



入江文平 [bunpei.irie@toshiba.co.jp](mailto:bunpei.irie@toshiba.co.jp)  
(株) 東芝 e-ソリューション社

## まえがき

世界初の光学的文字読み取り装置 (OCR: Optical Character Reader) は米国で1956年に発表された。日本では電子技術総合研究所 (現、産業技術総合研究所) で1959年に初めて試作された。それ以来精力的な研究が重ねられた結果、今日では文字認識技術は帳票用OCRや郵便区分機、PDAをはじめとする各種機器の文字入力インタフェース等さまざまな形で製品化されており、パターン認識技術としては最も実用化が進んでいる分野の1つである。

本稿ではこれから文字認識の研究を始めようと思っている読者、あるいはその概要を知りたいと考えている読者を対象としてこの技術について簡単な紹介を行う。

## 文字認識技術の分類

文字認識は筆記中の筆跡の時系列情報を用いるか、書かれた後にできた静的情報のみを用いるかによって区別される。前者をオンライン文字認識、後者をオフライン文字認識と呼ぶ (ただし後者の呼び方はあまり一般的ではなく、前者と対比する場合以外にはあまり用いられない)。

オフライン文字認識は画像入力デバイスによってOCRの他に磁気インクを用いる磁気インク式文字読み取り装置MICR (Magnetic Ink Character Reader) もあるが、通常はオフライン文字認識=OCRと考えてよく、本稿でもその

ように扱う。

オンライン文字認識の場合、入力情報はサンプリングして取得された筆記点の座標列とペンアップ/ダウン情報である (デバイスによっては筆圧やペンの傾き等が利用可能な場合もある)。オンライン文字認識は時系列情報を使えるため入力と辞書のストローク (画: かく) をマッチングさせる手法が有効であるが、実際には続け書きや筆順違いがあるためそう簡単ではなくOCR的な手法を併用するケースが多い。以下、本稿ではOCRを主に扱う。

## 認識対象の種類

認識対象セットはいくつかの観点から分類され、その違いは認識手法にも関連する。認識の難しさを不等号で表して (難しい方が大) 箇条書で示すと、

- 文字種：数字 < 英字 < カナ < 漢字
- 生成方法：印刷活字 < 手書き
- 手書き字体：楷書 < 行書 < 草書  
制限手書き < 常用手書き < 自由手書き<sup>☆1</sup>
- 手書き英字：ブロック体 < 続け字筆記体<sup>☆2</sup>

となる。不等号で表した難しさの順は一般論であり逆転する場合もある。また、通常は認識問題として簡単なほど要求される認識性能は高くなる。

☆1 制限手書きとは、カテゴリごとに特定の筆記方法を指定する場合で、常用手書きはある程度の制限を設ける場合、自由手書きは制限がない場合である。

☆2 ブロック体と続け字筆記体は混在して書かれることが多い。

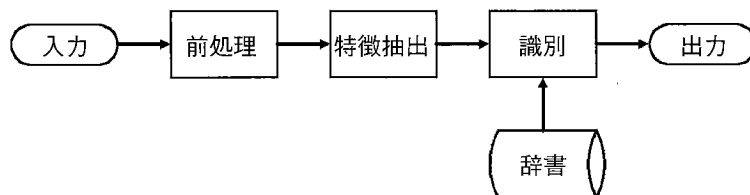


図-1 文字認識の構成

## 認識性能

パターン認識の性能を表すには認識率，すなわち与えられた入力のうち正しい認識結果が出せる割合を用いる。学会論文等で単に認識率R%と発表している場合，リジェクトをしないケースが多い。この場合残り $e = 100 - R\%$ は誤読となる。しかし，実用上は認識結果のうちどれが正しくどれが誤っているか不明では困る場合が多い。そこでむやみに認識結果を出すのではなく，その信頼度が低い場合はリジェクトすることが多い。この場合でも，リジェクトしなかったものの中に誤読が混じることは避けられず，正読率C%，リジェクト率r%，誤読率e%， $C + r + e = 100$ となる。しかし，誤読率はリジェクトのロジックを入れなかった場合に比べてはるかに低くすることができる。rとeの比率はリジェクトのパラメータで制御可能な場合が多く，目的に応じて適正な値を設定する。認識率（正読率）の数字は当然認識対象の文字セットによって大きく変動する。認識屋をやっていると「認識率はどれぐらいですか？」という質問を受ける場合がよくあるが，リジェクトの有無，対象セット（文字品質がどれぐらいか）等の条件を明確にしないと返答できない場合が多い。

