

漢字認識機能をもった文書処理システム (1)

ー 基本機能モジュール ー

5J-2

高宮 白、 越智 英夫、 柴沼 英博

協同システム開発株式会社

1. はじめに

情報化の進展にともないワープロ、印刷物等の活字文書がオフィス内にも大量に出回ることになった。

これら活字文書処理の効率化は事務作業効率の向上にも大きく係わって来る。我々は、活字文書をタイプしなおす事なく、そのまま直接入力(文字認識)でき、編集、保存、検索等が容易に行える「文書処理システム」を開発した。¹⁾

本稿では、この「文書処理システム」の中核を占める基本機能モジュール(基本・支援ソフト)について述べる。

2. 開発システムの全体構成

開発システムは、基本ソフト、支援ソフト、応用ソフトの3種類に大別される。①基本ソフトは、入力から出力までの文書処理を構成する基本要素の中から、応用ソフトで共通に利用する機能を抽出した部品となるものである。②支援ソフトは、基本ソフトを組み合わせて応用ソフトを構築する事と、利用環境の構築及び運用の効率向上を支援するものである。③応用ソフトは、基本ソフト・支援ソフトをベースとし、さらに業種・業務に固有の機能、独自のユーザインタフェース等を付加して作成されるアプリケーションシステムである。²⁾

なお、開発システムの全体構成を図1に示す。

3. 基本機能モジュール

3.1 基本ソフト

(1) 文書入力

文書入力は活字文字と図形・表等のイメージの混在した文書を、定型・非定型の別なくイメージスキャナを介してコピー感覚で入力できる機能を提供する。

基本ソフト	文書入力	
	文字認識	
	文書管理	
	データ編集	
	レイアウト編集	
	文書出力	
支援ソフト	実行制御	
	業務手続き支援	
応用ソフト	標準システム	中小企業向け 文書作成管理システム
	専用システム	建築設計業向け 設計計算書作成管理システム
		情報サービス業向け 資格試験問題作成管理システム
		コンサルタント業向け 報告書作成支援システム
		ソフトウェア業向け マニュアル作成管理システム
		情報サービス業向け 実績報告書作成管理システム
		書籍小売業向け 検索システム

図1 システムの全体構成

イメージ領域と文字認識の対象となる活字文字領域は、マンマシンインタフェース(MMI)を介して分離する。印刷、表示に必要な表現属性(位置情報、文字・イメージの別等)はこの時点で取り込まれ、頁内のレイアウト構造を持った文書として保管される。文字領域はこの時点では印字パターンのままのイメージデータであるが、次の処理の文字認識の認識率を向上させるために画質改善(雑音除去、傾き補正)を行うことが出来る。イメージスキャナで読み取り可能な文書サイズはA4であるが、それ以上のサイズの文書(例えば大型図面)も考えられるため、一枚の文書を分割してスキャナより入力し、その後統合して保管する事を可能としている。

(2) 文字認識

文字認識はイメージのままの文字データを、文字認識装置を介して文字コード（JIS漢字コード）に変換しさらに、認識精度の向上を図るため認識結果の誤りを修正する機能を提供する。

文字認識自体はハードウェアで行うが、現在使用している認識装置は文字パターンを文字認識専用の辞書と照合しながら複合類似度法を用いて認識し文字コードに変換している。認識結果は1文字に特定される場合、複数の候補文字を含む場合及び認識出来ない場合がある。また、誤った文字に認識する場合も有り得る。認識結果に複数の候補文字が存在する場合は、対話的にその中から選択する事が出来る。

また、誤認識や認識出来なかった文字は、画面上に表示されたイメージのままの原文や認識後の文章を見ながら修正が可能である。

(3) 文書管理

文書管理はコード化された文字情報とイメージの混在した文書の維持管理及び検索のための機能を提供する。文書保管の環境はオフィス内で日常使用しているキャビネ、バインダ、フォルダ等を模擬しており、アイコンで表示される。マウス操作でアイコン指示することで文書の登録、検索、廃棄が行えるようになっている。また、検索の効率を左右するキーワードの付与は、文書の内容からキーワードの候補となる単語を自動的に抽出する事が可能となっており、登録時の負担軽減を図っている。

(4) データ編集

データ編集は文書の文字情報編集と、文章の論理的意味付けによる論理単位の編集を行う機能を提供する。文書は頁単位の配置情報と関連付け構造化されるレイアウト構造と、章、節、項といった論理的な構成と関連付け構造化される論理構造の2つの面で捕らえる事が出来る。この様な文書構造情報を統一的概念で体系的に表現する標準規格案にODA/ODIF (Office Document Architecture and Interchange Format) があり、当システムでも取り入れている。文書の編集において、内容の編集を行った文書は、決められたレイアウトに従ってページ内に割り付けられ表示・印刷の形式となる。そのために、レイアウト付けに必要な割付構造を生成する機能と、文章の内容をその構造に基づいて、禁則処理を施し配置する機能を持っている。

(5) レイアウト編集

文書の出力レイアウトの編集をディスプレイ画面上で行う機能を提供する。ここでは、データ編集とは逆に文書をレイアウト構造から捕らえている。レイアウト構造はページ、ブロックという編集単位で構成され、階層構造を持っている。ブロックは文章、図形等の内容を伴っており最小の編集単位である。文書内において、ページ、ブロックに対する削除、複写、移動、表示上の属性変更、さらにブロックに対して拡大・縮小、重なり合ったブロックの上下変更等の機能を持っている。また、文書間での編集単位の複写も可能である。

(6) 文書出力

編集した文書をプリンタへ印刷する機能及び文書交換用の標準形式に変換する機能を提供する。

オフィスにおける文書の印刷形態は、体裁、ページ数、部数も多種多様で、その条件も印刷時に変更される事が多い。当システムでは、このような要求を満たす機能を持っている。

また、他システムとの文書交換を可能とするために、データ交換用媒体を通し規格化された文書の入力、出力機能を持っている。規格としてはJISの「日本語文書交換用ファイル仕様」を採用している。

3.2 支援ソフト

(1) 実行制御

基本ソフトと応用ソフトを一元的に管理し、各機能の実行処理が円滑に行えるように制御する機能である。

(2) 業務手続き支援

利用者の作業負担を軽減するため、定型的な操作を「手続き」として登録し、実行する機能である。

4. おわりに

本稿では、情報処理振興事業協会の委託を受けて開発を行った「中小企業向け簡易文書入力システム」における汎用的な基本部品について述べた。

今後はこれらの部品を利用して業種、業務に対応した応用ソフトの一層の充実を図って行きたい。

[参考文献]

- 1) 弘重 款, 他 : 「中小企業向け簡易文書入力システム開発計画」テクニカル・レポート NO. 1、協同システム開発 1988
- 2) 平野 智久, 越智 英夫, 高宮 白 : 「漢字認識機能をもった文書処理システム」ニエアプリケーションシステムー、情報処理学会第39回全国大会 1989