6W-7

英文書認識システム
—文書構造認識—

滝本　照雄　　門前　弘邦　　井上　健一郎
㈱富士通大分ソフトウェアラボラトリ　　富士通株式会社　　㈱富士通大分ソフトウェアラボラトリ

1．はじめに

オフィスオートメーション（ＯＡ）分野における文書処理の占める部分は大きく，ハードウェアなどの進歩に伴い，マルチメディアとしての文書に対して高度な処理が要求されるようになった。

一般の文書は，テキスト，図表，写真などから構成されており，この構成要素ごとに特有の処理が存在する。　文書処理を効率良く行うには，事前に対象文書の物理的構造，論理的構造を認識する必要がある。　本稿では，文書の構成要素の二次元的関係から論理的構造を認識するシステムの試作を行い，英語論文（ＩＥＥＥ）に適用した結果について報告する。

2．システム概要

英文書の構造認識は，英文書認識システムのフロントエンドとして位置づけられる。　英文書の構造認識は，スキャナ入力された文書画像から抽象データを生成するデータ構造解析部と英文書構造に関する知識を用いて抽象データから文書構造（図表，テキスト，写真の分類，テキストの階層化など）の認識を行う推論部から構成される。（図１）

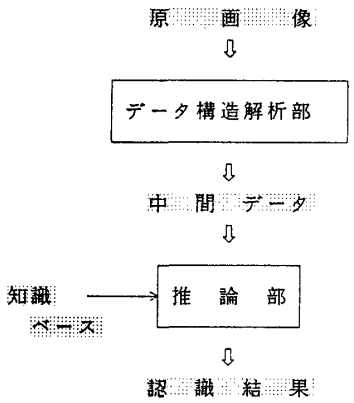

図１．システム概要図

3．文書モデル

今回の構造認識では，データを次のような４階層で考え，それぞれ利用する知識を区別した。

論　理　構　造　：IEEE固有の知識
⇧
物　理　構　造
⇧
中　間　データ　｝英文書共通の知識
⇧
原　画　像

物理構造のモデル記述は図２の通りである。

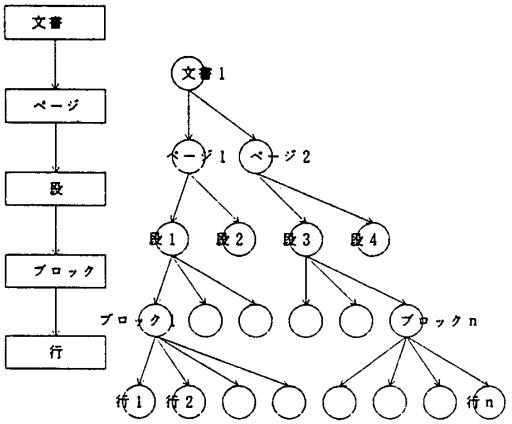

図２．物理構造モデル

論理構造のモデル記述は図３の通りである。

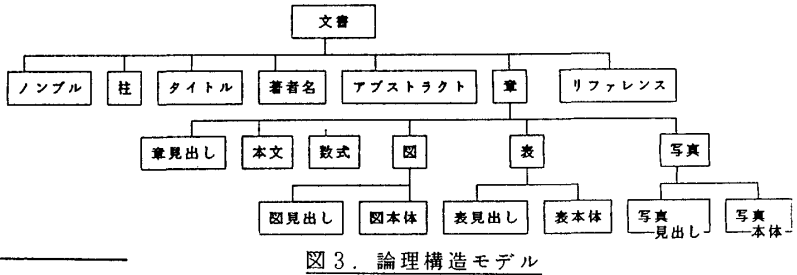

図３．論理構造モデル

A recognition system for printed documents -Recognizing structure of document-
Teruo TAKIMOTO[1], Hirokuni MONZEN[2], Kenichiro INOUE[1]
[1]FUJITSU OITA SOFTWARE LABORATORIES LTD.　[2]FUJITSU LTD.

4．処理概要

4.1　三要素分類

画像のミクロな特徴（黒画素連結領域サイズ，黒画素比など）をもとにマクロな意味のあるまとまり（文字⇨単語⇨行，写真，図形など）の三要素に構造化する。

4.2　物理構造認識

行の位置などをもとに領域の包含・近隣関係を計算し，ページ，ブロックなどの文書に従属した構造を認識（仮説を生成）する。

4.3　妥当性チェック

「ブロックとブロックは互いに重なりは発生しない」，「図領域ブロックの近隣の文字列は図見出しである」などの物理構造についての知識を用いて物理構造の妥当性の検証を行う。

4.4　バックトラック

「4.3　妥当性チェック」で仮説に矛盾が生じた場合，物理構造認識に戻り処理パラメタを変更して再認識を行う。

4.5　論理構造認識

物理構造認識の結果をもとに，節，章などの階層化，接続やタイトル，著者名などメディアに独立な論理構造を対象文書固有の知識を用いて認識する。物理構造を記述するフレームを論理構造を記述するフレームのリンクに書き変えることになる。

4.6　妥当性チェック

論理構造についての知識を用いて論理構造の妥当性の検証を行う。

4.7　バックトラック

「4.6　妥当性チェック」で仮説に矛盾が生じた場合，論理構造認識に戻り処理パラメタを変更して再認識を行う。

これらの処理をまとめたのが図4である。

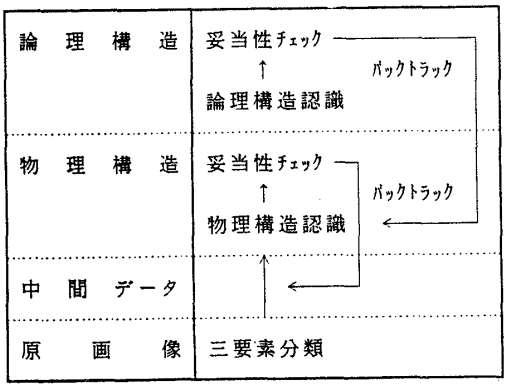

図4．処理概要

5．評価

今回のプロトタイプシステムは，パソコンFMR-60上に，C言語とＥＳＨＥＬＬ／ＦＭを用いてインプリメントされており，ルール数は４０である。

次の条件でテストした結果を表1に示す。

・対象文書　：IEEEの論文（A4版８ページ）
・原画分解能：120 dpi

認識項目	項目数	認識数	誤認識
ﾌﾛﾝﾄﾍﾟｰｼﾞ 要素	7	7	0
章見出し	9	11	4
図	5	5	0
図見出し	5	5	0
表	6	6	0
表見出し	6	6	0
表構造	6	3	3
写真	1	1	0

表1．認識結果

処理時間は約８分で，認識率９３％であった。表見出しの誤認識は論文中に記述された数式を見出しと誤ったためである。表構造の誤認識は罫線の組み合わせの誤認識によるものである。

6．おわりに

今回のプロトタイプシステムでは，文書のレイアウトをもとに，構成要素の二次元的関係（近隣・包含など）を利用し構造認識を行った。　今後は他の文書をテストし，文書の書式のルールを増やし，整理して認識率の向上を図りたいと考えている。

参考文献

(1)門前他：二次元的関係を用いた白地図認識エキスパートシステム，　昭和62年度人工知能学会 pp411 ～414
(2)辻本，麻田：文書画像理解による記事の自動抽出 ｺﾝﾋﾟｭｰﾀﾋﾞｼﾞｮﾝ研究会資料，52-13