## 文字認識のしくみ

文字認識の一般的な構成を図-1に示す。光学系等から入力された文字画像はまず前処理を受け，特徴抽出処理によって特徴ベクトルへと変換され識別部へ送られる。識別部では認識辞書を参照しながら識別を行い，結果（文字コードまたはその候補列，場合によってそれぞれに対応する評価値，リジェクトコード等）を出力する。認識結果の表現は通常文字コードであるが，1つ1つのコードに対応する文字種をカテゴリと呼ぶ。たとえば数字

のセットは「0」～「9」の10種類のカテゴリからなる。なお，前処理，特徴抽出，識別という区分は便宜的なものであり，たとえばニューラルネット等によって認識方式全体を設計する場合などはその区分はあいまいとなる。

### ○前処理

文字認識で用いられる前処理は光学系の補正，画像ノイズ除去，二値化，正規化等があり，最上流に位置するだけに認識性能への影響度は非常に大きい。正規化のうち代表的なものはサイズ正規化で，まず文字の外接枠を求め，その外接枠を縦横に伸縮させて一定の大きさに変換する処理である。伸縮のさせ方によって線形に縦横を伸縮させるオーソドックスな線形正規化と，画素やストロークの密度等に応じて非線形に伸縮させる非線形正規化とがあり，後者は手書きによる文字変形に対して有効であることが知られている。他には傾き正規化，線幅正規化等がある。

### ○特徴抽出

OCRで扱う入力はビットマップ画像等と同じく格子状の画素値を並べたもので，パターンベクトルとも呼ぶ。文字認識ではこの入力をそのまま，あるいは二値化しただけで認識部の入力として用いる場合もあるが，何らかの特徴量を抽出する処理を行う場合も多い。構造解析系（後述）の認識方式では輪郭抽出や細線化等による幾何学的な特徴抽出を用いる場合が多く，上述の前処理との区別はあいまいとなる。重ね合わせ法（後述）の場合は各種空間フィルタや判別分析，主成分分析等の統計手法によることが多く，この場合は後述の識別部との区別はあいまいとなる。

### ○識別

識別部はいうまでもなく文字認識の主役である。識別方式は重ね合わせ法（テンプレートマッチング）と構造解析法に大別されることが多いが，いずれにもあてはまらないものもある（特にニューラルネット系のもの）。数学的には識別器（classifier）と呼ぶ。

重ね合わせ法のアルゴリズムのうち最も基本的な単純類似度法はあらかじめカテゴリごとに用意されたテンプレート（辞書ベクトル）と入力文字の角度（類似度）を計算して類似度が最大のカテゴリを認識結果とするものである。それを発展させた手法としてテンプレートを単一のベクトルでなく部分空間とする部分空間法や複合類似度法等さまざまな方式がある。重ね合わせ法を用いた識別方式は、識別器としては複数のテンプレートのうち識別対象に最も近いものを識別結果とする最近傍識別器に分類される。後述の構造解析法も最近傍識別器の形態をとる場合が多い。重ね合わせ法の辞書は統計的な手法によって大量のデータから自動で（学習により）作成することが多い。

構造解析法は、芯線や輪郭の形状や位置関係等の幾何学的な性質をもとに字種を判定するタイプの方式である。辞書の作成は一般に自動化が難しく、手作りとなることが多い。構造が複雑で字種の多い漢字の場合、幾何学的な構造のバリエーションが多様であることと辞書作成の手間がかかることから、通常は重ね合わせ法が用いられる。

