解　説

文字および記号読取によるデータ入力*

森　　健　一**

1. はじめに

文字は人間同志の情報の交換や記録の媒体として，人間社会の中で重要な役割をはたしてきた．電子計算機の利用が進むにつれて，人間同志の情報の媒体としている文字を機械に人間と同様に認識させ，電子計算機が取扱えるコード化された情報に変換させる装置―文字認識装置―の研究開発が行われ，実用に供されるようになってきた．

文字認識装置の研究は電子計算機の出現より古く，1930 年代には早くも電信文をモールス符号に変換する特許が米国で出願されている．それ以来，大学，企業で精力的な研究開発が行われ，記号読取装置，活字読取装置，手書文字読取装置が次々に市販されるようになってきた．米国では 1960 年代より，米国銀行協会を中心にして，小切手の下部に磁気インクで文字コードを印刷し，これを読取る磁気式記号読取装置が普及し，機械と人間が同じ情報媒体を共用する最初の社会的なシステムとなった．我国では昭和 42 年より始まった郵便番号制度において，手書きの郵便番号を読取る自由手書郵便番号自動読取区分機が導入され，多くの郵便局で使用されるようになり，文字読取装置の有用性が国民に知られる大きな契機となった．

現在では米国において活字文字読取装置が多くの企業でソースデータの入力，特に伝票の入力に用いられている．我国では伝票の多くが手書きされているため，最近の手書き英数字カナ文字読取装置の発達を機に，普及の速度が早まろうとしている．

記号や文字の読取装置が人間と機械とのインタフェースとして用いられる場合を分類してみると，次の 4 つの場合がある[1]．

（1）　インプット

不特定多数の人が手書きした伝票や，タイプライタその他の機器を用いて印字した文字・記号情報（原始データ）を入力する場合（図-1（a））

（2）　ターンアラウンド

機械が処理した結果をプリンタで印字出力した伝票や書類などの文字・記号情報（2次データ）を，人間に渡し社会的な流通を経た後，再び機械内の情報に戻す場合（図-1（b））

（3）　ファイル

人間側にすでに文字の形で大量に蓄積されている

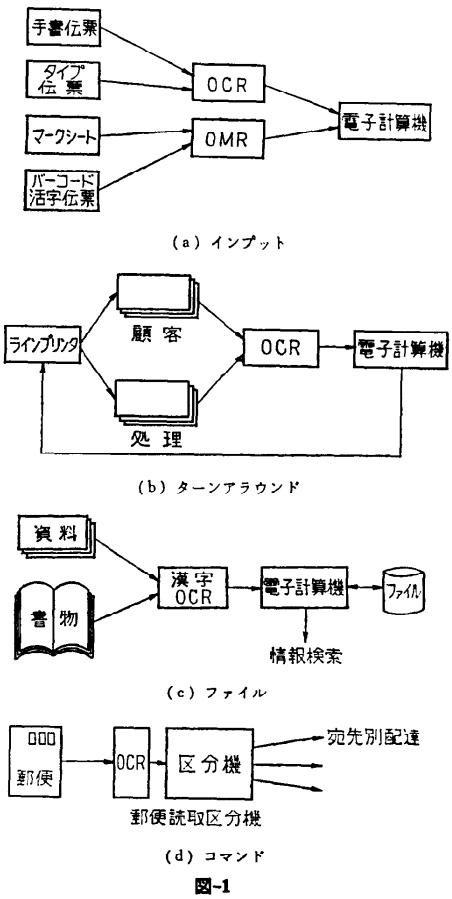

図-1

* Data Input by Optical Character Recognition and Optical Mark Reader by Kenichi MORI (Toshiba Research and Development Center).

** 東京芝浦電気(株)総合研究所

資料や書物などの情報をコード情報に変換して，ファイルデータとし高度な利用を図る場合（図-1(c)）

（4） コマンド

機械への命令・指示を文字や記号などで書いて与える場合（図-1（d））

インプットの場合，不特定多数の人が手書きした文字には，無限に多数の変形文字が出現するし，機械印字する場合でも，タイプライタや印字機が異なれば，字体や印字品質が千差万別に変化する．このように標準化されていない原始データの入力は不可能と考えられていた．しかしながら文字認識技術の進歩と同時に標準化の気運も進み，活字については OCR-A 字体，OCR-B 字体よりなる英数字記号と，OCR-K のカナ文字が JIS 化された[2]．現在，ほとんどの計算機メーカはこれらの字体を標準字体として採用している．さらに，手書き文字についてもカナ英数字記号の標準字体が JIS 化されようとしている[3]．このような標準字体が普及するにつれて，「インプット」としての文字読取装置の正確度も向上し，多くの伝票，プログラムなどがそのまま計算機に入力するようになってくるであろう．

