OCR 認識技術の動向と展望

Trends in OCR Recognition Technologies and Future Prospects

清野　和司　K. Seino
古屋　勝彦　K. Koya

特集II

現代の情報化社会において，文字認識技術の役割は大きい。応用商品としては，郵便区分機や汎(はん)用 OCR(光学式文字読取装置)，あるいはモバイルコンピュータの入力手段として使われるオンライン手書き文字入力などがある。ただし，商品の性質に依存してその技術課題は異なる。例えば，OCR にとっては紙の上から文字を抽出する技術が重要であるが，オンライン文字認識ではこの技術は不要である。おのおのの商品がそれぞれに高度な要素技術を必要とする。

われわれは，OCR 市場においてつねに最新の技術を開発し多くの商品を創出し続けてきた。現在もなお市場ニーズに適合する新しい技術の開発を続けている。ここでは，OCR の認識技術に的を絞り，現状レベルと今後の展望について述べる。

Toshiba has long contributed to the growth of the OCR market. The driving force behind this has been high-performance character recognition, based mainly on the multiple similarity method. We have been providing practical solutions to the OCR market using this technology.

In recent years there has been strong demand for the construction of intelligent databases, accompanying the popularization of high-performance and low-price personal computers and the network environment. In this connection, expectations on further enhancements of character recognition and high-performance document understanding are greatly increasing. However, effective solutions cannot be provided using the conventional technology.

With the above trends as a background, this paper describes the actual levels of some of the latest recognition technologies and future prospects in this field.

1 まえがき

OCR の認識技術は多岐にわたり非常に複雑である。しかし，その最終目標は明解で，人間の認識能力そのものの実現にある。究極の OCR を作るには，文字認識，形状認識，論理構造理解，単語認識，文章理解など，高度な人間の認識機能を，アルゴリズムで表現する必要がある。この技術開発の先には，現在のコンピュータを超越する機械が期待される。しかし，人間の認識の原理はほとんど解明されていない。なぜ文字を見つけ出せるのか，なぜ文字を読めるのか，なぜ同じ形や違う形と感じるのか。人間は，単純な化学反応の集積で，このような複雑な認識をいとも簡単に行っている。われわれは，この問題に部分的な解を与えつつ，それを市場ニーズに応じて商品化することにより，着実に OCR の市場を発展させてきた。

さて，最近のパソコンの爆発的な普及に伴い，欧米を中心としてイメージスキャナが急速に普及し始めた。さらに，インターネットの普及も手伝って紙メディアの情報を電子化する広範囲なアプリケーションが想定されている。このような市場環境のなか，OCR には，単なる文字認識装置ではなく，紙メディアからディジタル情報への変換手段としての期待が急激に高まりつつある。われわれは未知の新たな世界への第一歩を踏み出そうとしている。今後 21 世紀に向け，その根幹から技術の衣替えをする必要がある。

ここでは，以上の状況を踏まえ，①文字の認識，②帳票の認識，③文書の認識のおのおのについて，われわれが実現した最新の技術を紹介し，今後のユーザニーズにとって，ポイントとなる幾つかの技術課題について考察する。まずはその総括として，OCR 認識技術の課題を図 1 に示す。この図の中で使われている幾つかのキーワードについては，以降の章で詳細に説明する。

2 文字の認識技術

長年にわたる文字認識の研究における成果として，手書き・活字を問わず，品質の良い文字については，われわれはほぼ完全な認識性能を実現した。しかしながら，走り書きや崩し字などの文字については多くの課題を残している。OCR において，文字認識の性能は重要なポイントであることは言うまでもなく，これが人間と同じレベルに到達するまで，永遠の開発テーマである。

