印刷文書読取システムの試作

4E-2

川又 武典, 松浦 一巳, 岡田 康裕, 依田 文夫, 伴野 浩三, 小林 啓二

三菱電機株式会社 情報電子研究所

1．まえがき

オフィスでは新聞・雑誌・書籍などの既存の文書を電子化して有効利用する作業や、パソコン・ワープロで作成した文書を再利用する作業が増加している。従来は、このような場合もワープロ等を用いて再度入力しなければならず、入力効率が悪かった。そこで、筆者らは開発中の文字認識及び文字切り出しアルゴリズムを組み込んだ印刷文書読取システムを試作し、新聞・雑誌等を読取対象として評価を行ったので報告する。

2．アルゴリズムの概要

処理の流れを図1に示す。まず入力した文書画像から文字列イメージを切り出し、切り出した文字列イメージから個々の文字パターンを切り出す。次に切り出した文字パターンを認識する。最後に単語・文法情報を用いて認識結果を修正する知識処理を行う。

2．1 文字列切り出し(1)

本文と図表・写真・表題などとが混在する複雑な構造（書式）の文書や段組のある文書に対処した。

まず文書画像を分割して得た各小領域で白／黒特徴をと直線特徴を抽出し、これを用いて段領域を分離する空白と罫線のセパレータを検出する。次に検出したセパレータに挟まれた黒特徴の連結成分から成る矩形領域を検出する。更に隣接する矩形領域を結合して段領域を検出し、検出した段領域内の矩形領域を並べて本文文字列を検出する。

2．2 文字切り出し

印刷文書に多くみられる文字間接触や、均等割付・禁則処理・欧文混在・半角／倍角文字の混在による不定文字ピッチに対処した。

まず文字列と直交する方向の周辺分布値に基づいて文字列イメージを分離し、基本矩形を検出する。次に検出した基本矩形の大きさから接触文字を含むと判定した矩形に対して、対応する周辺分布値の極小点の情報に基づいて分離候補位置を求め、基本矩形を分離する。更に連続する基本矩形を結合した各矩形に対して、矩形情報と認識情報（差分類似度）を併用して求めた文字評価値（文字らしさ）を用いて文字パターンを切り出す(2)。

RECOGNITION SYSTEM FOR PRINTED DOCUMENTS
Takenori KAWAMATA, Kazumi MATSUURA, Yasuhiro OKADA,
Fumio YODA, Kozo BANNO, and Keiji KOBAYASHI
Mitsubishi Electric Corporation

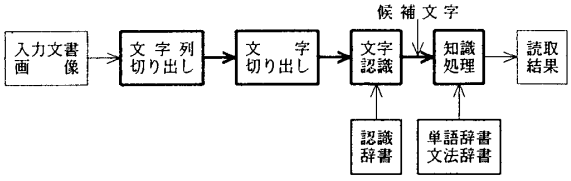


図1 処理の流れ

2．3 文字認識

文字認識では文字のつぶれ、かすれに対処するため、まず走査範囲を限定した背景特徴および文字の輪郭方向コード分布特徴を抽出する(3)(4)。次に文字の外側の特徴（文字の上部，下部の輪郭方向コード分布特徴および水平方向の背景特徴）を用いて大分類を行う。最後に文字の内部の特徴を併用して詳細分類を行い、候補文字を決定する。

2．4 知識処理

類似文字や大文字・小文字の識別能力を向上させるため、単語・文法辞書を用いた知識処理を行う。ここでは、候補文字の組合わせから構成される単語で単語辞書に登録されており、かつ単語間の接続が文法辞書で許容されているものを選ぶ(5)。なお、入力文章中に辞書未登録語がある場合のエラーに対応するために認識類似度をも併用する改良を加えた。

3．システム構成

先に報告した手書き文書用のシステム(6)を改良して試作した本システムの概観を図2に示す。本システムはハンディ・スキャナと文字認識装置と制御装置（パソコン）で構成される。文字認識装置は認識コントローラ

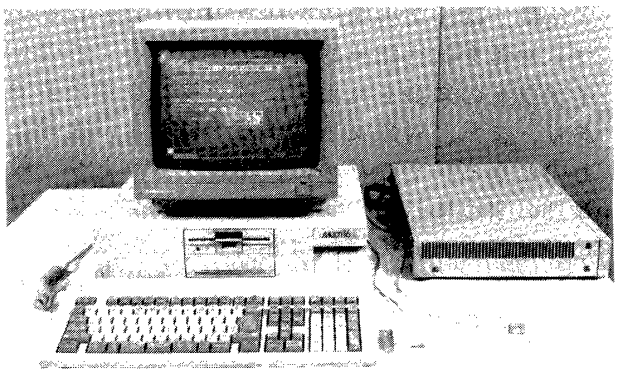


図2 試作システム

（ＲＣ）と辞書メモリのボードと、認識プロセッサ（ＲＰ）のボードの計２枚から成る。ＲＣはＲＰの制御を行うと共に、スキャナ及び制御装置とのＩ／Ｆを行う。また、スキャナから読み取った文書イメージ全体を記憶するためのイメージバッファメモリを有し、文字列の自動切り出し処理等に用いる。ＲＰは前節で述べた文字列切り出し、文字切り出し、文字認識、知識処理を行う。読取結果はＲＣを経由して制御装置に送られ、候補文字の表示による修正後、ワープロ文書ファイルに変換される。

