論文 / 著書情報
Article / Book Information

題目(和文)	文書画像の構造認識と実用システム構成法に関する研究
Title(English)	
著者(和文)	岩城修
Author(English)	
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:乙第3274号, 授与年月日:1999年2月28日, 学位の種別:論文博士, 審査員:
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:乙第3274号, Conferred date:1999/2/28, Degree Type:Thesis doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

文書画像の構造認識と
実用システム構成法に関する研究

平成 11 年 2 月

岩　城　　修

目　次

第 1 章　序論

1.1　研究の目的

情報の伝達や保存には，長い間紙をメディアとする文書が用いられてきたが，近年のパーソナルコンピュータの発達・普及によって，多くの文書はワードプロセッサ等によって電子的に作成されるに至った．このことにより，今後紙をメディアとする文書は次第に減少していくであろうか．今日，どこのオフィスにおいても，電子メールを含めた種々の文書はパーソナルコンピュータを用いて作成，交換，保存される．しかしながら，同時に日々多くの文書がプリンタから印刷出力され，紙はむしろ増大しているとも言える．その理由の一つは，紙がもつ可読性, 可搬性, 簡単に書き込みができるなどの利便性ゆえであろう．1980 年代にはペーパレス・オフィスの実現が叫ばれたが，今日の解釈は紙と電子メディアが相互にその特性を活かしつつ，統合化された環境を創り出すことと考えられる．このことは，コンピュータで印刷物と電子ファイルが同じように扱えることを期待するものである．すなわち，紙に印刷された文書がまるで磁気ディスクなどと同じようにコンピュータに読み込まれるようになることで，人間にとってもコンピュータにとっても優れた情報メディアとなり得る．そして，人間は紙と電子メディアの都合の良い方を何時でも自由に選択して用いることができるようになる．

さて，コンピュータが印刷文書や手書き文書を自動的に読み取る技術については，文書画像の認識技術が古くから研究されてきた．最も成功した例は，光学文字読み取り装置 (OCR: Optical Character Reader) の実用化である．文書はこれまで文字情報が主体であり，特に数値等の文字情報が中心の帳票に関しては，印刷活字や手書き文字を読み取ることにより，まさにコンピュータにも人間にも扱い得るメディアとなり得た．一方，今日のオフィスで扱われる文書は文字情報のみならず，図形やイメージ情報をも含んでいる．このことから，今後さらにこうしたマルチメディア文書を扱い得るための文書認識技術が不可欠である．

そこで本研究では，オフィスで扱われる種々の印刷文書を自動読み取りし，ワードプロセッサ等で作成される電子文書と同様に扱うことができる形態に変換する技術を確立することを目的とする．特に，近年文書のデータベース化が進められる中，文書検索時の利便性等を考慮すると，文書に内在するレイアウト構造や意味構造を反映した文書画像の構造認識が不可欠である．本研究で得られる成果をもとに，オフィスに大量に存在する遡及的文書のデータベース化を可能とするため，本研究において実際に文書認識システムを構築することに

より，実用面での課題を解決することを目標とする．

1.2 文書画像認識技術の現状

文書画像の認識技術は，文字認識技術のみならず，文書処理，図面認識，画像認識，言語処理技術のほか，符号化，データベース，マンマシンインタフェースなどのさまざまな技術分野と関連している．古くはファクシミリの符号化および伝送画像の画質改善を目的に研究が盛んとなったが，その後先に触れたOCR の実用化に伴って意味的な処理を含めた文書画像認識・理解の研究へと発展した．

文書画像認識のモデルの構築に重要な役割を果たすのが文書の書式である．文書画像における書式の重要な意味は，書式が文字列等の属性を指定している点である．すなわち，文字列を読んで意味を理解しなくとも，その配置等で例えば表題であるとか著者名であるなどその文字列が表している属性を規定することができる．したがって，書式のある文書画像に対しては，書式とのマッチングによって文字列の位置の同定や文字の切り出し，文字認識が可能となる．一方，書式を規定することのできない文書も存在する．例えば，ポスターやカタログ，チラシなどは文字に様々な変形が加えられ，また背景や写真など他の情報と重畳するなど，問題が極端に難しくなる．こうした書式のない文書に対しては，文字を認識し，単語や文章としての意味を把握することが不可欠であり，そのための意味的なモデルの構築が必要とされる．

ところで，現在の研究の中心は書式のある文書画像を対象に，書式を信号レベルのモデルとして扱う手法の提案である．書式のある文書画像の例として帳票[1-1)]，図書目録カード[1-2)]，名刺[1-3)]などを対象とした研究がある．含まれる文字列の属性が分かると，現れる文字の種類に拘束条件を付加できる場合があり，このような拘束条件が文字認識の効率と誤り率を大きく改善することも知られている．また，書式を知識として定義し，入力された文書画像をトップダウン的に解釈して対応づける手法として，トップダウン処理による記事抽出の研究[1-4)]，書式定義言語の研究[1-5)]などがある．