ターンアラウンドの場合には，文字を印刷するのが電子計算機システムのラインプリンタであることが多く，使用する字体，印字品質，用紙，リボンなどを，十分に管理することができる．したがって汚れが少なく，単一の字体で鮮明に印字した伝票を作成することができる．従来，文字読取装置が利用されていた形態の多くは，このターンアラウンドの範囲に属するものである．

たとえば，電力検針による電気料金徴集業務では，あらかじめ計算機により各家庭のコード番号を印字した検針伝票上に，検針員が電力計の指示値をマークで記入し，文字読取装置でコード番号とマークを読取る．この場合，コード番号は「ターンアラウンド」であり，マークは原始データの「インプット」になっている．電子計算機で前月までの消費量と当月の指示値との差から当月の電力使用料金を計算し，請求書および払込伝票を印字する．これらの伝票は各家庭に送られ，料金の支払と同時に払込伝票を切離し（2次データ），計算機室にターンアラウンドされ文字読取装置による入力と請求内容の消込みが行われる．

第3の「ファイル」は，長い期間かけて大量に蓄積されている文字情報，たとえば特許公報，判例集，顧客カード，文献カード，書籍，切抜き資料などが入力対象となる場合である．これらの対象の文字情報は，文字読取装置で入力することを考えて印刷されているわけではないので，字体も対象により異なり，書式も一定しない．文字読取装置としては，漢字，片仮名，平仮名，英数字記号など多種類の文字を読取れなければならないし，ページの中から図面や写真，数式などを除外する機能も持たねばならない．このような意味でこのタイプの文字認識装置の実現には飛躍的な技術進歩がなければ不可能である．後述するように最近になってこの技術難関が突破され，今後はこのような大量の文字情報を計算機に入力することができるようになりデータベースを利用した新しいサービスが発達するものと思われる．

最後の「コマンド」は，機械への命令を文字で与えることを意味している．前述した日本の郵便番号読取区分機はその典型例で，郵便物上に記入された郵便番号がコマンド情報となり，その郵便物自身をどの区分箱に入れるかの命令文になっている．その他，電話機にタブレットを接続しておき，住所や氏名，品物の名前を手書きすると，買物の注文や電報の発信ができるシステムなどが将来装置として考えられている．コーディングシートに書かれたプログラムを文字読取装置によって入力し実行させることも「コマンド」入力に属している．

文字読取装置を導入して事務の省力化や機械化を図ろうとする場合のシステム設計法として3つの方法がある．第一の方法は，現行の業務形態や伝票の流れ，伝票の書式，記入様式などをすべて保存し，性能の高い装置がなんとか現行システムをそのままカバーすることを要求する方法である．第二の方法は，機械本位に徹して新しい事務システムを設計する方法である．両者の方法は，人間か機械かいずれかの側に過大な負担がかかる．手書き伝票を対象とするときは第一の方法を要求する人が多い．ターンアラウンドを用いた初期の文字読取装置の利用では第二の方法に近い設計法が用いられた．しかし成功した多くのシステム設計では第三の方法，すなわち現行システムを尊重しつつ，機械にとってキーポイントになる最小限の標準化を行い，機械の特性に合わせ，人間にも機械にもうまくゆく妥協点（トレードオフ）を見つけてゆく方法がとられている．日本では文字読取装置を導入すると，伝票の100%を機械処理することを考えるが，米国では読めるものを高速に文字読取装置によって処理し，印字品質の悪いものや，標準外の伝票は人手で処理するよ

うなマン・マシン併用システムで成功したものが多い．

本稿では活字 OCR，手書 OCR，OMR の最近の技術動向を解説するとともに，今後の発展方向について考察してみた．

2. 活字データの入力

2.1 活字文字読取装置によるデータ入力

1章で述べたように，英数字カナ文字の字体については JIS で標準化され，この字体を装備したラインプリンタ，タイプライタが普及している現在，ターンアラウンド入力に活字文字読取装置を用いることは，ほとんど問題がない．装置のコストも 1960 年代の3千万円～1億円のものが，700 万円～2千万円と5分の1近くに下がってきている．さらに後述するハンド OCR では 50～100 万円の簡易型の装置も市販されるようになってきた．

