特集

図書の電子化のための文書認識技術

Document Recognition Method for Digital Libraries

あらまし

電子図書館のコンテンツ作成を目的として，雑誌や新聞を含む印刷図書を高精度に電子コード化する文書認識技術を開発した。図書から文字コードを認識するには，イメージスキャナにより得られた文書画像から文章，表，図などの領域を抽出した後，文章や表の中の文字を電子コードにする文字認識を適用する。領域識別処理においては，「仮説検証方式」により，横書と縦書が混在した図書に対しても高精度に識別可能にした。文字認識処理では，「精度保証型高速分類方式」により，候補文字を抽出する大分類を行わず，直接に詳細な文字識別を行った場合の文字認識精度を達成可能にした。さらに，将来的にカラー図書の電子化を目指して，一様色の文章画像を高精度に2値画像に変換できる「情報保存型2値化方式」を開発した。富士通では，これらの技術を利用して電子図書館システムを開発している。

Abstract

This paper describes our newly developed document recognition method by which printed materials such as books, magazines, and newspapers are accurately converted into electronic code data to create the contents of a digital library. After the domains of sentences, tables and figures are extracted from document images using document layout analysis processing, then sentence and table text are recognized using character recognition processing. Document layout analysis processing allows multiple columns to be precisely extracted using the hypothesis test method, even in a document containing horizontal and vertical lines. Therefore, characters are recognized precisely at the recognition accuracy at which characters are accurately recognized by using only a detail classification method. We have also developed a binary conversion method by which uniform color text images are precisely converted into binary images. This method will be used in the future to convert color books into electronic code data. Fujitsu is now developing digital library systems using the methods above.

鎌田　洋（かまだ　ひろし）

1981年広島大学大学院博士課程前期了。同年(株)富士通研究所入社。以来画像処理，文書認識の研究開発に従事。1992年および1993年画像工学コンファレンス優秀ポスタ賞受賞。技術士(情報工学部門)。
マルチメディアシステム研究所ドキュメント処理研究部

藤本克仁（ふじもと　かつひと）

1987年東京大学工学部計数工学科卒。1992年(株)富士通研究所入社。以来コンピュータビジョン，文書画像認識技術の開発に従事。
マルチメディアシステム研究所ドキュメント処理研究部

黒川浩司（くろかわ　こうじ）

1991年福井工業高等専門学校工業化学科卒。同年(株)富士通研究所入社。以来材料分析，文書画像認識技術の開発に従事。
マルチメディアシステム研究所ドキュメント処理研究部

ま　え　が　き

電子図書館ではコンテンツをいかに整備するかが重要である。コンテンツの整備方法としては，従来からの紙の図書をもとにした下記(1)(2)(3)の方法が有力な候補である。

(1)　全文を完全にコード化して従来の図書代わりに用いる。
(2)　目次などの概要情報をコード化して用いる。
(3)　検索のために図書をコード化した情報を持つ。

いずれの方法にしても，読み取り範囲の指示を人手で行うことなく，効率的な入力処理を実現するために，図書から読み取るべき領域を自動的に判断して必要な領域をコード化する技術が重要である。

このための文書認識技術の構成を図-1に示す。文書画像から文字コードを認識するには，文章・図・表などの個別の領域を抽出する領域識別処理を行った後，文章や表の中における文字を認識する文字認識処理を行う。この一連の文書認識処理において，高精度性を追求した。さらに，最近，普及が著しいカラー図書を将来的に電子化するために，一様色の文章画像を，文書認識処理の入力となる2値画像に高精度に変換する技術を開発した。

本稿では，著者らが開発した文書認識技術において，高精度な領域識別技術，高精度な文字認識技術，カラー文書画像の2値化技術を順に取り上げて紹介する。

高精度な領域識別技術

世の中には様々なレイアウトの図書が存在し，世の中のすべての図書を精度良く領域識別することは，困難な課題である。

そこで，著者らは，多くの図書で妥当と考えられる「長方形枠で区切って個別の領域を分離できる」という制約を満たす図書を認識対象にして，高精度な領域識別技術を開発した。

開発した領域識別技術では，様々な領域を抽出した後，文章領域において段を抽出する。領域の抽出では，ひとかたまりになった描画部分を抽出した後，この描画領域の並び方に規則性があれば文章領域，マス目状の罫線があれば表領域といったように，領域に分類する。

