JICSTファイル作成におけるデータの OCR入力の実際

北井 隆*

Practice of on-line input to computer using optical character reader for JICST files

KITAI Takashi

〔著者抄録〕 JICSTにおいて、OCR技術を利用したパソコンベースのデータ入力システムが開発され、平成5年8月より試験的ではあるが実稼働を開始した。この入力システムでは、OCR機器およびOCRソフトについて市販のものを活用する方式としたことにより、比較的安価に構築できた。このOCR入力システムを用いた、これまでの半年余の運用経験について振り返り、印刷媒体の電子ファイル化・データベース化における実効性について報告した。

〔著者付与キーワード〕 OCR〔読取装置〕, 文字認識, データ入力システム, パーソナルコンピュータ, JICST, 実用化

〔Author Abstract〕 A PC-based data entry system using optical character readers (OCR) is developed at the Japan Information Center of Science and Technology (JICST). This type of system is designed to help JICST feed data into a computer for the purpose of database construction, and has been successful since the start of test operation in August, 1993. The effective use of OCR hardware and software available has made the cost of constructing such a system much less expensive. On the basis of the six month's experience in test operation gained up to the present, it should be concluded that the system will be really beneficial not only for the construction of databases but also for the electronic filing of printed materials.

〔Keywords by author〕 OCR, character recognition, data entry system, personal computer, JICST, implementation

* (株)ディック・アルファ 技術部 (〒100 千代田区永田町2-5-2) Tel.03(3581)6411
DIC・Alpha Company. Technical Division. (5-2, 2-chome Nagata-cho Chiyoda-ku, 100)

1. はじめに

JICSTでは、現在、データベース作成方法の改善に取り組んでおり、その一環として、パソコンを使用して抄録を作成したり、索引を行う「抄録・索引支援システム」(第1バージョン)が、平成5年8月よりスタートした。このシステムでは、論文記事の標題や著者抄録、記載ページなどの書誌事項データの作成も行えるようになっている。

これにより、従来パンチ業者に委託していた書誌および記事原稿のパンチ入力工程を省略できるようになった。特に、書誌事項データの作成では、その入力手段として、市販のOCR用機器とOCRソフトを上記システムに組み込み、データ入力を効率良く行えるようにしている。

この書誌事項データの入力支援としてのOCR入力方式の概要と、これまで半年余にわたって運用してきた経験を実例として報告する。

2. OCRとは

OCRは、Optical Character Readerの略で、「光学的文字読み取り(装置)」と訳される。このOCR技術の先駆的実用化の例としては、郵便番号の読み取りが一般にもよく知られ、他には、金融機関での通帳番号や伝票番号の読み取りに利用されている。いずれも数字が読み取りの対象である。数字は字形が単純かつ個数も少ないことから、アルファベットや仮名はもちろん、漢字に比べてはるかに実用化しやすい面があったと言える。

現在では、印刷活字であれば、漢字までほぼ完全に読み取りが可能であり、最近においては、手書き文字の読み取りについても限界はあるものの、一部可能になっており、今後のさらなる実用化が期待される。

このように文字認識については、実用の一定レベルに十分に達し、認識に必要なイメージスキャナや認識ソフト等のハード・ソフトの性能は向上し、その価格についても個人使用レベルで十分手の届くところまできている。

文字認識の基本的なプロセスは、図1のような流れで示される。

- (1) イメージスキャナによる文字の読み込み
雑誌、論文集、会議録などの認識対象データを画像データとしてパソコンに読み込む。
- (2) 文字の切り出し処理
1文字ずつの切り出しを行う。
- (3) 正規化処理
認識処理を容易にするため、文字の大きさを整える。
- (4) 特徴抽出
これも認識処理を容易にするため、文字の持つ特異なパターンや、他の文字との識別パターンを抽出する。
- (5) マッチング(照合)
入力した対象文字のパターンを登録された標準パターンと比較し、照合・認識を行う。
- (6) 同定(検証)
以上のプロセスで得られた文字認識結果を検証する。
また、イメージスキャナは、次の原理で動作する。
 - ① スキャナは、光源からの光を原稿にあて、その反射光の強弱を光学センサによって電気信号に変換する。さらに、この信号をコンピュータで取り扱えるデジタルデータにしてコンピュータに送る。
 - ② 読み取りは画像を多数の点に分解することによって行われる。キャリッジ上の光学センサで、1ライン分の読み取り(主走査という)を行い、

