6W-6

文書理解における論理構造抽出の一手法

屋代 寛　村上達也　東野純一　嶋 好博　中野康明　藤澤浩道
(株)日立製作所 中央研究所

1．はじめに

一般に，文書には章・節等の意味的な構造を表わす論理構造，および，ページ内の印刷上の配置を表わす割付け構造がある。文書の内容を読者に正確かつ迅速に読者に伝えるため，これらの構造が有効に用いられる。特に，情報価値が高いと考えられる文書は，これらの構造を明確にしてある場合が多い。たとえば，学術論文や特許明細書などが，これに相当する。

ハイパーテキストの出現などを見てもわかるように，計算機における文書処理においても，構造を用いた目次の自動生成，インデックス情報の付加など高度な文書の利用が要求されている。このため，既存の紙面イメージの文書についても，構造を付加することが要求されている。

これまで，文書画像の書式(割付け構造)に着目し，内容を解析する方式を報告してきた。今回，文書画像から文書の論理構造を抽出する方式を検討し，基礎的な実験によって有効性を確認したので報告する。

2．文書画像理解システム

本システムのシステム構成を図１に示す。本システムは，文書画像を入力し構造化する解析部，適切な出力形式に変換する生成部，そして構造記述部から成る。解析部の処理・認識モジュールは構造記述部によって制御される。文書の明確な書式は書式定義言語ＦＤＬ（Form Definition Language）[1,2]で記述し，比較的あいまいな規則についてはルール形式で記述する。

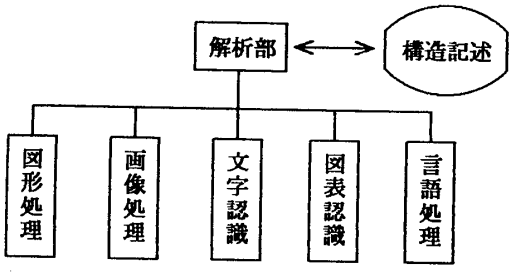


図１．文書画像理解システムの構成

3．論理構造抽出

ここで対象とする文書は，割付け構造，論理構造ともに比較的に明確である学術論文や特許明細書などを考える。これらの文書は情報の価値が高く，文書理解を行なう価値も高いと考えたためである。

（１）幾何学的情報を用いた抽出

文書の割付け構造は読者に内容を効果的に伝えるために，論理構造を反映している場合が多い。たとえば，章題は本文領域よりも行間が広い，また，段落の最初の行は字下げされている。このような割付け構造に関する幾何学的情報を用いて，章・節の抽出を行なう。

（２）文字認識と協調させた抽出

書式が明確である文書についても，上記（１）の方式では論理構造の抽出を行なうのは限界がある。幾何学的な情報を用いても論理構造の抽出ができない場合には，文字フォント情報，さらには，文脈情報の解析が必要である。文字認識と協調させた抽出の第１ステップとして，文字フォント情報を用いた解析を行なう。

上記（１），（２）で述べた処理は，まず，カラム単位の外接矩形領域に分離し，次いで行単位の外接矩形領域に分離する。この行単位の外接矩形領域を基本として，行の幾何学的な情報（空間的な配置など）の解析処理や文字認識処理を行なう。

4．実験

論理構造の抽出を計算機上で行なう。実験に用いた文書は雑誌「Hitachi Review」（1986年4月号vol.35, NO.2)を対象とした。図２（ａ）は実験に用いた文書の一例である。図２（ｂ）は章題・節題の抽出を行なった結果，図２（ｃ）は章“CONCLUSIONS”に対してパラグラフ単位に構造化した結果である。

図３は文字認識と協調させた章・節構造の抽出を行なった結果である。実験に用いた文書（図３（ａ）参照）には節題“Workstation hardware”があるが，幾何学的な情報だけでは節題として抽出できなかった。

An Extraction Method of Logical Strucure for Document Image Understanding
Hiroshi YASHIRO, Tatsuya MURAKAMI, Jun'ichi HIGASHINO, Yoshihiro SHIMA, Yasuaki NAKANO, Hiromichi FUJISAWA
Central Research Laboratory, Hitachi, Ltd.