文章領域における段の抽出では，文章中の文字の大きさや位置という多数の関係に合致する段を抽出する必要がある。

標準的な段抽出の方法としては，文章中のすべての文字の大きさや位置関係を統計解析して，精度良く行方向を検出した後，行方向を基準にして文字，行，段を順に抽出するものがある。しかし，この方法は横書か縦書の文章に対しては適用できるが，横書や縦書の段が混在し

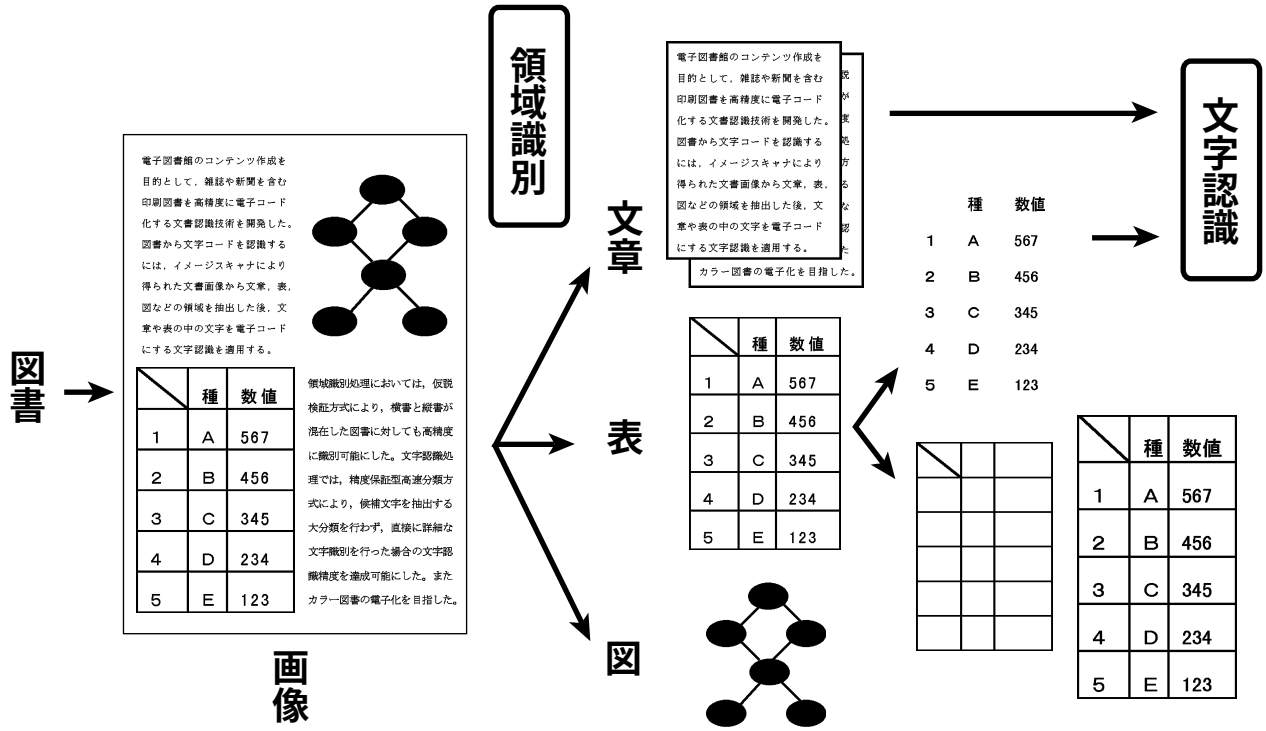

図-1　文書認識技術の構成
Fig.1-Configuration of document recognition method.

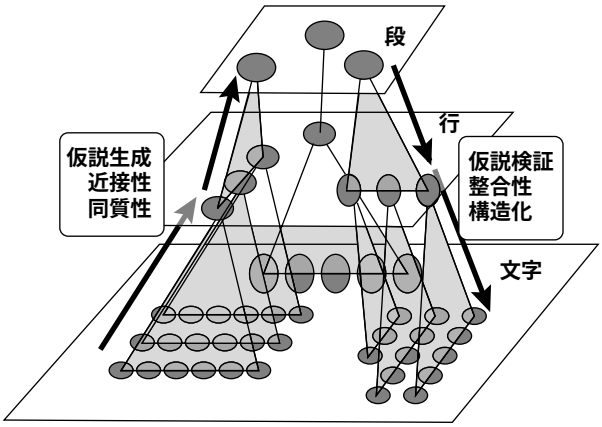

図-2　仮説検証方式による段抽出
Fig.2-Multiple columns extraction using hypothesis testing method.