図1 文字認識のプロセス

キャリッジの移動で次のラインの読み取り（副走査という）を行う。これによって、画像を縦横のマスメに区切ってデジタルデータとする。この一つ一つのマスメに相当する部分は画素と呼ばれる。

3. システム構成

OCR 技術は、「抄録・索引支援システム」の中の書誌事項データ入力サブシステムにおいて、データ入力手段のメインとして組み込まれている。そのシステム構成図を、図 2 に示す。

本システムを構成する機器と OCR ソフトは以下の通りである。

(1) 機器

(A) パソコン：Macintosh Centoris 650

(B) 卓上型イメージスキャナ：エプソン GT-8000

（主な基本仕様）

読み取り原稿サイズ：最大 A 4 サイズ

原稿条件：写真、印刷物など、表面が滑らかなもの

解像度：主走査 400 D P I

副走査 400 D P I

読み取り時間：文字・線画で、
約 1.5 msec / line

(C) プリンタ：MICROLINE 801 PS II + F

(2) OCR ソフト

(A) 英文ソフト：OmniPage Professional v.2.0

[CAERE 社/株ヒューリンクス]

（主な仕様）

① 最小 6 ポイントから最大 72 ポイントまでのほとんどの書体に対応。

② 原稿の品位や紙質に応じてイメージスキャナの濃度を細かく調整可能。

③ インタラクティブなスペルチェック機能を持つ。

④ 文字認識処理で対象の印刷文字の周辺をウインドウ表示する。

(B) 和文ソフト：MacReader plus v.2.0

[株メディアドライブ研究所]

（主な仕様）

① JIS 第 1 水準文字、第 2 水準の一部の漢字（約 200 文字）を認識。

② 書体は、明朝体、ゴシック体などに対応。

③ 日本語、英語のどちらにも対応したバイリンガル OCR。

(3) システムの概要

書誌事項データ入力サブシステムでは、データ入力画面を開いて、所定のデータ項目エリアに順次データを入れていけば、自動的に資料単位かつ記事単位に書誌事項データが作られる仕組みになっている。その仕組みでは、データ入力の手段の一つとして OCR を使うようにしてあり、これは随時利用できる。また、OCR 部分のみの単独利用も、もちろん可能である。

本サブシステムの入力項目は、次の通り。

① 記事言語

② ページ

③ 著者名、そのフリガナ、機関名リンク番号

④ 著者所属機関名、著者名リンク番号

⑤ 和文標題

⑥ 和文同所抄録

⑦ 英文標題

⑧ 英文同所抄録

これらのうち、OCR を使うのは、⑤～⑧の標題、抄録データであり、他のデータ項目は作業効率上、ほとんどキーボード入力で処理している（著者名やそのフリガナ等の短いデータは、パンチ業務の

図 2 OCR システム構成

非専門家でも、OCR よりキーボードから入力する方が速いため)。

4. 実施状況

平成5年8月から実稼働に入り、これまで(平成6年2月末時点)に約10,000論文の書誌事項データをOCR入力した。

作業方法と処理体制を以下に示す。

4. 1 作業方法

標題、抄録などの和文データはMacReader plus をもっぱら使用し、英文データの場合は、OmniPage Professional を使用する。MacReader plus でも英文データの処理は可能であるが、作業効率の違いから、和文データ専用としている。

実際の作業は、次の手順で行われる。

1 OCR 先行入力

OCR ソフトを起動してから、原稿をイメージスキャナにセットして原稿の画像取り込みと自動文字認識を行う。文字認識終了後、画面上で誤認識文字の校正を行い、最後にフロッピーディスクへの保存を行う。本作業で、標題と同所抄録の作成(従来のパンチ入力に相当)が行われる。

