UDC〔681.327.12:003.324.2〕:〔681.32:621.397.3〕

日本語ドキュメントプロセッサ XP-70S

本郷 保夫(ほんごう やすお) 木内 哲夫(きうち てつお) 岸 正弘(きし まさひろ)

1 まえがき

ワードプロセッサ（ワープロ）やデスクトップパブリッシング（DTP）の普及により，オフィス内での印刷文書の量は増大している。これらの印刷文書をコンピュータが処理できるデータに変換する装置として，日本語印刷文書読取り装置 XP-50S[(1)] を開発し，昭和63年から市場に供給してきた。XP-50S は日本語 OCR（光学式文字読取り装置）として，日本語で書かれた多量のテキストを入力し，テキストデータベースを作成するのに利用されている[(2),(3)]。また，電子出版システム，機械翻訳システム，テキスト朗読システム，点字印刷システムなどのテキスト入力にも利用されている。

XP-50S の利用分野が拡大するとともに，

(1) 英文の読取り機能

(2) 横長文書の読取りのための90度回転機能

(3) 図面や名簿などの大形文書（A 3 判）の読取り機能

の必要性および要求が強まった。これらの機能をサポートし，さらに日本語 OCR の適用範囲を拡大すべく，和文および英文を高速で読み取る日本語ドキュメントプロセッサ XP-70S を開発したので，以下にその装置の機能，原理などの概要を述べる。また，XP-70S を応用して数表や名簿などを読み取る表読取りソフトウェア TPal（ティパル）も紹介する。

2 日本語ドキュメントプロセッサ XP-70S

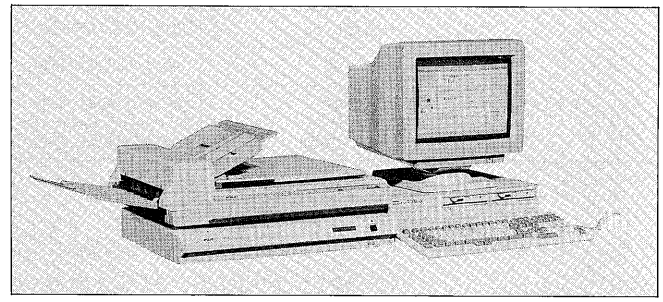
2.1 概 要

XP-70S は，日本語および英語で書かれたワープロ文書，書籍，新聞などの印刷文書を，高速でしかも高精度で読み取る日英2か国語 OCR である。XP-70S システムの外観を**図1**に示す。

XP-70S システムは，XP-70S 本体，イメージスキャナ，そしてパーソナルコンピュータ（パソコン）の三つで構成される。

XP-70S 本体は，90度画像回転機能とA 3 判文書を入力する画像メモリ拡張機能とをオプション機能として持っている。

図1 XP-70S システムの外観



N89-5852-6

イメージスキャナはB 4 判用とA 3 判用の2種類があり，A 3 判の大形文書も入力することができる。

XP-70S は30文字/秒の読取り速度で，明朝体やゴシック体で書かれた印刷文書に対して，マルチフォント混在読取りができる。また，認識対象文字に含まれない文字でも，使用頻度の高い文字を100文字まで，自由に学習・登録により追加ができる。

XP-70S の認識した結果は，パソコン上で動作する基本ソフトウェア XPal（エックスパル）により編集して，MS-DOS〈注1〉 のテキストファイルまたは OASYS〈注2〉 の文書フロッピィとして出力することができる。XP-70S の基本仕様を**表1**に示す。

2.2 ハードウェア構成

(1) ホストパソコン

XP-70S 基本システムのホストパソコンとしては，FM R60/70/80 シリーズを使用している。XP-70S 本体とパソコンとは，高速パラレルインタフェース（SCSI：Small Computer System Interface）で接続しているので，パソコンだけでなくワークステーションとの接続も可能である。