## 文字を認識するための他の技術

前章では単一の文字に対応する画像が入力として与えられた場合に認識結果カテゴリを求めるための文字認識アルゴリズムを説明したが、文字認識技術を実際に応用してシステムを構成するためには関連するさまざまな技術が必要となる。本章ではそれらの技術について説明する。

### ○大分類

漢字のようにカテゴリ数が多い（JIS第一水準で3,000程度）場合、全カテゴリに対して処理時間のかかる本格的な認識アルゴリズムを適用するのは得策でない。大分類では比較的軽い処理によって候補カテゴリを絞り込む処理を行う。「大分類」というといくつかのサブクラスのうちから選択するという印象を受けるが、実際は候補カテゴリを絞り込む処理であり、残るカテゴリは毎回異なるサブセットとなるのが普通である。

### ○知識処理

認識対象の文字の並びが単語や文をなす場合、1文字1文字は必ずしも完璧に読めなくても単語や言語に関する知識を用いてより確からしい結果を出せる場合がある。このような処理を知識処理と呼ぶ。通常は認識結果

の候補を求めてから知識を用いるため、後処理と呼ぶこともある。

### ○文字切り出し

文字認識技術応用システムの入力は単一の文字に対応する画像の場合もあるが、複数の文字を含む画像である場合も多く、文字かどうかさえ不明の場合もある。入力画像中から単一の文字を切り出す問題は狭義の文字認識に匹敵する難問とされている。画像入力単位によって必要に応じて文字領域抽出、文字行切り出し、文字切り出しという順序で切り出す。単語単位に分かち書きされるアルファベット系言語の場合は単語切り出しという階層が加わることもある。

### ○レイアウト理解

OCR技術を応用して雑誌や新聞を認識する場合、名刺をイメージ入力して名簿管理に使う場合、あるいは未知あるいは既知でも分類不明の帳票を読み取る場合には見出し、段組み、セル（帳票における書き込み欄）の従属関係等、文書の幾何学的構造を理解する機能が必要となる。それを実現するのがレイアウト理解技術である。幾何学的構造を理解することにより認識結果をしかるべきレコードに入れることができ、また、適切な知識処理を施すことができる。この技術は文書画像理解の分野の主要な課題のうちの1つである。

### ○修正インタフェース

OCRの応用製品では文字を認識させっぱなしというケースは少なく、誤認識の修正やリジェクト文字に対するコードを入力するためのインタフェースが必要となる。ストレスなく高速、正確に修正を行うためのインタフェースの設計も文字認識技術の重要な課題の1つである。

## 情報源

### ○学会・研究会

国内で開催される学会で文字認識や文書画像処理関連を含み専門性が高いものとしては隔年開催の画像の認識・理解シンポジウムMIRU (Meeting on Image Recognition and Understanding) がある。定期開催の研究会としては情報処理学会のCVIM (コンピュータビジョンとイメージメディア) 研究会、電子情報通信学会のPRMU (パターン認識とメディア理解) 研究会がある。雑誌では情報処理学会論文誌、電子情報通信学会論文誌DII分冊、PRMUの研究報告等がある。

国際会議としてはパターン認識分野では最も伝統のあるICPR (International Conference on Pattern Recognition) の他に専門性の高いものとしては隔年で交互に開催されるICDAR (International Conference on Document Analysis and Recognition) とIWFHR (International Workshop on Frontiers in Handwriting Recognition) があり、前者は文書画像処理全般、後者は手書き文字認識専門である。海外の雑誌でパターン認識専門のものではPattern Recognition, Pattern Recognition Letters等があるが、その他、IEEE Transactions on Systems, Man and Cybernetics, Neural Networks, Neural Computation等にもしばしば文字認識関係の論文が掲載される。