これらの光学式文字読取装置 (OCR) には，最近の LSI 技術が全面的に採用されている．OCR の目玉に当る走査装置には，ホトダイオードアレイや CCD アレイが用いられ長寿命化と低コスト化がはかられている．論理回路においても LSI の採用により，相当複雑な認識方式でもコンパクトな装置にまとめることができるようになり，OCR の精度も著しく向上した．

一方では，OCR 技術の中であまり発達していない部分もいくつかある．用紙を供給する機構部や，文字行の位置や1行中の1文字分のパターンを検出する前処理部などがその例である．これらの技術は 1960 年代と本質的には変わっていないため，OCR の性能がこれらの技術によって制限を受けるようになってきている．

用紙の問題では従来の OCR は，OCR 用紙の使用を前提にしていた．以前は用紙中に含まれるゴミや油滴が OCR の正読率に影響をもっていたためこれが少ない OCR 用紙の使用を必要としていた．しかしながら，OCR の認識部が発達した今では少しぐらいのゴミは全く問題でなくなってきた．現在，上質紙ではなく OCR 用紙の使用を条件としている OCR は，機構部の用紙搬送上から生ずる問題が主たる理由になってきている．OCR 用紙はパルプの含有量が多く腰が強いので，用紙の正確な搬送が楽にできるからである．機構部の技術の発達は遅々たるものであるが，ようやく最近になって，LP 用紙程度の薄い上質紙でも正確に搬送できる技術開発が行われるようになってきた．

文字パターンを検出する前処理部も昔からほとんど同じような手法が用いられており，同じような問題点をかかえている．タイプライタ等では印字機構のガタのために隣り同志の文字が接触した状態で印字されることがある．またラインプリンタでは上下の文字行や左右の文字の印字の際にゴースト（印字していない文字の活字パターンの一部が，影のように印字される現象）が発生することが多い．これらの雑音が加わった印字パターンから，1文字分のパターンを検出する方法についてはいくつかの研究はあるが[4),5)]，その論理が複雑すぎて一般的でないため，一般の OCR ではもっと簡易型の文字パターン分離方法が採用されているのが現状である[6)]．文字パターンの分離を必要としない文字認識方法の開発，認識部より前処理部へ認識結果をフィードバックすることにより文字パターン分離の正確度を高める方法などの多くの方法が提案されている状態である．

これらのいくつかの問題点をかかえてはいるが，活字 OCR は完全に計算機入力装置として安定した地位を占めるようになってきた．最近では「ターンアラウンド」ばかりでなく，データの発生した時点で計算機入力するために端末 OCR も発達し，「インプット」にも用いられるようになってきた．文字情報の良さは，人間と機械とが同じ情報媒体を共用できる点にあるので一度，安定な OCR が出現すればその適用範囲は非常に広いものになる．端末 OCR ではゴム印で押した文字や，ドットプリンタで印字した文字であっても読取れるものが市販されている．

2.2 ハンド OCR

最近，市販されて話題になっている OCR に，ハンド OCR (図-2) がある．これは活字 OCR のコストを極端に下げて行ったときの一つの形態で，OCR から機構部を取除き，スキャナと，認識部だけで構成さ

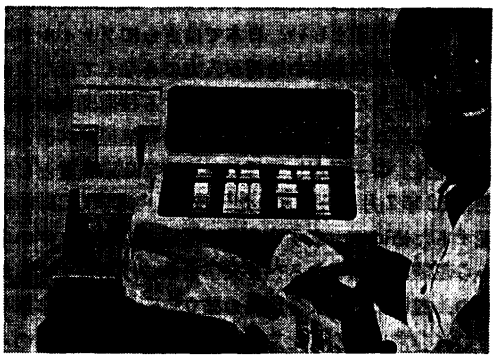

図-2 ハンド OCR

れたOCRである．ヘアドライヤ形のスキャナ部を手で持って，文字行に沿って走査すると，文字情報の入力を行うことができる．読取る文字の種類は数字を中心にいくつかの英字記号よりなり，25～30字のものが多い．