(2) 日本語ドキュメントプロセッサ XP-70S

〈注1〉 MS-DOS：米国マイクロソフト社の登録商標

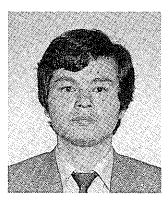
〈注2〉 OASYS：富士通(株)のワードプロセッサ

本郷 保夫

昭和52年入社。ビデオセンサ，日本語OCRなどのパターン認識応用装置の開発に従事。現在，富士ファコム制御(株)技術本部第五技術部課長。

木内 哲夫

昭和54年入社。画像処理・文字認識アルゴリズムの研究開発に従事。富士ファコム制御(株)技術本部第五技術部課長補佐。現在，カーネギーメロン大学に留学中。



岸 正弘

昭和50年入社。空力プセルチェッカ，日本語OCRなどのパターン認識応用装置の開発に従事。現在，富士ファコム制御(株)技術本部第五技術部。

XP-70S は，32ビットマイクロプロセッサを使った管理モジュールと文字認識モジュールとで構成されている。管理モジュールはパソコンからのコマンドにより，文書画像をスキャナから入力し，文章領域の行や文字を次々と抽出する。文字認識モジュールは，低次・高次特徴抽出部，分類処理部，認識処理部で構成されており，パイプラインによる並列処理で高速文字認識を実現している。また，XP-70S は次の二つのオプションカードを実装できる。

(a) 90度回転カード

入力した画像を90度回転して読み出せるので，横長の原稿でも読み取ることができる。

(b) 拡張画像メモリカード

最大A 3 判原稿の画像まで入力することができる。

(3) イメージスキャナ

イメージスキャナはB 4 判用とA 3 判用との2種類があり，さらにそれぞれ自動給紙機構（ADF）付きとフラットベッドタイプの2種類がある。A 3 判用の場合には，XP-70S 本体に拡張画像メモリカードが必要である。

表1　XP-70S の仕様

項　目	仕　様
原稿の種類	ワープロ文書，書籍，雑誌などの日本語印刷文書（縦書き，横書き）
原稿サイズ	B 4 判以下（オプションでA 3 判まで可能）
認識対象文字	印刷文字約3,800字種
読取り速度	30文字/秒
文字フォント	明朝体，ゴシック体，教科書体などのマルチフォント混在読取り
文字サイズ	7～28ポイント（2.5～10mm）
学習文字数	100文字まで登録可能
出力形式	MS-DOS テキストファイル，OASYS の文書フロッピィ

イメージスキャナは 400dpi（約16本/mm）の高解像度であり，2.5mm 角の小さい文字まで読取りが可能である。また，濃度も 8 段階で設定できるので，紙質や印刷状態に合った濃度が選択できる。さらに，新聞などの低品質文字に対して最適な 2 値化を行う自動 2 値化カードもオプションカードとして利用できる。

2.3　基本命令

XP-70S はホストパソコンから以下の基本命令を受けて，文書の読取り処理を実行する。

(1) 画像入力命令

XP-70S はスキャナを駆動して，ホストパソコンから指定された濃度で画像を画像メモリに入力する。

(2) 画像転送命令

XP-70S からパソコンへ画像データをアップロードする。またはパソコンから XP-70S へ画像データをダウンロードする。この画像転送では，原画像または縮小画像（1/2，1/4，1/8）を指定でき，転送する範囲も任意である。

(3) テキスト読取り命令

XP-70S は指定された読取り枠内について文字行および文字を抽出して，認識した結果（候補文字10位までを含む）をパソコンへ送信する。

(4) 文字登録・削除命令

文字登録命令は，XP-70S が認識用辞書にない文字パターンの辞書を自動作成して，その文字パターンを指定の文字コードで登録する。これは，XP-70S で最大100文字まで登録できる外字辞書領域を持っている。また文字削除命令は，外字辞書に登録した文字パターンの辞書を削除する。

