4Y-10　知的ファイリングモデルシステムの開発（その３）－自動ファイリングのための文書理解の一方式－

中野康明　藤澤浩道
（株）日立製作所　中央研究所

はじめに　文書画像をファイルに登録するとき、各種の情報をキーボードから入力する必要があるが、この省力化のため自動ファイリングが強く望まれている〔1〕。自動ファイリングは検索を高度化するため、さらに重要となっている。この目的に、文書理解・文字認識技術の適用が期待される。文書理解は広く研究されているが〔2～4〕、実用上の観点からは未だ問題があると思われる。筆者等の開発した書式定義言語ＦＤＬ〔5〕も処理量で問題があった。本稿では実用的な文書理解の一手法を提案し、その実験結果について述べる。

文書識別と自動インデキシング　われわれの提案する文書の自動ファイリング方式では、複数ページからなる文書を区分ごとにまとめて、ヘッダシートと呼ぶ特定の用紙の後に続けて入力する。光ディスクへの文書の入力時（または入力後）、ヘッダシートを検出して、その上に記入した文書のクラス番号Ｃ＃とクラス内文書番号Ｄ＃などを認識する。Ｃ＃、Ｄ＃、及びページ番号により各文書画像が同定され、これらを介して書誌事項データベースと対応付けられる。

次に、文書識別により各文書画像のページ属性を決定する。ページ属性とは、表紙、図面、参考文献のページなど、検索時に効果的に使用できるような特徴を言う。ページ属性は文書の書式として画像の中に反映しているので、文書画像と書式を記述した知識ベースとを照合すれば、各画像のページ属性を決定できる。文書識別の副効果として、表題・ノンブルなどに対応する部分画像領域が特定されるので、文字認識による自動インデキシングも可能となる。

簡易な文書識別方式の提案　われわれは、強力な知識表現言語であるＦＤＬの思想を生かしつつ、実用性を考慮した書式定義言語ＳＦＤＬ（Simplified Form Definition Language）とその処理系を開発した。ＳＦＤＬでは、単純な前処理により画像を著しく単純化して現実的な処理量としたこと、書式定義に際して必須領域のみを指定するようにして、記述を単純化したことが特徴である。

一般の文書理解手法では、まず文字（文字成分）に対応する要素を検出し、その配列を調べて、行領域あるいはテキスト領域にまとめて行くが、ＳＦＤＬでは、連長フィルタリングにより文字成分の間の空隙を埋めた後、直ちに行領域を抽出する。

連長フィルタリングの概要を述べる（文書は横書きとする）。まず小さな孤立雑音を除去した後、走査線ごとに、白画素の連続する長さ（連長）を測る。この連長がある閾値より小さいとき、その白の連を黒に置き換える。この処理により、行領域がほぼ忠実に抽出される。連長フィルタリングは〔6〕に例があるが、本報告の方式は書式照合の前処理として用いる点が異なる。

連長フィルタリングを施した画像から、輪郭抽出などにより単連結領域を抽出し、行の候補とする。前処理により画像が単純化しているので、処理速度やメモリの問題は少ない。図１に文書画像Ａ、連長フィルタリングを施した結果Ｂ及び抽出した行領域候補Ｃ（原画像に重ねて表示）を示す。

ＳＦＤＬにおける書式定義では、各書式に出現が必須である領域のみを記述する。行領域を記述対象の単位とし、各書式の必須出現領域について絶対位置、大きさ、領域間の相対位置の値の範囲を記述する。書式の記述は主文、修飾文、関係文などの文を用いて行う。主文は領域を定義し、修飾文は領域について

Development of a Knowledge-Based Filing Model System (3)　Automatic Filing by Document Understanding
Yasuaki NAKANO, Hiromichi FUJISAWA
HITACHI, Ltd.