用途は主としてPOS (Point of Sales) 用であり百貨店などの小売業で売上品の値札の読取入力に用いられている．従来売場での管理は金銭的なものだけで売上げた品物の管理は別途行われていたが，ハンドOCRを用いれば両者を同時に行うことができる点で優れている．人手で走査するため，缶詰や衣料品のような曲面上に印刷されている文字や，透明な包装紙の中に入った値札でも読取れるなどの融通性があることも特徴である．スキャナは文字行に対して ±8° 程度の傾き，文字面からの垂直度が ±10° 程度の傾き，文字の高さに対し2倍程度の位置ずれ，および走査速度の変動幅が5倍以内程度を許容するように設計されている．また文字行の前から走査しても，後から走査してもよい．文字行の最初の文字が1行中の桁数と，文字行の役割(金銭行，品物番号行などの区別)を同時にあらわしている．1行中の全ての文字がリジェクト (OCRが読めない文字パターンが入力されたとして読取拒否すること) されずに読取ることができ，桁数が第1桁目に指定された数と一致するときに，ランプやブザーで入力が完了したことを知らせるようになっている．

OCRの読取精度は99.9%以上で，リジェクトが生じたときは，再度文字行を走査すればよい．誰でもちょっと練習すれば正確に入力できるようになる点が普及を早めている原因でもあろう．

2.3 漢字 OCR

「ファイル」入力のためには，標準化されていない文字や文書であっても読取れなくてはならない．また，大量なデータ入力を必要とするため，読取速度も高速であることが望ましい．日本ではさらにファイルデータのためには日本語の情報が入力できなくてはならない．このような困難な条件を克服する技術開発が通産省の大型プロジェクト「パターン情報処理システムの研究開発」の一環として行われ，漢字認識装置として昭和52年7月に公開された[7] (図-3)．漢字OCRを実現するためには，まず2,000字種以上の対象文字の中から正確に候補となるべき文字をみつけ出してくるための，機械による大分類法の確立[8]と，複雑な字形や互いに非常に類似した文字を正確に識別するためのパターン認識理論の確立[9]が必要であった．

図-3 漢字 OCR

我々が字典で漢字を検索する場合，音訓や部首を用いることによって，能率よく目的とする文字をみつけ出すように，機械が2,000字種の文字を端から調べるのではなく，文字パターンの性質を用いて候補文字の範囲を限定するのが前者の大分類法である．図-3の漢字OCRでは，漢字パターンの縦，横の線密度にもとづく複雑指数と漢字パターンの四周辺部の文字線量をコード化した四辺コードを併用することにより，雑音にもつよく効率のよい大分類法を実現している．入力漢字パターンに対しての候補文字数が100字以下になれば，英数字カナ文字認識の場合と同程度になるが，漢字認識ではきわめて類似した文字組が多くなっているため，英数字認識程度の能力をもつ認識原理ではきわめて鮮明に印刷されている特別の場合を除いては，通常の印刷物，例えば特許公報や新聞を読取らせると満足な認識率が得られない．通常の印刷物はOCRのために特別に印字品質に考慮をはらっているわけではないので，種々の雑音を含んでいる．このため一種類の邦文タイプライタで印字した漢字パターンをシミュレーション実験で仮りに正確に読めたとしても，通常の印刷物をもってくると手も足も出ないということが起る．ファイル入力が可能となるか否かは，この困難さを克服することにある．

この目的で開発されたパターン認識の理論が混合類似度法[9]で，漢字パターンに含まれている雑音をモード関数によって直交展開し，主要成分を抽出するとともに，これらの雑音に不変な類似度関数を与えている．さらに，王一玉，間一閒のように互によく類似している文字組がある場合に，その差分パターンによって類似文字間の類似度値の差異が強調されるような理論体系を与えている (図-4) (次頁参照)．混合類似度法は計算機を用いて見通しよく漢字OCRを設計するために用いられており，種々のハードウェア上の工夫により

コンパクトな形で漢字 OCR が実現され，高い識別能力が実証された．漢字 OCR の認識性能は，人間のオペレータが邦文タイプや漢テレ鍵盤を打鍵する場合よりずっと高く，速度は 100 倍以上早い．識別文字数／1字当りの認識時間で評価すると現存する OCR の中でも最高速度のものに属している．