(5) 外字辞書転送命令

文字登録で作成した外字辞書を，XP-70S とパソコンの間で転送する。

図2　XP-70S の文書処理の流れ

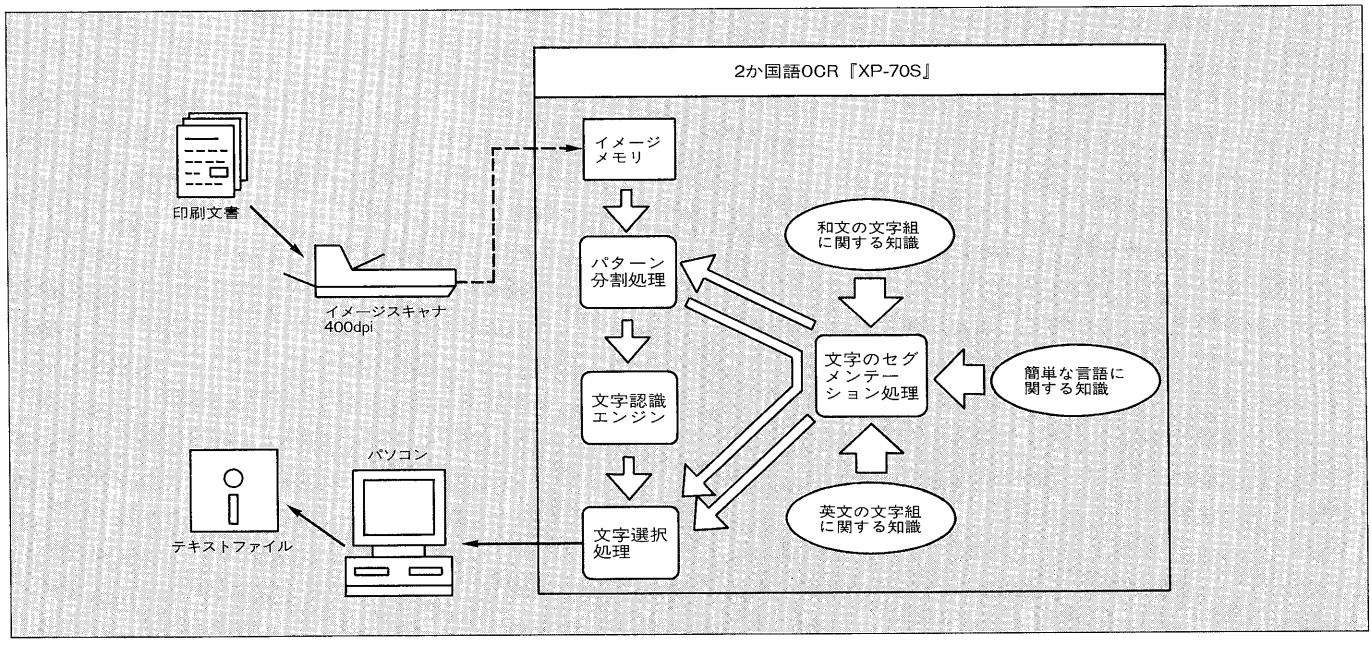


図3　セグメンテーション処理の難しい文字パターンの例

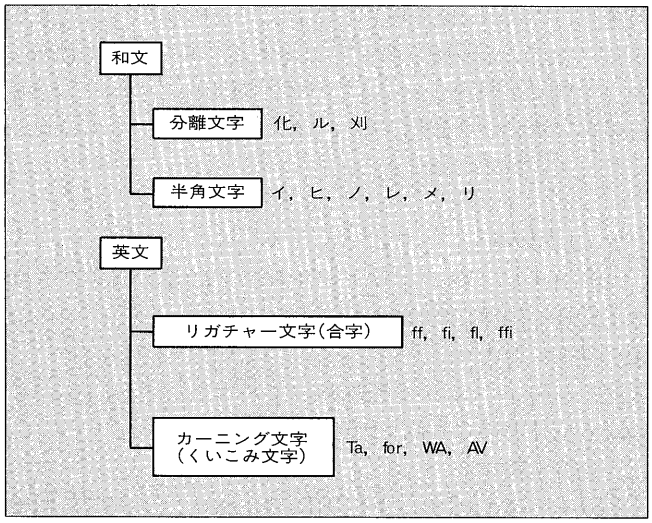


図4　二次元特徴量の例

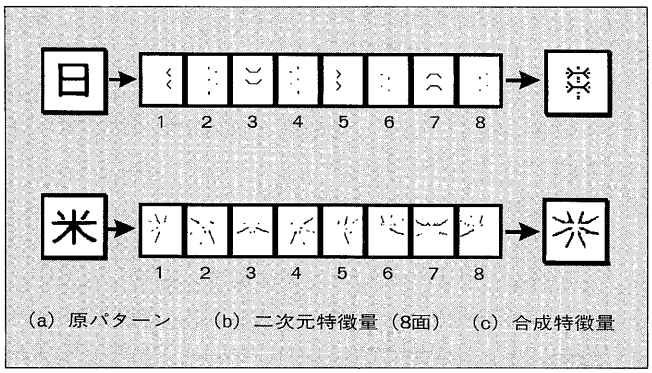


図5　読取り操作画面

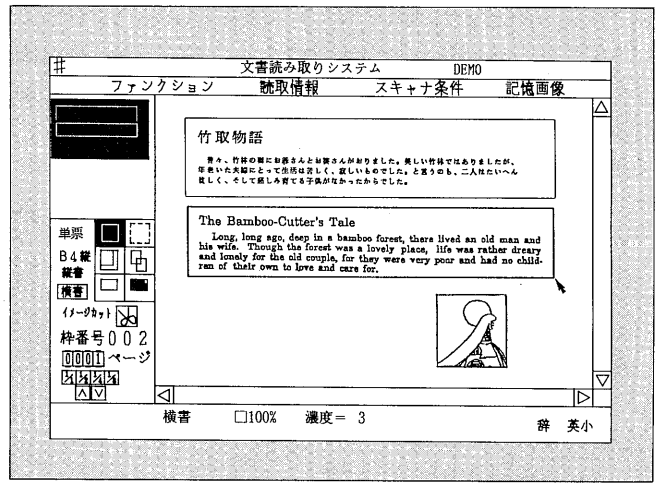
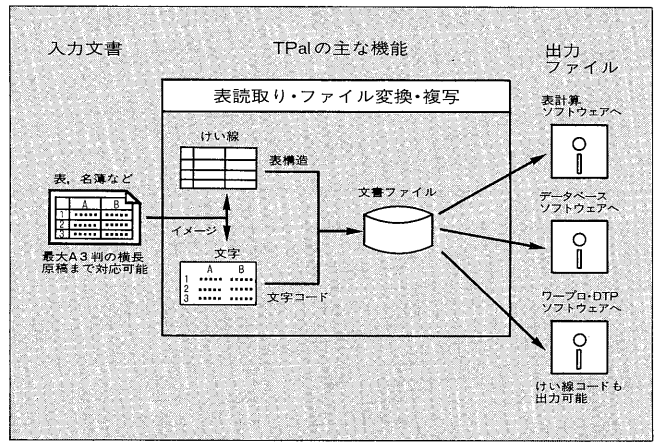


図6　TPal の処理フロー



2.4　文書読取り処理

XP-70S の文書読取り処理は和文と英文とのテキストに対応して2種類のセグメンテーション処理[(4),(5)]が実行できる。その文書処理の流れを**図2**に示す。和文と英文との違いは，文字の形と大きさ，そして文章を構成する文字の配置，つまり文字組とにある。和文での文字組は定ピッチであるのに対して，英文ではプロポーショナルスペーシング(文字間スペースの均等配分)となっている。したがって，1文字ずつにパターンを分離するセグメンテーション処理が，和文と英文とで異なっている。

和文および英文において，セグメンテーション処理の難しい文字パターンの例を**図3**に示す。和文の場合では，漢字の「化」のように，半角文字の「イ」と「ヒ」とに分離して読み取れる分離文字がある。また，英文の場合には複数の文字が合体したリガチャー文字や文字同士が接近してくいこんだカーニング文字がある。XP-70S では，これらの文字を認識するのに，文字の形や大きさだけでなく，文字組に関する知識や簡単な言語的知識を利用して，読取り精度を向上させている。

