新聞構成ブロックの読み順に対する一考察

1D－8

浦田敏道†‡　海老名毅‡　猪木誠二‡　井上彰†
†(株)エム研　‡郵政省通信総合研究所

1. はじめに

イメージスキャナで取り込んだ新聞画像をコンピュータで解析し、音声出力により内容を伝達するシステムにおいて、領域分割処理([1])により抽出された各矩形領域(ブロック領域)に対し読み順を割り当てる処理は、重要な役割を果たす。この処理が正確に行われなければ、話の繋がりの悪い領域を続けて読み上げてしまい、聞き手(ユーザー)に、理解しづらい内容を伝達してしまう。

我々が今回提案するのは、日本語新聞に限る読み順割り当て手法である。数社の新聞について調査したところ、「ほとんどの新聞レイアウトにおいて、各ブロック領域の読み順はセパレータ（罫線）の配置によって決定できる。」という結論を得た。そこで本手法では新聞中のセパレータの配置に着目して、ブロック領域の読み順を決定している。以下の章で本手法の詳細を述べる。

2. ブロック領域に対する読み順割り当て手法

2.1 用語の定義

本手法について述べる前に用語の定義をする。

- 水平距離

二つの任意のブロック領域 X と Y を、$X=(xl,xt,xr,xb)$　$Y=(yl,yt,yr,yb)$ とする時、ブロック領域 X , Y の水平距離を

$d\_horz(X,Y)=\min(|xl-yr|,|xr-yl|)$　と定義する。

但し原点は対象画像の左上隅で右方向を水平座標正の方向、下方向を垂直座標負の方向とする。

- 垂直距離

二つの任意のブロック領域 X と Y の垂直距離を

$d\_vert(X,Y)=\min(|xb-yt|,|xt-yb|)$、と定義する。

- 左側に存在、下側に存在

$xl \geq yr$ が成り立つ時、領域 Y は領域 X の左側に存在するといい、$xb \leq yt$ が成り立つ時、領域 Y は領域 X の下側に存在するという。

- 可読領域

領域分割識別処理により新聞画像中から抽出された領域のうち、読み上げ対象となる領域を可読領域と呼ぶ。我々が開発したシステムでは、表、写真、図、文字領域を可読領域としている。

2.2 読み順割り当て処理

以下に、提案する読み順割り当て処理手法について説明する。初期設定として、自然数 n に初期値 1 を代入しておき、処理は基本的に下記の処理１、処理２、処理３、....、処理 7 の順に実行するものとする。途中指示があれば指示された処理に移行するものとする。

処理１

対象画像の右上隅点からもっとも近い未読ブロック領域を抽出し、それを領域 A とし、領域 A に読み順 n を割り当て n に 1 をたす。未読ブロック領域が抽出されなければ、対象新聞画像中の全ての可読領域に対し読み順の割り当てが終了したとし、読み順割り当て処理を終了する。

処理２

A と、A の左側に存在する A との水平距離が最短のセパレータを抽出しそれを s0 とする。s0 が存在しなければ、対象画像領域の左端境界線を s0 とする。

処理３

A と s0 の間に存在する、未読な可読領域のうち最上部に存在し、A と s0 の水平距離が最小のブロック領域を次に読むべき可読領域として抽出しそれを

"A proposal of the method for assigning each rectangle range constructing the newspaper the order for reading", Toshimichi Urata†, Tsuyoshi Ebina‡, Seiji Igi‡, Akira Inoue†

Aとし、読み順nを割り当てnに1をたす。

Aとs0の間に可読領域が存在すれば処理2へ、そうでなければ処理4へ移行する。

処理4

Aの下側に存在し、Aとの垂直距離が最小となり、かつ右辺がAの左辺より右に存在するセパレータ全体を集合S_Aとする。

処理5

Aの下側に存在し、Aとの垂直距離が最小となり、かつ右辺がAの左辺より右に存在する未読な可読領域全体を集合C_Aとする。この時、C_Aが空集合であれば、処理1へ移行する。

処理6

S_Aの要素のうち、Aとの水平距離が最も短いセパレータsを抽出。S_Aが空集合である時、対象画像領域の右端境界線をセパレータsとして抽出する。

処理7

C_Aの中からsの左側にある、sとの水平距離がもっとも短い可読領域を選び、それをcとする。cが存在しなければ、処理1へ、存在するのであればcをAとし、読み順nを割り当てnに1を足し、処理2へ移行する。

3. 結果と考察

A4サイズの新聞を400dpiの解像度でスキャナから取り込み、その画像に対し領域分割識別処理を施し、各ブロック領域に読み順割り当て処理を実行した結果が図1である。図1の新聞画像中の枠で囲まれた部分が領域分割識別処理により抽出されたブロック領域で、各領域の中央部に読み順割り当て処理により割り当てられた読み順を表示している。この新聞画像に対する読み順割り当てはほとんど正確に割り当てられているが、読み順「2」のブロックから読み順「13」のブロックの順に本来ならば読み順を割り当てるべきところを誤って割り振られている。このような誤りは縦書き見出し文字領域の上に小見出し的な横書き見出し文字領域がレイアウトされている場合によく生じる誤りパターンである。また、処理7で選出された可読領域cの選び方において、大体の新聞レイアウトはこれに従うが、cが見出し文字領域である場合には、本来次に読むべき可読領域はそれより右に存在する場合と、また、次に読むべき可読領域として新しい記事の見出し文字領域cを読み上げてもかまわない場合と二つの場合があり、これについては、文章の繋がりを判断する処理が必要でレイアウト情報のみで正確に読み順を割り当てることは不可能である。我々が開発したシステムでは、処理7で選出された領域cが誤りであれば、セパレータ領域sをS_Aから除外し、処理6へ移行し、正しい領域cが抽出されるまでこれを繰り返すといった、処理が簡単なキー操作でできるユーザーインターフェースを装備することによってこの問題を回避している。

図1 処理後のブロックの順序

4. まとめ

新聞構成ブロックの順序を決定するアルゴリズムを検討した。本手法では、領域分割識別処理終了時において罫線領域が正確に抽出されていることが前提である。今後の課題としては、罫線抽出処理の正確さが要求される。

5 参考文献

[1] 浦田敏道 海老名毅 猪木誠二 井上彰：“空白領域に着目した文書画像の分割法とその識別法”,デジタルドキュメント 8-1 (1997.7.18) pp.1-pp.8