マルチメディア処理における文書構成要素抽出処理

5Y-3

樋野 匡利、　福田 浩至、　田畑 邦晃
(株) 日立製作所　システム開発研究所

1.まえがき

現在、文書処理、文書管理システムを構成する基本的な要素として、ワードプロセッサ(WP)、文書画像ファイル等が用いられている。WPはコード情報を基本処理単位とし、文書の作成、編集が容易にできる点に、文書画像ファイルは画像情報を基本処理単位とし、入出力の容易さ、データベースとの結合による大量文書管理機能に特徴がある。

今後、これらの機能をマルチメディアの概念に基づき結合した文書処理システムが重要になると考えられる。

この様な文書処理システムの実現を目的に、画像処理からのアプローチとして、文書画像から文書構成要素を分離抽出する処理について検討した。

2．処理の概要

文書構造の分離抽出は、次の手順で行なう。

(1) メディアの分離

文書画像から文字、図形、写真・画像の領域を分離抽出する。この処理については、著者らを含め多くの方式が提案されている。[1]-[5]

(2) 文書構成要素の抽出

メディア分離の結果に対して、文書構成要素の性質に着目して、抽出処理を行なう。

本稿では、メディア分離処理のうち連続行、段落の抽出処理と文書構成要素の抽出処理、その応用としての文書の再構成処理について報告する。

3．連続行、段落の抽出

メディア分離処理で抽出された行（以下、縦書、横書にかかわらず、文字列、文字行を単に行と呼ぶ）について、一連の文章を構成すると考えられる行を連続行として抽出する。抽出は離散的弛緩法を用い、同様の高さを持ち、等しい行ピッチで並ぶ行を連続行とすることにより行なう。

以下、説明を簡単にするために、横書の場合について述べるが、縦書の場合も同様である。

各行に対して、上側で隣接する行、下側で隣接する行と、連続行として統合可能かどうかを示す2つのラベルLup、Ldwnを設ける。ラベルは統合可能な場合 '1 (○)'、不可能な場合 '-1 (×)'、決定できない場合 '0 (△)' とする。

具体的には、ある行iについて、まず上側で隣接する行jを求め、行iとjの高さ比rh_1を求め、その値がある範囲内（$1/\delta \leqq rh_1 \leqq \delta$、$\delta \geqq 1$）、即ち行の高さが類似していれば、2つの行の下辺間の距離Dupを求める。高さ比が範囲外、及び、行が存在しない場合には、Dupを無限大とする。下側についても、同様にして、Ddwmを求め、Dup、Ddwmの比の値により、図1に示す様に初期ラベルを設定する。

次に、隣接する2つの行のラベルの組み合わせにより、ラベルの値を更新する。ラベルの更新結果、互いに統合可能な行を、連続行として抽出する。
（図2参照）

抽出された連続行に対して、各行の開始位置、終了位置に着目して段落を抽出する。[5]

図3に段落の抽出結果を示す。

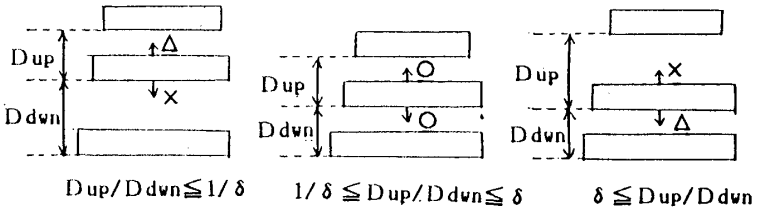

図1 行統合の初期ラベル

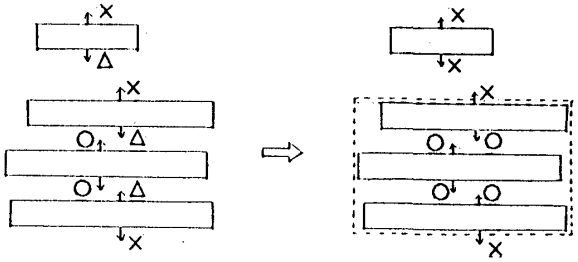

図2 ラベルの更新と連続行の抽出結果

Document Structure Extraction Method by Multi Media Technology
Masatoshi Hino, Kouji Fukuda, Kuniaki Tabata
HITACHI,Ltd.