XP-70S の文字認識エンジンで採用している認識アルゴリズムは，工業技術院電子技術総合研究所の技術指導を得て開発した二次元特徴相関法[(1)]であり，その二次元特徴量の例を**図4**に示す。二次元特徴量は，文字線の境界から垂直方向に波を伝搬させて，波が衝突してできる波頭の形を抽出したものである。波の衝突は文字線同士の距離が変化しても変わらないので，この二次元特徴は文字の字体に対するフレキシビリティが高いと言える。

2.5　基本ソフトウェア XPal

XPal はホストパソコン上で動作する XP-70S 用のマンマシンインタフェースソフトウェアであり，読み取ったテキストデータを編集，保存，管理する機能を持っている。

(1)　読取り操作

XPal の読取り操作画面を**図5**に示す。まず，イメージスキャナの濃度を指定して画像を入力する。部分拡大機能により，入力した画像にかすれやつぶれがないことを確認して，読取り範囲を枠で指定する。この際の読取り条件としては，和文か英文かにより文字種類の指定を行う。また，文字の変形率(文字の高さと幅の比)の指定も必要により行う。例えば，一般の文書では文字は正体で，変形率が100%であるが，新聞の場合は文字の幅の方が高さより大きい平体であり，変形率は80%である。読取り枠と読取り条件とを設定して，XP-70S に読取りを実行させると読取り結果がテキストデータとして作成される。