図３（ｂ）では，文字認識を行ない，その際に得られたフォントの情報も抽出している。実験に用いた文書の章題・節題には，本文で用いているフォントと違うものを用いており，この情報を用いることによって章題・節題の抽出が可能になった。

５．むすび

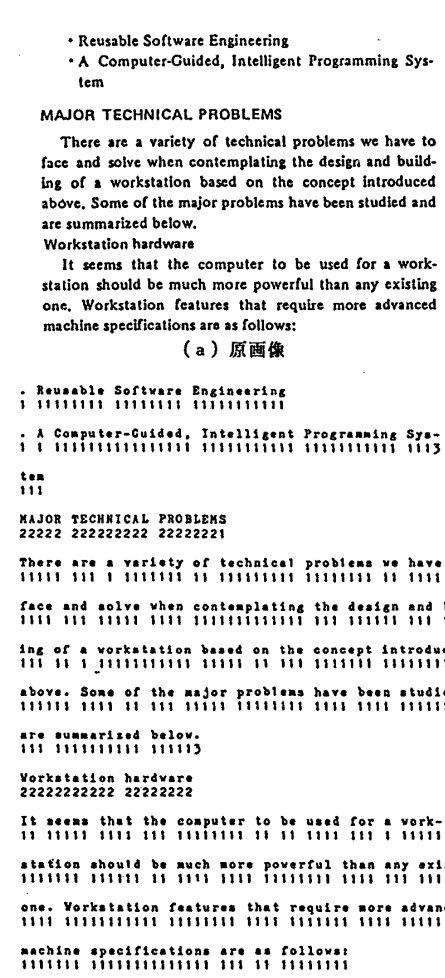
文書画像から文書の論理構造を抽出する方式を提案した。この方式は，文書画像の幾何学的な位相関係を用いた解析だけでなく，文字認識との協調的な処理となっており，より強力で柔軟な構造解析の手法である。今後は，文書理解システムの構築のため，画像処理，文字認識，言語処理など多面的な処理を統一的に行なう。

なお，本研究は通商産業省工業技術院大型プロジェクト「電子計算機相互運用データベースシステムの研究開発」の一環として行なわれたものである。

参考文献

[1] J.Higashino et al.:“Knowledge-based Segmentation Method for Document Understanding”, 8th ICPR, pp.745-748

[2] 東野他:“マルチメディア文書画像理解システム”, 信学全大 1470(昭60年3月)



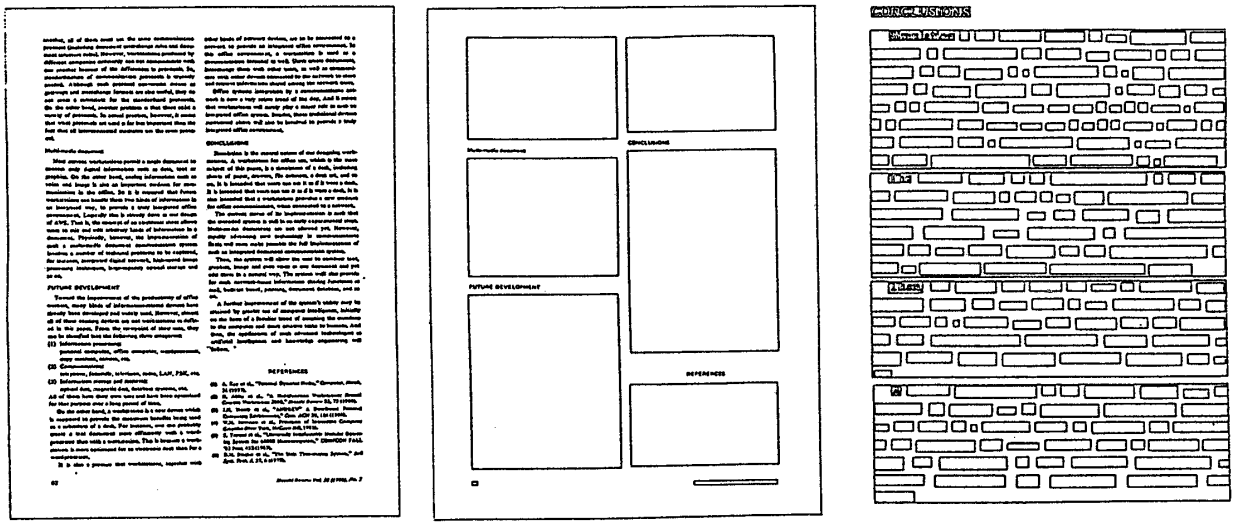
（ａ）原画像

（ｂ）文字認識結果

文字の下にある数字はフォント情報を表わす。

（１：ローマン体，２：サンセリフ体，３：イタリック体）

図３．文字認識と協調させた論理構造抽出



（ａ）原画像　　（ｂ）章題・節題の抽出　　（ｃ）パラグラフの抽出

図２．幾何学的情報を用いた論理構造抽出