4．文書構成要素の抽出

処理の対象文書を論文、研究報告等の科学技術文献とし、文書構成要素を次の様に仮定する。文書は、文字部と非文字部から成り、文字部は文献名等の書誌事項、要旨、本文、章節の見出し、図表のキャプション、参考文献リスト等から成るものとする。さらに本文は、章節、段落、行という階層構造を持つものとする。

第３章までの処理で、非文字部の図表・写真と文字部の行、段落の情報が抽出されている。これらの情報に基づき、行の特徴、要素間の位置関係といった文書構成要素、各々の性質に着目して、行や段落が、文書構成要素の何であるかの同定を行なうことにより、文書構成要素の抽出を行なう。

本稿では、本文、章節の見出し、図表のキャプションの抽出処理の例を示す。図４に、抽出処理の結果を示す。

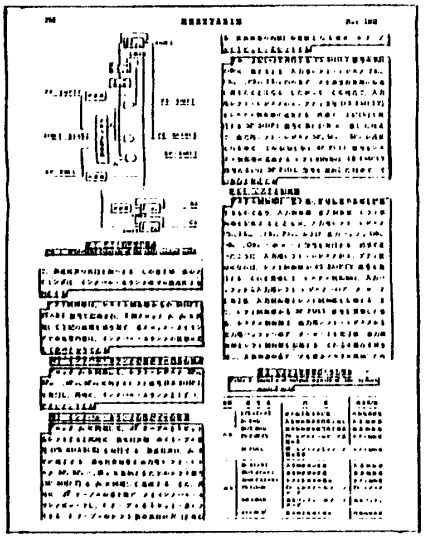

図３ 段落抽出の結果

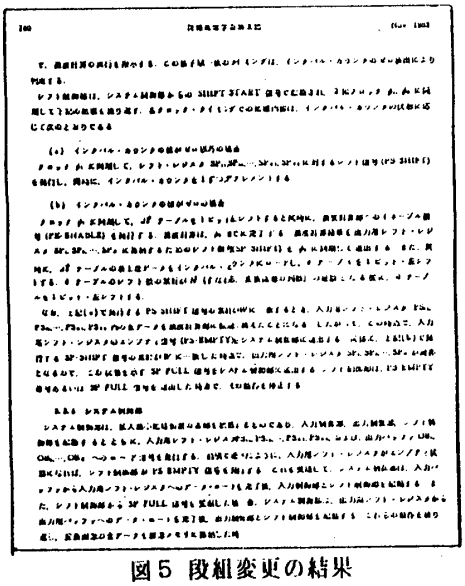

図５ 段組変更の結果

5．文書再構成への応用

文書構成要素抽出処理の応用として、文書再構成処理の例を示す。

図５に文書の段組変更を行なった結果を、図６に図表・写真の抄録を自動作成した結果を示す。

6．結び

文書画像を、文字、図表、写真等にメディア分離し、文書構成要素の性質に着目して、分離結果から各構成要素を抽出する方式を提案、その応用例を示した。

参考文献

[1] 村尾、坂井：情処第21回全大、7H-1
[2] 秋山、増田：信学論Vol.66-D No.1
[3] 岩城他　　：信学研資PRL83-63
[4] 野口、豊田：情処第23回全大6C-1
[5] 樋野他　　：情処第32回全大、3K-2

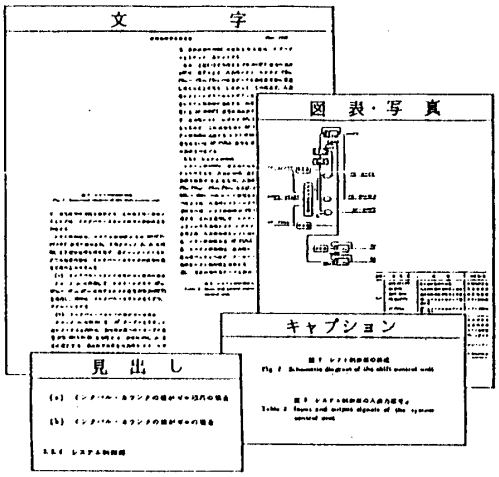

図４ 文書構成要素の抽出結果

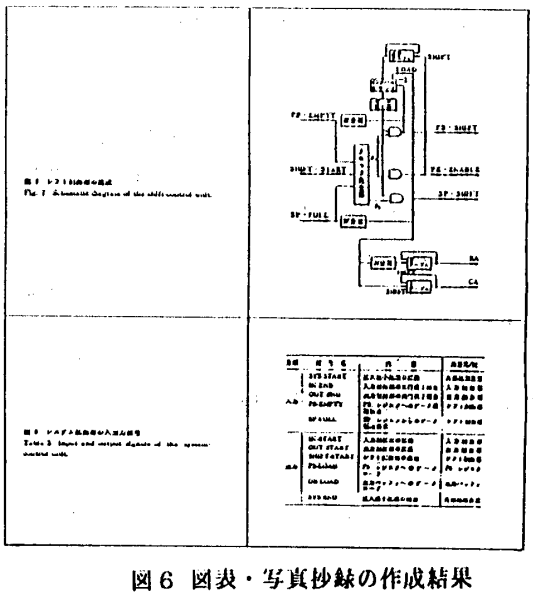

図６ 図表・写真抄録の作成結果