3. 手書文字データの入力

3.1 手書 OCR によるデータ入力

日本では伝票の大部分が手書きで記入されることが多いため，タイプライタを多用する米国に比較して OCR の利用が遅れている．それでも最近の手書 OLR の発達によって導入の気運が高まってきており，電算メーカの各社からカナ英数字の読取のできる端末型 OCR の発表が相ついでいる．

手書文字認識の問題は主として日本で研究が進み，前述したように全国民の書いた自由手書きの郵便番号を毎日処理しているのは世界でも日本だけである．**図-5** は JIS 原案として検討が進められている手書きカナ英数字の標準字体案である[3]．手書文字認識の最大の困難さは，ある人の書いた A の字と，他の人が書いた B の文字が，書いた本人は異なるカテゴリーの文字の字として書いたつもりでも全くよく似ている場合にどうするかという問題である．5—S，8—B，7—クーワ，0—D などの文字パターンの間では連続的に字形を変化させることができる．まぎらわしい文字を全てリジェクトにすると認識率は 95% 程度になってしまう．それほど人の書く文字は1字ずつみると相当あいまいな文字パターンであることになる．一般的に我々が手書文字をみて，99% 以上の正確さで判読できるのは，人間が1文字ずつのパターンをみて判断しているのではなく，前後関係から色々の情報を用いて判断しているからである．

標準字体を判定することによって記入者間の曖昧さを減少させた場合の手書文字を，常用手書き文字と呼んでいる．これは英語で Handwritten と Handprinted が区別されているように楷書体で書くことに相当している．現在市販されている端末型の手書 OCR はいずれも常用手書文字を対象にしている．

手書文字を記入する人が限られた範囲内の人であれば，標準字体を守るのも容易であるが，不特定多数の人が記入する場合にはその徹底は仲々難しい問題である．さらに経理関係の人は独特の数字を書くことにより，同じ職業の人同志の間では数字の読み誤りを防止する努力をしている．このような点から手書 OCR は，標準字体の他に簿記書体や小学校の教科書の書体など，多くの書体を判読できる機能をもたなくてはならない．

他の問題として活字 OCR の項でも述べたように，現在の OCR の弱点として前処理部があり，隣り同志の文字がつながっているとその分離に失敗し，リジェクトを発生することがある．したがって，帳票設計の

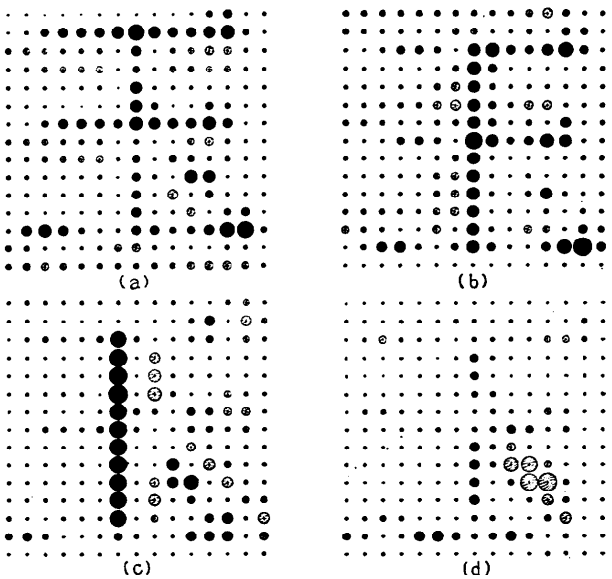

図-4 「王」のための混合類似度法による辞書パターン

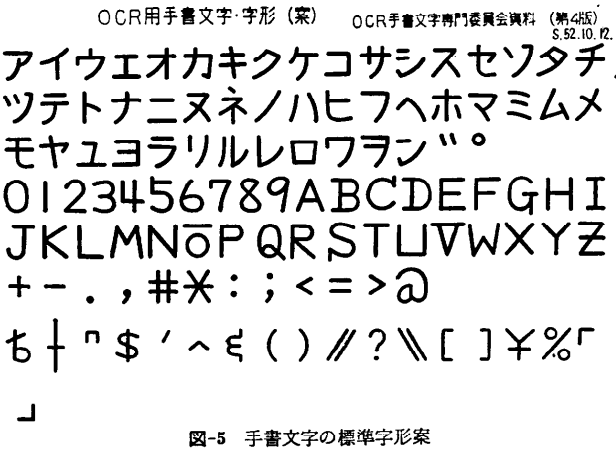

図-5 手書文字の標準字形案

際に記入者が無意識のうちに1文字ずつ分離して記入するように，記入枠の設計を行うのが普通である．記入者が標準字体で記入することに慣れる以上に，1文字ずつを枠からはみ出さないように注意することの方が認識率を大幅に向上させることに役立つ場合すらある．手書文字の文字ピッチは数字で 25.4 mm 当り5字，英字，カナ文字の場合で4字程度が最も妥当であるという調査結果が出ている．