４．読取例

図３から図６は制御装置の画面に表示された処理結果の例である。

まずスキャナで文書を走査すると入力された文書画像が表示される。そこで読取領域を指定すると（図３）指定された領域の文字パターンを切り出して認識し、知識処理を行った結果が表示される（図４）。読取結果に誤読がある場合は、その文字を指定すると候補文字が表示される。この中に正解文字がある場合は、その文字を指定すると文字の置換が行われる（図５）。文字修正を終えた文章はワープロ文書ファイルに変換され、以降ワープロ機能を用いて自由に編集することが出来る（図６）。

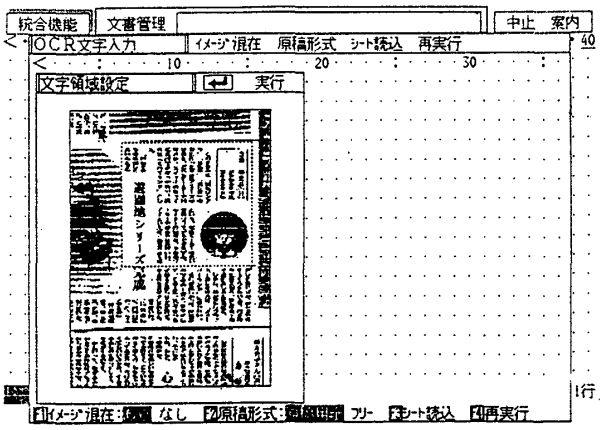


図３　読取領域の指定（　　指定領域）

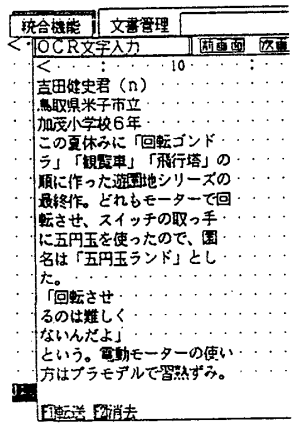


図４　読取結果の表示

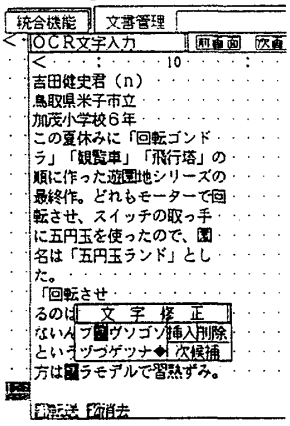


図５　誤読文字の修正

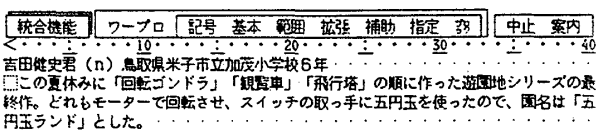


図６　ワープロ文書ファイルへの変換

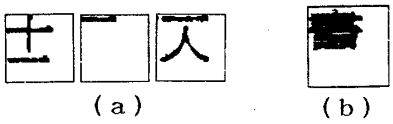
５．評価結果

新聞，雑誌を対象にしてシステムの性能評価を行った結果を表１に示す。表中（）内は誤読文字数を示す。

表１　評価結果

対象文書	文字数	認　識　率	知識処理後認識率
新　聞	984	89.6(102)	94.9(50)
雑　誌	942	97.9(20)	99.5(5)

縦書と横書では分離文字の性質が異なる。このため新聞では文字列「十三人」が図７(a) に示すように誤切り出しされ、「士一入」に誤読した。また、新聞は雑誌に比べ縦方向の文字サイズが小さい。このため横線の密度が高い文字などで図７(b) の文字「書」のようにつぶれが発生し、認識率が低下している。



（ａ）　（ｂ）

図７　誤読の例

６．むすび

印刷文書読取システムを試作し、ワープロ文書・雑誌などの読取が可能となった。しかし、一般文書を読取るには多様な文書構造に対応する必要がある。今後は更に各アルゴリズムの改良を図ると共に、より使いやすいシステムとする予定である。

文　献

(1) 松浦ほか：“印刷文書の文字列切り出し”，昭63信学秋全大，D-202(1988)．(2) 依田ほか：“認識情報を併用した文字切り出し方式”，昭61信学総全大，1519(1986)．(3) 依田ほか：“大局的特徴を併用したストロークマッチング法による手書き漢字認識の検討”，信学技報，PRL82-30(1982)．(4) 加藤ほか：“ナンバープレート認識技術”，三菱電機技報，Vol.62，No.2(1988)．(5) 水上ほか：“単語の文法的接続情報を利用した日本文認識の後処理”，昭61信学総全大，1547(1986)．(6) 伴野ほか：“日本語文書読取システムの試作”，昭59信学総全大，1613(1984)．