(2) テキスト編集操作

読取りデータは，候補文字（最大10個まで）を表示させて，その中から正しい文字を選択して修正することもできる。このとき，原稿のイメージを同時に表示することができ，修正が容易に行える。組込みのワープロ機能により追加・修正することもできる。

(3) ファイル操作

修正編集したテキストデータは，OASYS 変換機能により，OASYS 文書フロッピィとして出力できる。また，テキストをそのまま MS-DOS のテキストファイルとして，利用することもできる。

3 表読取りソフトウェア TPal

TPal はパソコン上で動作する XP-70S の応用ソフトウェアであり，数表，名簿，住所録などのさまざまな表形式文書を読み取ることができる。TPal の処理フローを**図6**に示す。イメージスキャナから表などの原稿の画像データを入力して，TPal が表のけい線と行間の空白領域を自動認識し，表構造を抽出する。TPal は表の各セル内を XP-70S に文字認識させて，表形式の文書ファイルを作成する。この読取り結果は，市販の表計算ソフトウェア，データベースソフトウェア，そしてワープロ・DTP ソフトウェアで処理できる形式にファイル変換されて，利用される。したがって，TPal と表計算ソフトウェアを使えば，統計表からグラフを新しく作成することもできる。また，TPal には，行列の編集機能や文書ファイルの修正確認作業でのイメージ表示機能などが完備している。TPal の基本仕様を**表2**に示す。

表2 TPal の基本仕様

項目	仕様
原稿の種類	統計表，数表，名簿，住所録などの表形式文書
原稿のサイズ	B4判以下（オプションでA3判まで可能）
表タイプ	3タイプからの選択方式 （一部けい線のない表も読取り可能）
読取り速度	1枚/分（20行5列，1,000文字相当）
表の行列数	最大128行50列
文書ファイルのサイズ	最大99ページ（1表＝1ページ）
自動処理枚数 (ADF付スキャナ使用)	A4判普通紙で最大50枚（用紙高さ4mm以下）
出力ファイル形式 (MS-DOS ファイル)	市販の表計算ソフトウェア，データベースソフトウェア，ワープロ，DTP ソフトウェアに対応

4 あとがき

日本語ドキュメントプロセッサ XP-70S は和文と英文の両方を読み取れる2か国語 OCR として開発し，従来の日本語 OCR の適用範囲をさらに拡大してきた。また，近年，漢字を読む OCR ソフトウェア[6]が登場しており，5～10文字/秒と低速ではあるが，パソコンユーザーにも OCR への理解と関心を高めるのに役立っている。さらに日本語 OCR の普及には，①文字認識速度が速く，②一般の文書を高精度で読み取り，③パソコンからワークステーションまで幅広くシステム化できる低価格化の実現が必要である。

マイクロエレクトロニクスの進歩により，マイクロプロセッサの高速化やメモリの高集積化が一段と進んでおり，日本語ドキュメントプロセッサのコストパフォーマンスも確実に向上している。今後の課題として，文書のレイアウトを理解して，自動的に文字領域を抽出する自動領域抽出機能と，読み取ったテキストデータの中の誤読箇所を自動的に検出する単語照合機能[7]との開発を進めてゆきたいと考えている。また，手書き文字を高精度で読める文字認識アルゴリズムの開発にも注力してゆく所存である。

参考文献

(1) 本郷保夫ほか：文書・画像処理における FA パソコンの応用，富士時報，Vol.62，No.7，p.479-483（1989）

(2) 堀池博巳ほか：日本語テキスト・データベースの作成・利用ツールについて，1990年情報学シンポジウム，p.153-160（1990）

(3) 柏原斌紀ほか：和文印刷物の読取りと再編纂，電総研彙報，Vol.55，No.2，p.215-227（1991）

(4) Ogura, I. et al.：A New Segmentation Method for Japanese Printed Documents Using Layout Knowledge, VISION'90, November 12-15, 1990, MS90-611, p.15-59～15-77（1990）

(5) 岡田正年ほか：英文文書における単語間空白検出の一検討，1990年電子情報通信学会秋季全国大会，D-387，p.6-389（1990）

(6) 本間健司：漢字を読む OCR ソフト登場，日経パソコン，1990年10月29日号，p.118-122（1990）

(7) 紺野章子ほか：日本語 OCR のための誤読修正方式の一検討，情報処理学会第41回全国大会，46-L，p.2・144-2・145（1990）

e-Front runners

＊本誌に記載されている会社名および製品名は，それぞれの会社が所有する
商標または登録商標である場合があります。