以下では，文字認識技術における最近の注目すべきトピ

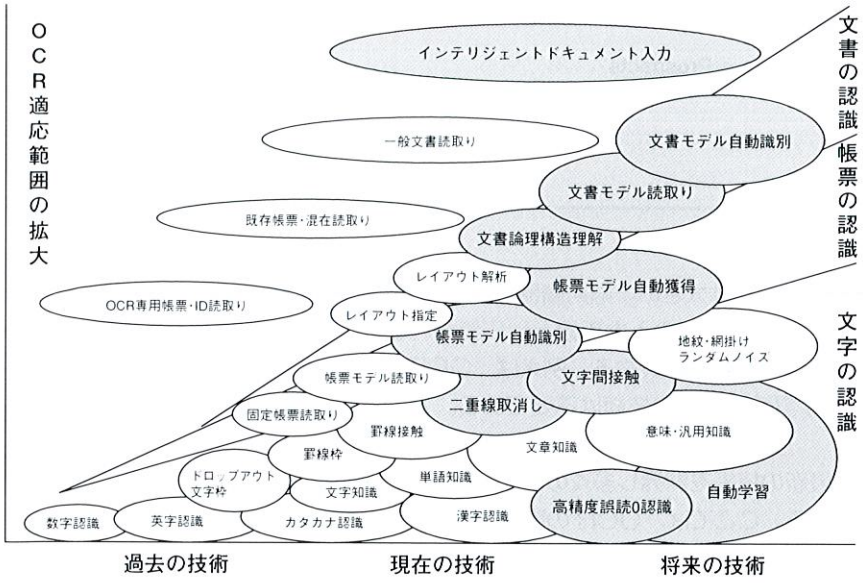

図1．OCR認識技術の課題　今後のインテリジェントなドキュメント入力を実現するために，特に注力すべき技術課題を網がけで示した。
Technical subjects of OCR recognition

ックスとして，①統合化文字認識方式，②文字抽出，③学習機能について述べる。

2.1　統合化文字認識方式

最近の文字認識方式は，単一の方式に頼るのではなく，複数の方式を有機的に統合し，おのおのの方式のもつ短所を補い合うことにより総合的に高い認識性能を得ようとするアプローチが注目されている。われわれは，すでに1970年代から，複数の特徴量や識別手法を使用するハイブリッドな方式を採用しており，さらに改良を続けている。このアプローチは，誤認識の抑制において非常に効果がある。われわれの現在の目標は，手書き数字認識において，正読率を現状のまま，誤読を100万分の1以下に押さえることである。限りなく誤読0に近い状態で認識することにより，オペレータはリジェクト文字以外をチェックする必要がなくなり，入力の生産性が非常に向上する。

このような誤読低減を目的とした認識技術の開発では，従来に増して大量な文字パターンデータを必要とする。われわれは，長年の製品展開において，多くのユーザに協力を戴き，実運用データを大量に集積してきた。これは，われわれの文字認識技術を支える重要な基盤となっている。

2.2　高性能な文字抽出

文字認識の過程において，文字パターンの認識に先立ち，1文字ごとを抜き出す処理を文字抽出と呼ぶ。この処理では，非ドロップアウトカラーの罫(けい)線と文字，あるいは文字どうしの接触などが問題になる。罫線と文字の接触については，おのおのの濃度の違いを用いた多値イメージで解決する方法，その太さの違いを用いる方法，文字の輪郭の連続性を用いる方法など幾つかの解を得ている。さらに，最近では二重線取消しのあるフィールドについて，その上に記入された訂正文字を読み取る機能も実現した（図2）。また，文字どうしの接触については，手書き数字に限定した手の動特性から生ずる文字パターンの連結性の知識を使用する方式，あるいは手書きの住所・姓名などの場合，それらの単語知識を用いる方式などを開発した。文字抽出の技術は，特に既存帳票読取りにおける認識性能を大きく左右する。今後のさらなる技術向上が望まれている。また，地紋，罫線，各種画像にオーバラップした文字の抽出などでは，今後，カラー画像処理の技術を積極的に応用すべきである。

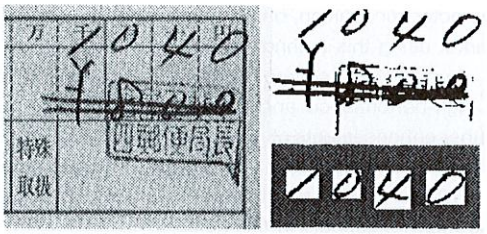

図2．高度な文字抽出　多値イメージ処理により罫線を消去し，さらに二重線取消しを検知してその上の訂正文字列を抽出する。
Advanced character extraction

2.3　学習機能

文字認識の性能は，業務内容や地域性などの影響による文字品質の傾向の違いに大きく依存する。そこで，状況に応じて最適な認識性能を実現するための手段として，実運用時の学習機能が期待されている。しかし，本格的な学習には多くの解決すべき問題がある。最大の問題は学習による悪影響の抑制である。これを本質的に解決した学習機能をもつOCRは今のところ存在しない。しかし，われわれは部分的にこれを解決し，読取対象字種の性質に応じて二つの学習機能を開発した。

一つは文字認識辞書への学習である。オペレータが非正読文字を修正すると，まず，その文字パターンを解析し，

学習の可否を自動判定する。このとき，学習による悪影響を引起こす可能性のあるパターンを除外する。その後，学習可と判断された場合にかぎり，そのパターンを認識辞書へ追加登録する。パターンの登録は，その帳票を修正している間の一時的な学習（短期記憶的）と永久的な学習（長期記憶的）とがある。前者の学習は帳票内の同形パターンの修正効率の向上を目的とする。