た文章に対して適用することはできない。

段の抽出は文字の抽出に不可欠な処理であり，しかも横書や縦書の段が混在した図書も重要である。そこで，領域識別技術においては，段の抽出処理を重点課題として開発した。この結果，「仮説検証方式」を考案することにより，横書と縦書が混在した複雑なレイアウトの段組をも抽出可能にした。

仮説検証方式による段抽出の原理を図-2に示す。仮説検証方式では，段を行の集合，行を文字の集合としてとらえて，文章領域を段，行，文字の3階層の構造としてとらえる。下位の文書要素から上位の文書要素を仮説として生成し，上位の階層の整合性により仮説を検証することを，信頼度の高い行や段を増加させるように制御しながら，すべての文字が段を構成するまで繰り返して，段を抽出する。

まず，繰り返し処理の1回目の処理を説明する。

お互いに近接しかつ大きさ・間隔が同程度である文字の集合を行とする仮説を生成する。生成した行は一般に重なっており，矛盾を除くための検証を行う。重なっている行のそれぞれについて，文字を多く含む行を信頼度が高いと判断し，大きさが小さい行を信頼度が低いと判断して，最も信頼度の高い行のみを残す。残った行については，行と文字との整合性のために，行に含まれる文字を行に垂直な方向に統合して，行を抽出する基礎となった文字を再構成する。

同様に，互いに近接しており，大きさ・間隔が同質な行の集合を段とする仮説を生成する。生成した段は一般に重なっており，矛盾を除くための検証を行う。重なっている段のそれぞれについて，含んでいる行や文字の情報を用いて，最も信頼度の高い段のみを残す。残った段については，段と行の整合性を向上するために，段に含まれる行を段の中の行方向に統合して，段を抽出する基礎となった行を再構成する。

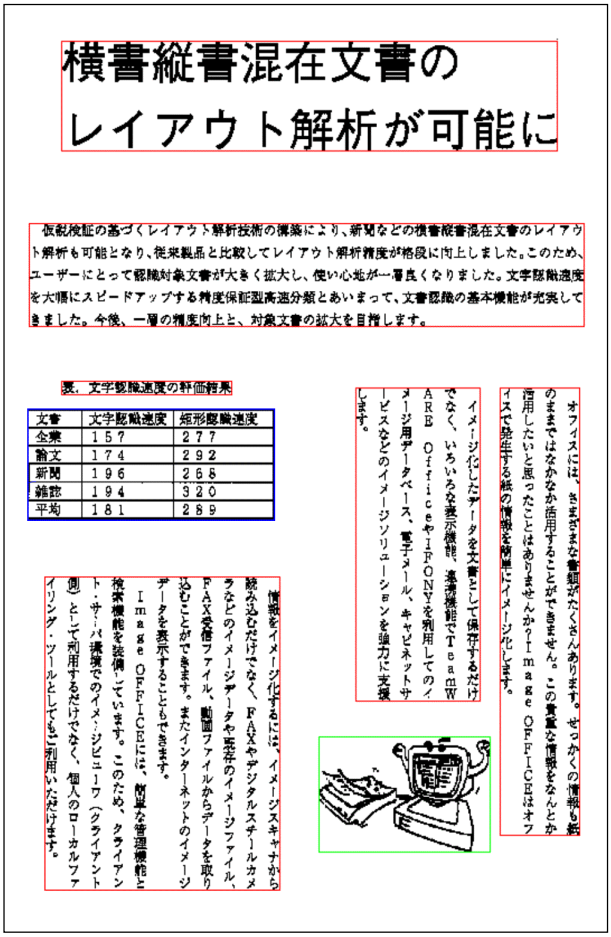

図-3　領域識別の処理例
Fig.3-Example of results of document layout analysis.