書式のある文書画像に対しても，書式を事前に知ることができない場合，できるだけ汎用的な書式に基づくモデルが必要となる．しかしながら，こうした汎用的なモデルを定義することは一般的には極めて困難である．したがって，処理対象文書を限定したり，あるいは人間による介在，すなわち対話的に処理結果を修正するなどを前提とせざるを得ないのが現状である．こうした対話処理を前提としても，読み取り誤りの修正作業の軽減を図るための認識精度の向上は不可欠である．

今日，実用化が進められている文書画像の認識システムは，予め書式定義さ

れた帳票を認識対象としたOCRと区別され，ドキュメントリーダと呼ばれる場合がある．ドキュメントリーダが具備する機能として，原稿走査機能，画像変換機能，レイアウト解析機能，文字切り出し機能，文字認識機能，言語処理機能，ファイル出力機能などがあるが，OCRの実用化における文字認識技術の進歩と対比して，ドキュメントリーダにおけるレイアウト解析等の技術は実用性の観点から必ずしも十分と言えない．これは，先に述べた文書画像の書式の汎用的なモデルの構築が困難であるからと言える．すなわち，今後の文書画像の認識技術の進展，ならびに実用システムの構築に際しては，人間による対話処理を前提としても，オフィスで扱われる一般的な文書画像を対象に，書式未知の文書画像の認識精度の向上を図ることが望まれる．また，現状のドキュメントリーダが具備している文字認識を基本とした文字コード情報への変換機能のみならず，読み取り結果の様々な応用を可能とするため，文字列の属性など意味に関わる処理機能を実現することも望まれる．

1.3 研究の概要と論文の構成

本研究論文は8つの章から構成されるが，大きく3つのステップからなる．

第1のステップは，第3章における文書画像の前処理技術の高度化である．具体的は，文書画像の認識に先立って必要となる文字・図形の分離技術について述べる．

第2のステップは，第4章におけるオフィスで扱われる一般文書を対象とした文書画像のレイアウト構造認識技術の実現である．また，第5章において，提案アルゴリズムを実装する文書認識システムの構成法を検証する．

第3のステップは，第6章における学術論文誌を対象とした文書画像の意味構造認識技術の実現である．また，第7章において，提案アルゴリズムを実装する文書認識システムを実用化し，その評価を行う．

図 1-1 に本研究論文の構成を示す．以下，本研究論文の構成に基づき，各章の概要を説明する．

第 2 章「文書画像の構造認識」では，文書画像の構造認識に関する基本的枠組みとして，文書のレイアウト構造および意味構造について述べ，本研究で対象とする文書の範囲について言及する．また，文書画像認識に関する従来の研究を俯瞰し，本研究の位置づけを明らかにする．

第 3 章「文字・図形の分離」では，文字・図形が混在した文書の図表領域を対象として，文字と図形を分離抽出する新しい手法について論じる．オフィスで扱われる文書・図面は多種多様な書式を有しており，また文字と図形が混在していることが多い．これらの文書・図面を認識入力するために，文字・図形

が混在した領域の文字と図形を分離抽出することを目的とする．

第 4 章「文書画像のレイアウト構造認識」では，文書をイメージ情報として入力し，文字・図形を分離抽出してそれぞれを文字認識，図形認識することにより，コンピュータ処理に適したコード情報に変換する文書画像の認識手法について論じる．文字・図形の認識に先立って文書のレイアウト構造を抽出し，文字・図形の分離抽出や一文字切り出しを行うことにより，多種多様な文書の認識を可能とすることを目的とする．

第 5 章「レイアウト構造認識に基づく文書認識システムの構成法」では，イメージ形式の文書をコンピュータ処理に適したコード情報に変換する文書認識システムについて論じる．まず，文書認識処理の実現に必要なレイアウト解析，文字認識，図形認識，イメージ圧縮の各処理技術について，既存アルゴリズムの適用性を吟味し，文書認識の実現性について述べる．次に，これらの各処理を組み合わせて文書認識処理を実行する文書認識モデルを提示し，モデルに基づき文書認識システムの性能評価を試みる．

第 6 章「文書画像の意味構造認識」では，文書画像から電子的な文書へのメディア変換に際し，文書の意味構造の認識手法について論じる．従来の手法では，意味構造モデルに構成要素領域の位置関係を記述する必要があり，構成要素の相対位置が変動する文書画像に対しては適用できないなどの問題があった．そこで，構成要素の位置，行間隔，文字の大きさ，文字数などの特徴を画像特徴としてとらえ，これらを総合的に判断して文書画像の意味構造を認識するパターン分類に基づく手法を提案する．