手書 OCR を選定する場合に，筆記用具などのような制限がついているかも重要な点である．伝票は多くの場合，改竄を防ぐためにボールペンを用いるのが普通であるが，手書 OCR の中にはボールペンで書かれた文字を全く読取れない OCR がある．これは OCR のスキャナにどのような光電変換素子を用いているかによっている．最近の OCR は半導体系の素子を用いているものが多いが，これは赤外線領域での感度が高く，一方ボールペンのインクは赤外線領域では光を吸収しない．このため文字と紙とのコントラストがつきにくくなるという問題が生ずる．フィルタ，光電変換回路を工夫することによりこの問題を解決し，ボールペン文字でも安定に読取ることができる手書 OCR も市販されるようになってきた．

3.2 端末型手書 OCR

これまで3,000万円以上した手書 OCR が活字 OCR と同様に LSI 技術の利用，マイクロコンピュータの採用などによって安価になってきた．手書文字による伝票を計算機室に集め集中処理するシステムに代って，計算機への入力を分散化することにより，ピークロードの減少やデータ入力の時間的遅れを少なくすることが可能になってきた．このような端末型 OCR では読取速度は集中処理型 OCR に比較して遅いが，価格は十分に安く，かつ読取精度は集中処理型の OCR と同様程度に高いものである．この意味では OCR の立場からするとコスト性能比の大幅な向上を実現したものである(図-6)．

端末 OCR では対象文字が常用手書のカナ文字，英数字，記号である．対象文字数が100字近くになるので，漢字 OCR のように大分類法を用いているものが多い．すなわち，入力の手書文字パターンを直接識別するのではなく，文字パターンに含まれる端点，交叉点，ループの数や位置を調べて候補文字を限定する方法や文字全体の形状から候補文字を5文字以内に限定する方法などが考えられている[6]．手書文字ではラーヲ，アーマのように文字線がついているか，突出しているかあるいは文字線の傾き程度など，局所的な部分パターンの差異により異なる文字になるので，候補文字を絞ってから精密に部分パターンを調べるという方法が，認識の速度向上，メモリ量の減少から必要になってくる．

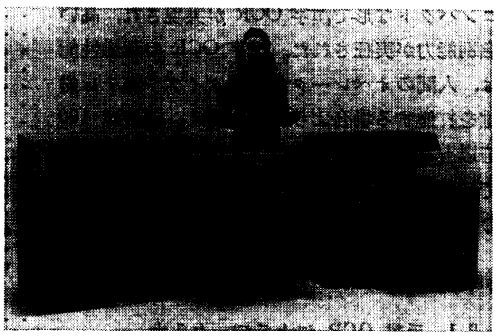

図-6 端末型手書 OCR

手書文字の書き方の制限を強めて，特定のルールに従った文字のみを読取対象とする手書 OCR は，制限手書 OCR と呼ばれ常用手書 OCR と一応区別されている．その区別は相対的なもので明確ではないが，常用手書 OCR に制限付の手書文字を読ませれば，正読率は一段と向上するのは当然である．ユーザとしてはシステム設計に際して記入者の曖昧さを減少させるためにどの程度の手書文字を書くことができるかという点と，導入する OCR の性能とのトレードオフをはからなければならない．制限手書 OCR は常用手書 OCR に較べて価格は一段と安いのが普通である．

4. マークによるデータの入力

文字の代りに棒状のマークによりデータ入力する方法は OCR の発達以前から用いられており，現在では POS，漢字データ入力にまで用いられている．最も古くは磁気インク文字読取(MICR) のために開発された E 13 B 字体，CMC-7 字体などは見かけ上は文字のような体裁をもっているが，実はコード化されたマーク読取の原理を用いている．