繰り返し処理の2回目以降では，すでに抽出した行や段に，行に含まれない文字や，段に含まれない行が統合されやすい規準で仮説生成の処理を行う。また，仮説生成，仮説検証の処理が繰り返されるに従い，徐々に行や段が抽出されやすくなるように，抽出パラメタを制御する。

このことで，行や段に含まれる文字の割合が徐々に増大し，最終的には，すべての文字がいずれかの段に含まれるようになり，段が確定される。

仮説検証方式では，文章全体の行方向を用いずに，文字の大きさや位置の関係のみを用いるため，横書と縦書が混在した図書からも段を抽出できる。

仮説検証方式の考案の結果，学術図書に対して，約9割の領域識別精度を達成することができた。領域識別の処理例を図-3に示す。横書と縦書が混在した文書について，領域識別技術により抽出できた文章の段，表，図の領域を，順に赤，青，緑の枠で囲って示した。それぞれの領域を正しく抽出できていることが分かる。

高精度な文字認識技術[(1), (2)]

日本語の印刷図書では，第1水準の漢字を含む約3,500の文字種を認識対象とするので，数十程度の候補文字種に絞り込む大分類を行った後，得られた候補文字種に対して高精度な詳細識別を行う2段階処理を行う。ところが，従来は，入力文字の特徴と認識用辞書に文字種ごとに保持した標準的な文字特徴との相違度の計算により大分類を行っていたため，大分類処理において，認識精度が低下する問題があった。

そこで，著者らは，文字認識精度を低下させずに高速に認識結果の候補となる文字を求める「精度保証型高速分類方式」を考案した。この方式は，直接に詳細識別したときに正解となり得る文字種を，入力文字の大分類用の特徴値から求める候補文字テーブルを事前に用意しておき，大分類の実行時には，候補文字テーブルを引く処理のみを行うものである。具体的には，下記(1)(2)の処理を行っている。

(1) 候補文字テーブルの事前作成：詳細識別用の文字特徴空間を統計解析して，文字特徴を分離しやすい複数の圧縮特徴軸を求め，特徴空間中の各文字の特徴領域から圧縮特徴軸への射影を推定して，各圧縮特徴から候補文字種集合への対応づけを候補文字テーブルに設定する。

(2) 大分類の高速実行：入力文字画像から抽出した各圧縮特徴の値から候補文字テーブルを用いて求めた複数の候補文字種集合の共通部分を候補文字種とする。

精度保証型高速分類方式の原理を図-4に示す。文字特徴空間における文字種「あ」「い」「う」の分布を圧縮特徴1，2に射影している。射影された文字種の組合せにより，圧縮特徴1，2の空間が分割され，候補文字テーブル1，2が作成される。例えば，黒点位置に入力文字特徴があった場合，候補文字テーブル1，2を検索して，二つの候補文字集合{「あ」「い」}，{「い」「う」}が得られ，集合の共通部分である「い」が認識候補文字になる。

精度保証型高速分類方式の考案により，直接に詳細識別を行った場合の認識精度を保持しつつ，高速に，約3,500の認識対象文字種から平均85文字種の認識候補文字を求めることができるようになった。この結果，社内評価用文書に対して，毎秒160文字程度（Pentium 166 MHz）の認識速度で，99%以上の認識精度を実現した。

カラー文書画像の２値化技術

近年，学術図書や雑誌においても，カラー印刷されたり写真が挿入されたりした図書の割合が増えている。

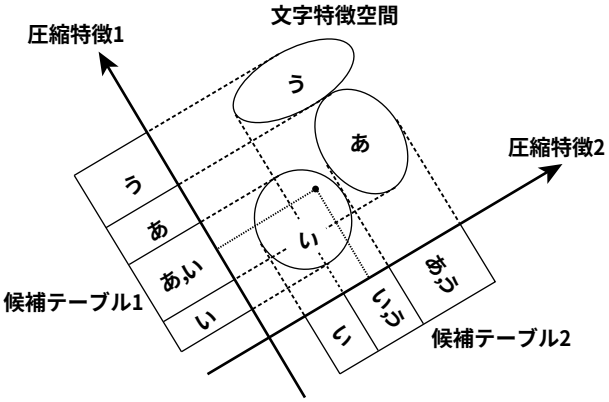

図-4　精度保証型高速分類方式の原理
Fig.4-Principle of fast precise pre-classification method.