### ○データベース

認識辞書を学習する際や未学習データで認識性能を評価する際には文字や文書の画像データベースが必要となる。辞書作成は実データによらず設計する場合もあるが、評価時には実データはほぼ必須となる。特に、他の文献で発表されている方式と比較を行うためには、共通のデータを用いることが客観的な性能比較の必要条件となる。ちなみに(客観的な性能比較の)十分条件を満たすためには評価データとシステムとが完全に独立でなければならない。したがって、決定する基準がないシステム中の未定のパラメータを決定するために、値をいろいろ変えてみて評価データで最も性能のよかった値に決めたりすると厳密にはシステムと実データは独立ではなくなってしまうので注意が必要である(よく行われていることではあるが)。

このような目的のために公開されているデータベースがいくつかある。切り出された文字のデータベースとして国内で最もよく知られているのは電子技術総合研究所(当時)によって作成された「ETL文字データベース」である。手書き英数字、ひらがな、カタカナ、教育漢字、JIS第一水準漢字、印刷漢字等のサンプルが収められており、ETL1~9の9種類がある。頁単位のイメージデータとしては、日本電子工業振興協会(電子協)によって作成された「JEIDA '93」があり、新聞、書籍、電話帳、説明書、特許明細書等のイメージが収められている。両者とも電子協の認識形入力方式専門委員会のWebページ(URL: <http://it.jeita.or.jp/jhistory/committee/humanmed/recog/database>)にアクセスすれば入手方法が書かれている。他に、郵政研が認識アルゴリズムコンテスト用に作成した手書き郵便番号、手書き宛名のデータベースもあり、入手方法は上記Webページに書かれている。海外のデータベースについてはメリーランド大学のWebページ(URL: <http://documents.cfar.umd.edu/>)を参照されたい。

### ○教科書

本稿で説明した内容よりさらに踏み込んで勉強されたい読者には実応用システムに関する記述が豊富な森<sup>1)</sup>、最新の研究成果まで解説している石井他<sup>2)</sup>等をお勧めする。次章で述べる今後に残された課題については小川<sup>3)</sup>を参照されたい。

## 文字認識は終わっているか?

文字認識は研究も商用化も歴史が古いため「文字認識の研究はもうとつくに終わっているのでしょうか?」と聞かれることがよくある。しかし現状の文字認識技術応用の現場ではよりよく認識するようにユーザ側もシステム側もいろいろと工夫しており、認識屋にとっては一種の過保護状態といえる。認識速度はともかく自由に筆記された手書き文字や多様なフォントの印刷活字文字を正確に読むという意味では、OCRの実力はまだまだ人間のレベルにはほど遠い。よくある例にならえば、鳥のように空を飛ぶために違った原理で飛行する飛行機が発明され、まっすぐ飛ぶ限りスピードの点では鳥をはるかにしのぐレベルに達したが、自由自在に飛び回る飛行術という点では鳥にははるかに及ばないのと同じである。

現在実用化されている文字認識技術の主流は学習データから自動作成した辞書を用いる重ね合わせ法の系統であり、特徴抽出の改良や識別器に関するさまざまな理論的成果の適用により性能を向上させつつある。このアプローチは上記の例における飛行機に対比できる。しかしこのアプローチは人間の視覚系とはかなり違った方向に発展してきており、「1文字単位に切り出された文字を認識する」という問題設定にかなり依存している。たとえば隣どうしが接触していききれいに切り出されない文字の認識、無地でない背景の上に書かれた文字の認識、未知フォントの認識等、制約を緩めたより難しい文字認識問題に対する解決方法がその延長線上にあるとは限らない。私見ではむしろ構造解析系の方式の方がこれらの問題に向いているように思われる。

今後の文字認識研究に求められるのは、従来の問題の枠組みの中でより高い性能を求めるのと同時に、より難しい文字認識問題にトライすることである。そうすれば応用上もはるかに大きな市場が期待され、人間の視覚系の解明にもつながっていくことが期待される。文字認識は終わっていない。

### 参考文献

- 1) 森 健一: パターン認識, 電子情報通信学会 (1988).
- 2) 石井健一郎他: パターン認識, オーム社 (1998).
- 3) 小川英光: パターン認識・理解の新たな展開, 電子情報通信学会 (1994). (平成13年5月11日受付)