なお、標題と同所抄録は、大部分の場合、論文記事の先頭ページか、または最終ページにあり、一括して標題と同所抄録を採取している。標題については、目次ページから採取する場合もある。

2 書誌事項データ入力

抄録・索引支援システムの書誌事項データ入力画面を開いて、記事言語、ページ、著者名などはキーボードから入力し、上記のOCR先行入力で作成した標題と同所抄録は、パソコン機能の“コピーアンドペースト”ないし、“カットアンドペースト”を使って、所定のデータ項目欄に貼り込みを行う。この例を、図3、その入力原稿および処理例を図4(1)、(2)に示す。

3 校正作業

キーボード入力とOCRデータの貼り込みによって完成した書誌事項データについて、抄録・索引支援システムのゲラプリント機能で校正用ゲラを出力し、記事単位で入力内容のチェックおよび修正を行う。

4. 2 処理体制

1日の作業時間を、午前、午後前半、午後後半の三つに分け、それぞれにローテーションで作業者を配置する3交代制で処理する。3交代制とした理由としては、作業者の健康に留意したこと、また作業効率を考慮したことが挙げられよう。

なお、機器構成はパソコン4台、イメージスキャナ2台であるため、基本的にパソコン2台はOCR先行入力専用、他の2台は書誌事項データ入力専用としている。

5. OCR 方式と従来方式との比較

これまで書誌事項データの inputs は、主に次の2段階で行われてきた(従来方式)。

[第1段階] パンチ入力指示作業

前記3.(3)の各パンチ対象項目について、そのパンチ範囲、著者や機関名の省略形の指示を、論文資料および補助資料に対して行う。

[第2段階] パンチ作業

上記の指示に基づいて、パンチ対象データを専門パンチャーが入力作業を行う。(入力結果は、指定のフォーマットの磁気テープにして納品される)

OCR方式は、この2段階の作業を一括して行うもので、工程の統合と若干の工期短縮を実現している。さらに、従来の方式では、パンチ専門の熟練作業者を必要とするが、OCR方式はそのような専門熟練者を必要としない。一般の事務作業レベルの人で十分であり、せいぜいワープロができれば有利と言えるかもしれない。

抄録文部分について、両方式の作業時間比較を以下に示す。

① OCR 方式による作業時間は、

The screenshot shows a software window titled "編集 OCR 機能テスト ウィンドウ アプリット". The main area is labeled "書誌情報登録・更新" (Journal Information Registration/Update). It contains several input fields: "受入番号" (Received Number) with value "940223811", "記事番号" (Article Number) with value "0002", and "ページ" (Page) with value "9-17". There are also fields for "言語1" (Language 1) set to "JA", "言語2", and "言語3". Below these are sections for "原文標題" (Original Title), "英文標題" (English Title) with the text "Residential Photovoltaic Power Generating System Connected to the Utility Line.", and "英文著者抄録" (English Author Abstract) with a detailed paragraph about a PV system in Japan. At the bottom, there are controls for "入力モード" (Input Mode) set to "指定なし" (None) and "表示フォント" (Display Font) set to "14".

図3 書誌事項データ入力画面（英文標題・同所抄録の貼り込み例）

図4(2) 英文 OCR 処理例（未校正）

UDC 621.311 : 621.383.4

地球環境特集

記事
番号 0002

論文

逆潮流あり個人住宅用太陽光発電システム

Residential Photovoltaic Power Generating System Connected to the Utility Line

Abstract

We are the first in Japan to develop and install a utility interconnect- ed photovoltaic (PV) power generating system with reverse power flow capability in an occupied private residence. The system consists of a PV array, a 3-kW inverter and a protection device. The inverter maintains high efficiency over a broad range, from high to low insolation, and is the smallest and lightest of its kind in Japan. We also report on the composition of the system and comment on future systems, according to the Technical Guidelines for Utility Intercon- nection Operation issued by Japan's Ministry of International Trade and Industry.