そこで，著者らは，カラー図書を認識可能にする第一歩として，グレースケール文書画像(注1)を含むカラー文書画像の情報を保存しつつ，高精度に2値画像に変換する「情報保存型2値化方式」を考案した。学術関係の書物に関しては，情報量が多いため最も重要な文章部分は，一様色の背景のものが大部分であるため，一様色の文章部分を対象にして研究開発した。

イメージスキャナでは，カラー画像を読み込むとき，文書の読み取りラインを移動しつつ，カラー3原色である赤，緑，青の値を順に入力する仕組みである。このため，3原色の値を読み取る位置が微妙にずれて，文字の周辺などに多数の微小な色ずれが発生するので，色を手掛かりに描画部分を抽出することは難しい。そこで，カラー画像をいったんグレースケール画像に変換した後，画素の明るさを手掛かりにして，2値化するようにした。

情報保存型2値化方式の処理構成を図-5に示す。イメージスキャナなどで行われている従来の2値化法による2値化の結果の比較も示した。本方式では従来法に比較して，文字の線をより忠実に抽出できている。

グレースケール画像の2値化は，下記(1)(2)の2段階処理にした。

(1) 文字の粗抽出：画像全体の明度値についてのしきい値処理により，周辺部を含めて文字を粗く抽出する。文字の背景部分を除くことにより，後段の処理における精度と処理速度を向上できる。

(2) 精緻な局所2値化：画素間の補間により画像を精緻化するサブピクセル化と，1画素ごとに2値化するしきい値を決定する局所2値化を統合した新処理を行う。

（注1）白色と黒色に加えて灰色を含む文書画像。

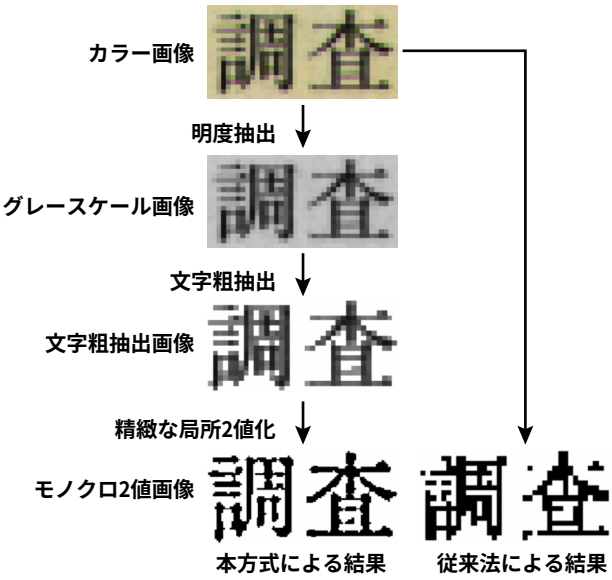

図-5　情報保存型2値化方式の処理構成
Fig.5-Configuration of precise binarization method.

この統合処理により，高精度な2値画像を得るとともに，処理速度を向上できる。

カラー文書画像(150 dpi)に対して，約3秒で高精度なモノクロ2値画像が得られ，2値画像を文字認識する実験において，従来法と比較して，誤読した文字数が半分未満になる効果が得られた。

む　す　び

電子図書館のコンテンツ作成を目的として，情報量が多いため最も重要な文字を中心に，従来からの図書を高精度に電子コード化する三つの特長的な文書認識技術を研究開発した。

第1に，イメージスキャナにより得られた文書画像から文章，表，図を抽出する領域識別技術では，「仮説検証方式」を考案して，横書と縦書が混在した複雑なレイアウトの図書に対しても，高精度に段組を識別できるようにした。

第2に，文章や表の領域の文字画像をコード化する文字認識技術では，「精度保証型高速分類方式」を考案して，直接に詳細識別を行った際の高い認識精度を保持しつつ候補文字種を求め，高精度に文字認識できるようにした。

第3に，カラー図書が増大しているため，「情報保存型2値化方式」の考案により，一様色の文章部分を高精度に2値画像に変換できるようにした。

富士通では，上記の技術を利用して電子図書館システムを開発している。

今後は，さらに文書認識精度を向上するとともに，認識対象となる図書の範囲を拡大していきたい。

参 考 文 献

(1)　鎌田ほか：圧縮文字特徴と原特徴を併用した高速かつ高精度な文書認識方式．電子情報通信学会技術研究報告，1997，PRMU96-205，pp.97-104.

(2)　藤本ほか：特徴領域の射影推定による高速高精度な大分類方式．電子情報通信学会技術研究報告，1998，PRMU97-220，pp.97-104.