もう一つは単語知識辞書への学習である。一般的に手書き漢字の読取りでは住所や姓名などの単語辞書を用いた知識処理を行っている。知識処理は，文字認識の精度，単語辞書のカバー率，単語の発生頻度に依存してバランスを調整する必要がある。そこで，これらの各パラメータを常時監視し学習状態に応じて最適な状態を保つような学習方式を開発した。

今後は個人レベルでの文字認識需要の増加が予想される。そこでは個人の癖字や筆跡の学習が期待される。その場合，より高度な学習方式の実現が必要となる。

3 帳票（フォーム）の認識技術

旧来のOCRは，その技術レベルの問題からユーザにOCR専用帳票の設計を強制せざるを得なかった。例えば，フォームをドロップアウトカラーで印刷する必要があったり，文字記入領域には1文字単位の文字枠を置く必要があった。したがって，ユーザが業務にOCRを導入する場合，既存の帳票を廃止し，改めてOCR専用帳票を設計する必要があった。しかし，最近の技術レベルの向上により，従来から使用していた既存帳票をそのまま読み取ることができるようになってきた。紙の厚さや紙質の制限，あるいは帳票の印刷色の制限は依然としてあるものの，既存の帳票がこの許容範囲内に入る場合，OCR導入時の初期投資が非常に低く押さえられる。

以下では，既存帳票読取において期待の大きい，①帳票モデル自動獲得，②帳票モデル自動識別，について述べる。

3.1 帳票モデル自動獲得

一般に帳票読取OCRでは，帳票内の読取位置や読取字種などの情報（帳票モデル：FC）をあらかじめパラメータで指定している。その負荷を軽減するために，OCR専用帳票を設計する場合，帳票設計CADで同時に帳票モデルの指定を行えるシステムもある。しかし，既存帳票の場合は，OCRに合わせた帳票設計が不可能なためこの機能は使えない。そこで，ブランクシート（あるいは記入のあるサンプルシート）の画像から自動的に帳票モデルを作成する機能が期待され，実現可能な段階に入っている。フォームを形成する罫線情報の自動抽出を行い，さらに，プレ印字されたフィールド名称などの文字を読み取り，その帳票の読取フィールドや読取字種を自動的に予測する。

3.2 帳票モデル自動識別

ユーザの運用において複数種類のフォームを必要とする場合，OCR専用帳票であれば，帳票IDという識別子をあらかじめ印刷しておき，それを読み取ることによりおのおのの帳票に対する読取処理を行うことができた。しかし，図3のような既存帳票の読取りにおいては，そのような明示的な識別子は期待できない。そこで，われわれは，フォームを形成する罫線情報を用いて，複数の既存帳票を自動的に識別する方式を開発した。対象とするすべての帳票について，罫線の骨格情報を含む帳票モデルを登録し，被読取帳票について，各モデルの中から罫線骨格的にもっとも類似するものを見つけ出し読取処理を行う。

この技術の開発により，複数種類の既存帳票における混在読取処理が可能となった。今後は，罫線の無い帳票に対応するため，帳票内に印刷された固定のマークやロゴなどをキーとする識別方式の開発が必要となる。

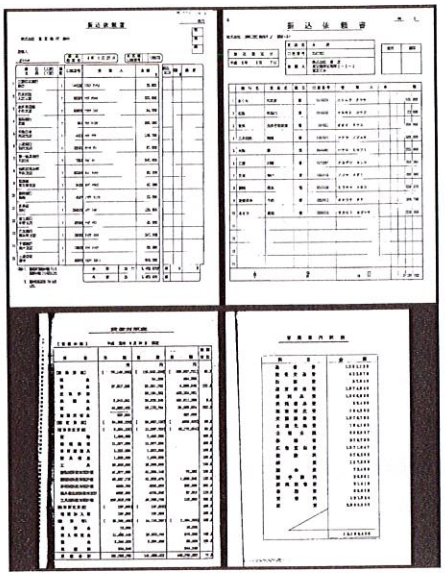

図3．帳票モデル自動識別　微妙な罫線構造の違いも自動的に識別し，混在入力された個々の既存帳票ごとに別々の読取り処理を行う。
Form model retrieval

4 文書（ドキュメント）の認識技術

一般の印刷文書を入力する場合，文書中の文字領域や図形領域などを自動的に識別する必要がある。この機能をレイアウト解析機能と呼ぶ。旧来の文書読取OCRは，文書から文字領域だけを抽出して文字認識を行い，その他の記載情報は除外していた（テキストリーダ）。しかし，最近では文字領域だけでなく，表，線画，写真，数式などのテキスト以外の属性をもつ領域を識別できるようになった（ドキュメントリーダ）。したがって，出力文書形式も，従来のテキ