マーク読取の最も多い応用例は手書きされたマークの位置を検出することにより，データ入力をする方法で，国勢調査や前述した電力検針，答案採点などに数多く用いられている．この方法は簡便であることが最大の長所であり，読取装置の価格も安い．反面，データ記入のミス率が高く，1枚の記入シート当りに書くことのできる情報量が文字に比較して極端に悪い点が欠点となっている．

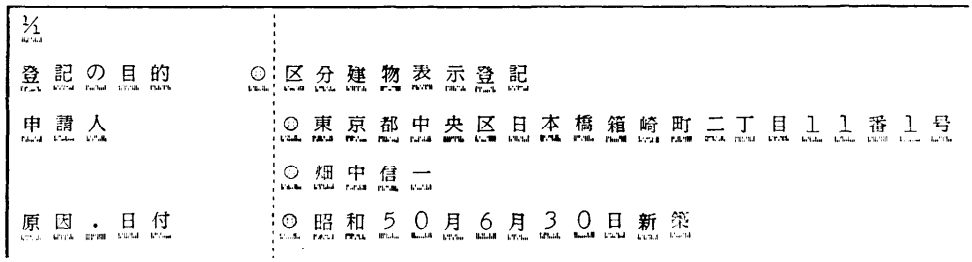

図-7　バーコード付漢字タイプライタ文字

米国では小売業の機械化のために品物コードをバーコードの形で印刷し，各個品毎に貼付したり，包装紙に印刷しておき，レーザスキャナ，ビジコン等を用いてマーク読取りを行う POS が普及しはじめている．スーパーマーケット等短時間に大量の買物客を処理したい業種に向いたシステムである．しかしながら，マークでは人間に読めないこと，情報密度が悪いため多くの桁数を入れることが困難である点などの欠点があるため，前述した活字のハンド OCR が開発され普及が始まっている．

マークリーダを漢字データの入力に用いることも考えられ，図-7 のように邦文タイプライタの活字の下方にバーコードを付与し，人間は上の活字を読み，機械は下のマーク読取を行う装置が開発されている．バーコードは3値の 11 本のマークで構成されているので，1万字種以上の漢字，仮名，記号をあらわすことができる．またコード自体にエラーチェック機能が含まれているので，バーコード読取装置の性能は 10 万字に1字以下の誤読率で，非常に安定した読取性能をもっている．

5. OCR によるデータ入力の今後の展望

手書 OCR の発達により，カナ文字，英数字のデータの「インプット」は今後5年間は増加の一途をたどることが予想される．それはキーパンチャの人手不足，コスト高の傾向は今後も続くと見られ，一方，手書 OCR は LSI 技術の発達，特にメモリ素子，マイクロコンピュータの発達により多種多様の構成の機器が開発され，標準字体の JIS 化と普及にともなって安定したデータ入力方法となることが見込まれるからである．

さらに漢字 OCR の出現によって「ファイル」入力のための道も開け，新しい計算機応用が進むものと考えられる．オンライン手書 OCR の開発，ファクシミルの発達によって，「コマンド」入力も新しい展望が開けよう．OCR の今後の技術開発としては，手書漢字 OCR の研究開発が残された問題である．さらに，手書英数字カナ文字の自由度を拡大する努力は続けられてゆくものと思われる．

参考文献

1) 森：文字認識装置，電子通信学会誌，56 巻 11 号，pp. 1524～1529 (1973).
2) JIS C-62550, C-6252.
3) 手書 OCR に関する調査研究報告書，日本電子工業振興協会 (1977).
4) M. B. Bartz: The IBM 1975 Optical Page Reader, Part II, IBM J. Res. Dev. Vol. **12**, No. 9, pp. 364～371 (1968).
5) R.L. Hoffman et al.: Segmentation Methods for Recognition of Machine printed. Characters, IBM J. Res. Dev. Vol. **15**, No. 3, pp. 153～165 (1971).
6) 安田：文字読取装置はどのように構成されているか，日経エレクトロニクス，No. 100, pp. 71～100 (1975).
7) 森，坂井：日本語文書を読み取る印刷漢字認識装置，大型プロジェクト，パターン情報処理システム講演会論文集，pp. 33～44 (1977).
8) 坂井，森：漢字パターンの大分類，電子通信学会研究会資料，PRL 73-14 (1973).
9) 飯島：混合類似度による識別理論，電子通信学会研究会資料 PRL 74-24 (1974).

(昭和52年12月10日受付)