牧野康弘* Yasuhiro Makino
井家健仁* Takehito Ike
石田健雄** Takeo Ishida
萩原龍蔵** Ryuzo Hagiwara
大川克実*** Katsumi Okawa

[キーワード]

太陽光発電, 太陽電池, システム連系, 分散電源, 逆潮流, インバータ

1. 緒言

長期的なエネルギーの安定供給を実現するには、既存の電力系統と自然エネルギー源による分散電源とを併用した電力形態の実用化が不可欠である。

近年、このような電力形態は、地球環境問題の解決策として、また夏場におけるピーク負荷平準化といった社会的要請により大いに期待されている。

このような背景の中で、特に太陽光発電システムは、太陽電池の製造コストの低減や出力容量によるシステム構成上の差が少なく、標準化が容易であること、また、資源の地域偏在性や設置場所の制約が少ないことから、実用化に最も近いシステムであると考えられる。一方、社会に定着させるための法的整備も急速に進み、1992年4月の電力10社による余剰電力買電制度の実施、1993年3月の逆潮流あり系統連系技術要件ガイドラインの策定などが実現している。

我々は、1981年から個人住宅用太陽光発電システムの実証実験¹⁾を開始し、また1986年から六甲アイランドにおける系統連系技術関係の実証試験用に、逆潮流あり太陽光発電システムの研究開発並びに実用化モデルの開発を行ってきた。本稿では、これらの成果を基盤にして、

我々が1992年7月に、日本で初めて個人住宅に設置した逆潮流あり太陽光発電システムについて述べ、実データを基に、システムの評価と今後の取り組みを報告する。

2. システムの概要

2.1 全体構成

本システムの構成は、図1に示すように、屋根とベランダに設置された太陽電池アレイが、3 kW 連系インバータ(2系列目の連系保護機能内蔵)と系統連系保護装置(1系列目)を介して、既設の主漏電遮断器及び売買用電力量計に接続されている²⁾。

システムの仕様を表1に、単線結線図を図2に示す。本システムは、施工時に、関西電力㈱から提示された「太陽光、風力、燃料電池等の自家発電設備が一般配電線へ逆潮流する場合の技術要件の暫定取扱」(以下、暫定ガイドラインと記す)に準拠した構成となっている。

絶縁トランスは、直流側と交流側との絶縁が必要であることと、受電連系点の電気方式と同一の方式(単相3線式)にしなければならないことなどの理由から設置されている。

また、電力会社の作業者の安全を確保するため、屋外開閉器を設置している。

*研究開発本部 メカトロニクス研究所
Mechatronics Research Center
**研究開発本部 ニューマテリアル研究所
New Materials Research Center
***半導体事業本部 厚膜事業部
Thick Film IC Division

英文抄録(700字)の場合:約45~50秒

和文抄録(400字)の場合:約70~90秒

(注)上記 OCR 作業時間は、読み取り開始から文字変換終了までの機械処理時間である。

②従来のパンチ方式による作業時間は、

英文抄録(700字)の場合:約190~210秒

和文抄録(400字)の場合:約200~240秒

単純な比較では、OCR 方式の方が大幅に短い作業時間で済む。ただし、ここであげた数字は、OCR ソフトでの辞書によるスペルチェックやパソコン画面上での校正などの後作業の時間は含んでいないし、従来方式の方もペリファイに要する時間を含んでいない。

なお、参考までに OCR 方式による 1 記事の作業時間(20分)の内訳を、図 5 に示す。これを見ると、校正時間が全体の 5 割強を占めていることが分かる。

6. OCR 入力における留意点

OCR 方式が実用になるか否かは、読み取り精度、すなわち、認識率が高いかどうかにかかって

図 5 1 記事の作業時間内訳

(1 記事 = 20分)

いる。そして、その認識率を左右する要因は、イメージスキャナーの種類、OCR ソフトの種類、対象資料(原稿)の印字品質・紙質など、いくつかあるが、ここでは前記 3. に述べたシステム構成で得たポイントを以下に示す。

