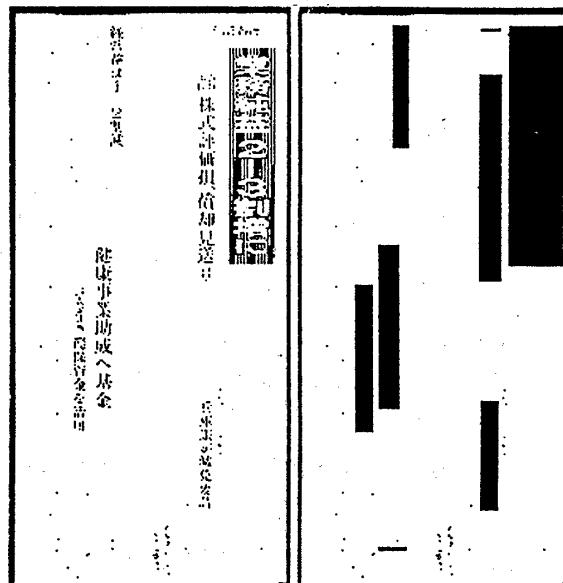


図1 低閾値画像データと抽出された領域

2.3 高閾値画像データによる領域抽出

低閾値画像データによって抽出された領域の部分を除いた高閾値画像データを、縦方向に面積投影して閾値処理することによって、写真領域、文章領域、空間領域、縦黒ライン領域と推定される領域を順次に抽出する。（図2参照）



図2 高閾値画像データと抽出された領域

2.4 最終紙面分割

誤って抽出されることの多い下記のような領域に対し、適当な処理を順番に施す。また見出し領域の周囲に存在する空間領域を見出し領域に接続し、その領域を最終的な見出し領域とする。

1. 抽出された見出し領域の中から広告領域、特に文章領域と同一の縦幅を持つ領域面積の小さい広告領域と思われる領域を抽出する。
2. 写真領域、図領域に重なっている見出し領域を削除する。
3. 見出し領域と重なっている空間領域を見出し領域に接続する。また同様に、見出し領域と重なっている文章領域を見出し領域に接続する。
4. 見出し領域の両隣に存在する空間領域を見出し領域に接続する。
5. 空間領域を、文章領域や写真領域に接続可能な場合は接続する。

2.5 記事別区分

見出し領域、文章領域、写真領域、図領域を番号付けして記事内容ごとに区分していく。このとき、抽出された黒の縦ライン領域のデータを使うことによって、より正確な記事区分が実行できる。(図3 参照)

1. 同じ記事内容の見出し領域と思われる領域に同一番号をつける。

2. 見出し領域の左隣の文章領域、写真領域、図領域は、その見出し領域番号と同一の番号をつける。

3. 縦の黒ライン領域と既に番号のついている領域のデータを利用して、この時点で番号のついていない文章領域、写真領域、図領域に番号をつける。

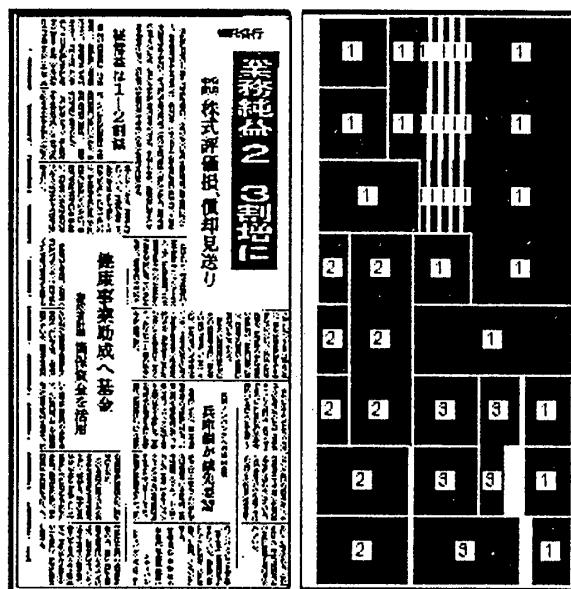


図3 記事区分後の画面

3 むすび

文章領域抽出による新聞記事の自動区分システムについて述べた。最後に、有益な御討論をいただいた本学マルチメディアラボの諸氏に謝意を表する。

参考文献

- [1] 朴, 海老名, 伊藤: “汎用的な文書画像の階層的領域分割と識別法”, 電子情報通信学会技術研究報告, PRU 91-46(1991).
- [2] 野口, 豊田: “新聞記事の切抜きを行うシステムに関する基礎的研究”, 情報処理学会第23回全国大会講演論文集, 6C-1(1981).
- [3] 坂井, 稲垣, 加藤: “複雑な構造を持つ文書画像の自動解析”, 情報処理学会第23回全国大会講演論文集, 6C-2(1981).
- [4] 牧野, 赤田: “文書画像の領域分割について”, 情報処理学会第23回全国大会講演論文集, 6C-3(1981).