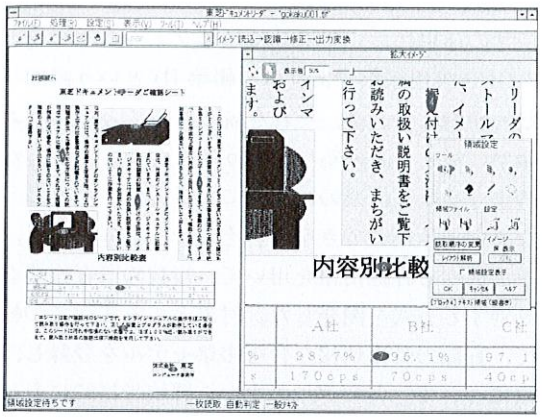

図4．文書レイアウト解析　表やイメージだけでなく，タイトルやキャプションを的確に分離しており，今後の論理構造理解機能の可能性を示唆する。

Document layout analysis

スト形式の他に表やイメージを含む複合ドキュメント形式をサポートできる。さらに，テキスト領域の縦書き／横書きの自動判定も行うことができるようになった。また，レイアウト解析の結果として得られた領域の認識結果について，その位置や大きさ，領域属性，読取順序などの修正を行うためのユーザインタフェース（レイアウト修正）も重要である。この機能の操作性の良さも，全体の入力効率に大きく影響する。図4に，ドキュメントリーダのレイアウト解析結果とその修正画面を示す。

以下では，ドキュメント管理システムとの連係を考慮した場合に重要となる，①文書論理構造理解，②文書モデルについて述べる。

4.1　文書論理構造理解

文書論理構造理解とは，次のような機能である。つまり，文書の物理的なレイアウト情報や文字認識の結果を総合的に判断し，そのページのヘッダ，フッタ，タイトル，見出し，キャプションなどの論理的なタグ情報を自動的に認識する。また，複数ページの文書については，表紙，目次，本文，章だて，節だて，文献，索引などのページにまたがる論理構造を認識する。

この認識においては，画像のレイアウト解析技術以外に，日本語形態素解析を駆使した総合的な知識処理を必要とする。この技術により，SGMLやHTMLなどの文書構造表記形式への本質的な自動出力が可能になる。この機能は，後方に控えるファイリングや検索などにとって必要不可欠である。今後の技術開発における期待が大きい。

4.2　文書モデル

文書モデルには，その文書種別に対するレイアウト構造，論理構造，領域の読取字種などの情報が格納される。さらに，定義として，複数ページ文書も対象とする。われわれはより高精度な領域認識性能を得るために，新聞，名刺，電話帳，財務諸表，あるいは英文や欧文など，個々の文書モデルを知識として使用するレイアウト解析機能を開発した。

現時点では，特定の文書モデルに特化したレイアウト解析の開発は重要である。これによりレイアウト修正はほとんど不要となり，領域の読取字種が限定される場合，文字認識の性能も向上する。ただし，モデルの選択はオペレータが行う。今後は，任意の文書モデルを外部から与えたり，入力された文書がどのモデルに対応するかを自動識別する機能の実現も期待できる。これにより，ユーザ固有の文書モデルの定義や複数種類の文書の混在読取処理が可能となる。この技術は，OCRと後方のドキュメント管理システムとの親和性をより向上させる。

5　あとがき

今後の情報通信サービス市場はその発展が計り知れない。それに伴い，OCR認識技術に対する期待度もエスカレートしていくことが予想される。そのような高度な要求にこたえるために，OCRの認識に関連する大規模な知識ベースの必要性を感ずる。単なるシート画像，文字パターン，単語，コーパスなどのデータだけではなく，人間の脳の内部処理を模倣するためのあらゆる情報を知識として貯える。そしてこれを活用するためには，文化，遺伝，生命，脳，心理などの工学的モデルが重要となる。

以上，現時点での最新のOCR認識技術の紹介を通じて，そのポテンシャルを再確認し，さらに，今後必要とされる新しい技術開発について考察した。今後，OCR市場のさらなる飛躍を目ざして，ここでの考察に基づく新技術開発を精力的に進めていく。

文　献

(1) 坂井邦夫，他：文字認識技術の発展とOCRシステムの広がり，東芝レビュー，47，2，pp.84-87 (1992)
(2) Y. Ishitani: Model Matching Based on Association Graph for Form Image Understanding, Proc. 3rd ICDAR, pp.287-292 (1995)
(3) 石谷康人：創発的計算に基づく文書画像のレイアウト解析，画像の認識理解シンポジウムMIRU'96，Vol.I，pp.343-348 (1996)

清野　和司　Kazushi Seino

青梅工場コンピュータマルチメディア設計部主査。
OCR認識技術の設計開発に従事。
Ome Works

古屋　勝彦　Katsuhiko Koya

青梅工場コンピュータマルチメディア設計部主査。
OCR認識技術の設計開発に従事。
Ome Works