絶対的な制約を、関係文は領域間の相対的な制約を指定する。文法の詳細については〔7〕を参照されたい。

書式照合は、書式中に記述された構造が入力文書画像に存在するか否かを判定する処理である。まず書式中の各領域について修飾文で指定された条件を満たす行領域（一般には複数）を入力画像中から候補として抽出する。全ての領域に少なくとも一個の候補が検出されれば、次に関係文のチェックを行う。すなわち、各領域の候補間の相互関係が関係文の条件を満たすかどうかを調べ、満足しない候補を除去して行く。除去の結果一個も候補の存在しない領域が生じた場合には、この書式との照合は失敗とする。全領域に候補が残っていて、除去される候補が存在しなくなれば、入力画像と書式との照合は成功である。以上の照合処理を全ての書式について行い、照合が成功する書式を求める。本照合アルゴリズムは、文字認識の一手法〔8〕と共通した点がある。

実験結果　本手法による文書識別実験を行った。対象文書は、光ディスクに格納された特許公開公報８１ページ、特許明細書控１０６ページの総計１８７ページである。文書はＡ４サイズ、走査ピッチは８本／ｍｍであり、文書１枚は１７２８×２２８７画素の二値画像で表現される。

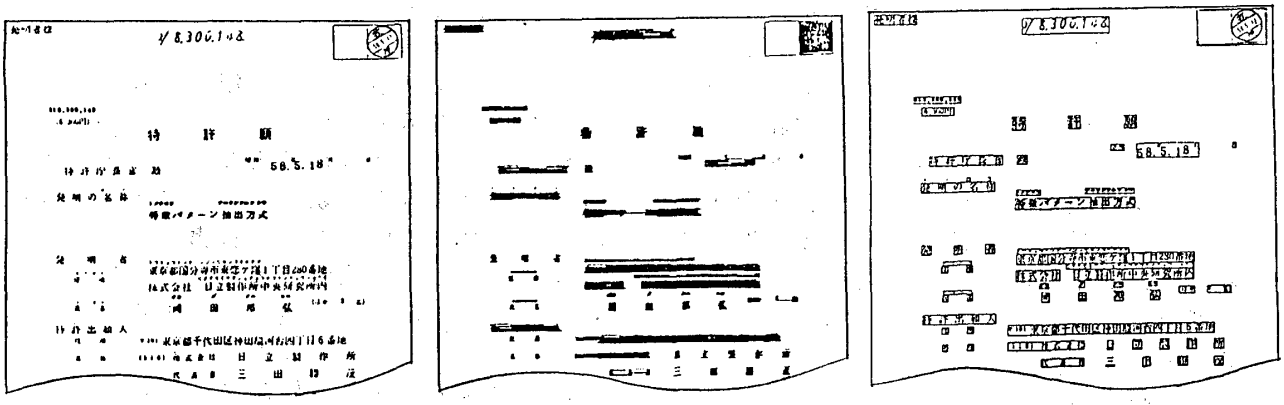
ページ属性は、特許公開公報では表紙、本文、図面の３種類とした。特許明細書では、特許願、実用新案願、請求の範囲、本文、図面、ヘッダの６種類とした。識別実験は公開公報と明細書では別々に行った。

書式定義は、机上で各ページ属性について作成したものを、識別実験結果により修正した。最終的な識別実験結果は、特許公開公報では未学習サンプルを含む８１ページ全てが正解となった。特許明細書では、全ページについて学習した書式を用いて、１０６ページに対し、正解１０４、拒絶２、誤り０の結果を得た。

なお、特許明細書では上記のページ属性のいずれでもないものがあるが、これらは一致する書式がなければ正解としている。

おわりに　自動ファイリングの一手法と実用的な文書理解手法を提案し、実サンプルに適用して良好な結果を得た。今後はＳＦＤＬの機能拡張、書式定義の対話的作成やサンプル画像からの自動学習機能などを検討したい。

〔1〕藤澤他，第31回情処全大2N-1，（昭60-9）
〔2〕野口他，第23回情処全大6C-1，（昭56-10）
〔3〕秋山他，信学論，66-D，（1），111（昭58-1）
〔4〕岩城他，信学技報PRL84-67（昭59）
〔5〕東野他，昭60信学総全大S10-2（昭60-3）
〔6〕K.Wong et al., IBM J. Res. Dev., 26，（6），647（1980-11）
〔7〕中野他，信学技報PRU86-30（昭61）
〔8〕森他，（昭50-8）信学論，58-D，（8），442（昭50-8）



A　文書画像　　B　連長フィルタリング　　C　抽出した行領域候補

図１　連長フィルタリング処理結果と行領域抽出結果