(1) 資料の印字品質

極端に薄い文字やかすれた文字は、例えば、人間の目では十分読めても、ほとんど読み取ってはくれない。また、字が途中で切れているもの、例えば、「d」の中央の一部が切れていると「c」と「l」に分けて読み取ってしまう。その他の誤読例を参考までに、表 1 に示す。

(2) 印刷物の紙質

通常の印刷用紙であれば、紙質はそれほど問題にならないが、その中でも紙面が平滑な、コート紙や微塗工紙などの塗工紙は認識率が非常に高い。非塗工紙でも、上質紙は OCR 入力に適している。これらに比べ、紙面がざらついているものは、認識率が低下するようである。また、読み取り面の裏や重ねてある紙の画像が裏映りする薄い紙のものも、認識率はよくない。

(3) OCR ソフトの設定

原稿のサイズ、読み取り方向、段組み、スキャン濃度などの作業条件は、資料毎に最初に 1 回だけ設定するだけでよいが、これらのうち、スキャン濃度は認識率に大きく関係する。認識率が悪い場合は、濃度設定を調整する必要がある。同じ原稿で、濃度を変化させた時の認識率の変化の実測例を、図 6 に示す。この例から、スキャン濃度設定が認識率に対して重要な要素を占めることがわかる。なお、この実測例で使用した資料のデータは次に通りである。

資料(A): 上質紙, 文字サイズ=11級

資料(B): コート紙, 文字サイズ=14級

資料(C): コート紙, 文字サイズ=12級

資料(D): 上質紙, 文字サイズ=12級

7. おわりに (今後の課題)

電算システムへのデータ入力、これまで専門

表1 OCRによる誤読の例

英 文		
項番	原稿の文字	読み取り後の文字
1	I	1 (アラビア数字)
2	Il	11
3	in	m
4	it	U
5	n	i i
6	±	+
7	×	x (英小文字)
8	℃	° C
9	%	。 7 0
10	0 (アラビア数字)	O (英大文字)
11	5	S

和 文		
項番	原稿の文字	読み取り後の文字
1	— (音引き)	— (漢数字)
2	- (ハイフン)	— (音引き)
3	+ (加算記号)	十 (漢数字)
4	1 (アラビア数字)	l (英小文字)
5	カ (カタカナ)	力 (チカラ)
6	ば	ぱ
7	ブ	昔
8	音	昔
9	果	果
10	管	菅
11	間	問
12	示	ホ
13	自	白
14	晶	品
15	層	屑
16	入	人
17	率	卒
18	力 (チカラ)	カ (カタカナ)

家によるキーボード入力主流であった。しかし、今後は、データ入力業務の分野に限らず、専門家だけに依存するデータ処理の時代から、一般の人間も参加できる時代に移行しつつあると言える。現在のオフィスコンピュータやパソコンの普及度合、また、情報処理技術や情報化機器の発展・革新を勘案すると、この傾向に一層拍車がかかることは間違いないと思われる。

本報告で述べたOCR方式は、現時点では、印刷物の印字品質がある一定水準以上であれば有用という前提条件がつく。印字品質が悪い場合には、

(1) 和文 OCR

(2) 英文 OCR

図6 認識精度グラフ

認識率が極端に低下するため、OCR方式は全く実用にならない。このような実用的な限界を踏まえた上で、OCR技術を利用すれば、印刷媒体情報の電子ファイル化、さらにデータベース化において、一般の人が作業するものとして考えた場合には、非常に優れた入力方式といってよい。

作業効率の観点からは、OCRで読み取った文字ないし単語を、元の印刷物と照合する校正作業でかなりの時間が必要であり、これをいかにして短縮するかが今後の課題である。

さらに、現状ではキーボード入力の方が効率が良い、短いデータ項目(著者名、そのフリガナなど)についても、システムをさらに発展させて、ワンタッチで入力できるシステムの構築を実現したいと考えている。