第 7 章「学術論文誌認識システムの実用化」では，学術論文誌に含まれる連続した複数のページからなる個々の論文の意味構造を認識し，データベースに登録するためのシステムを構築した結果について論じる．第 6 章で論じたパターン分類手法を応用して論文単位の意味構造を認識し，記述形式として SGML (Standard Generalized Markup Language) を用いた文書を生成する．現在発行されている学術論文誌を用い，実際に構築した学術論文誌認識システムの有効性を評価する．

最後に，**第 8 章「結論」**では，本研究で得られた成果を要約し，残された課題について述べる．

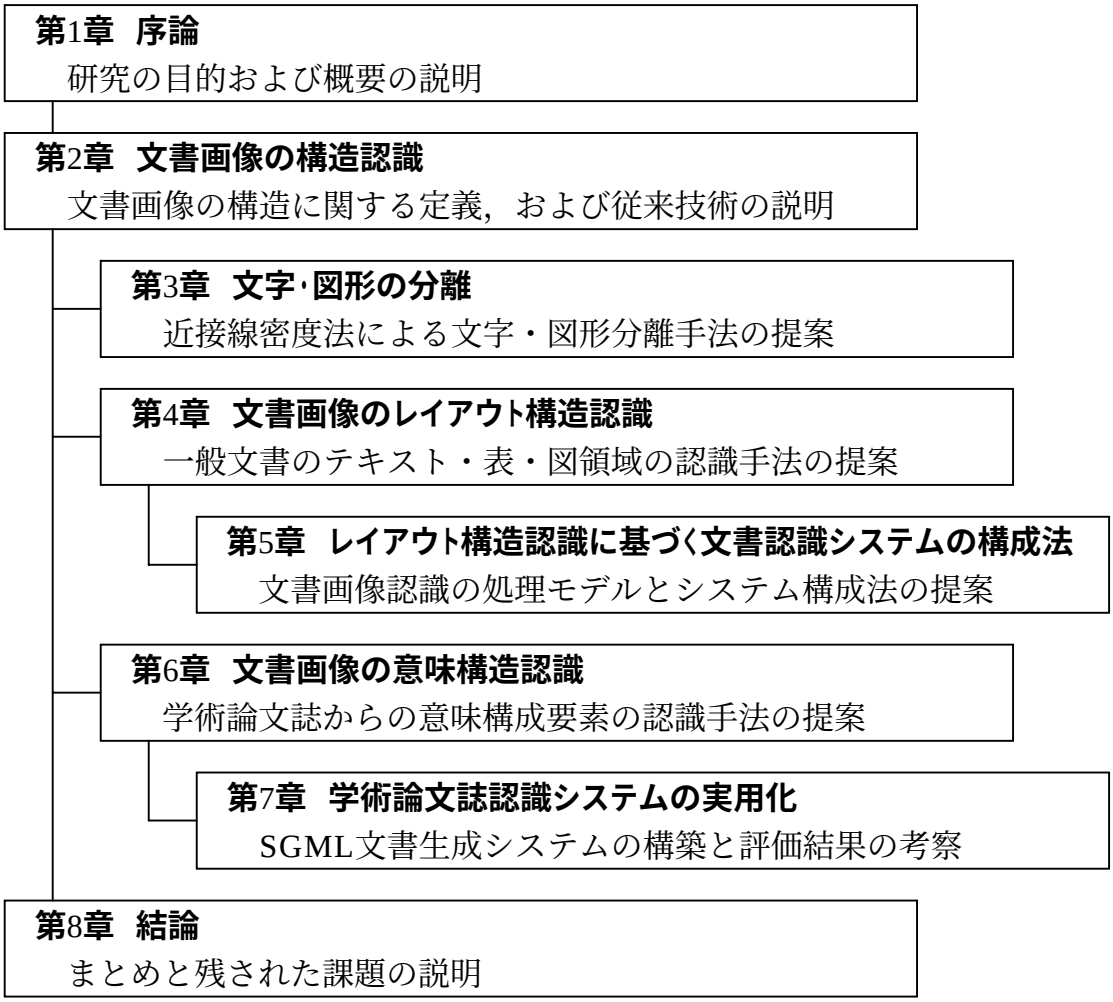


図 1-1　本研究論文の構成

Fig. 1-1　Chapters of the thesis.

第 2 章　文書画像の構造認識

2.1　文書のレイアウト構造と意味構造

文書は，本来伝達・保存したい情報を文字および図形を用いて記述し，その内容を読者に分かり易くするため，表題や著者名，章や節，図表などの意味的構造を持たせたものである．また，紙への印刷，あるいはディスプレイへの表示に際して，割り付け体裁や表示体裁などのレイアウト構造が加えられる．

紙で扱われる文書を電子文書に変換することにより，文字や図形をコード情報として扱うのみならず，これらのレイアウト構造，意味構造をそれぞれコンピュータで扱い得ることが必要である．すなわち，文書画像の認識は，これらの構造を抽出し，それぞれをコンピュータで処理可能な形式に変換することと定義することができる．

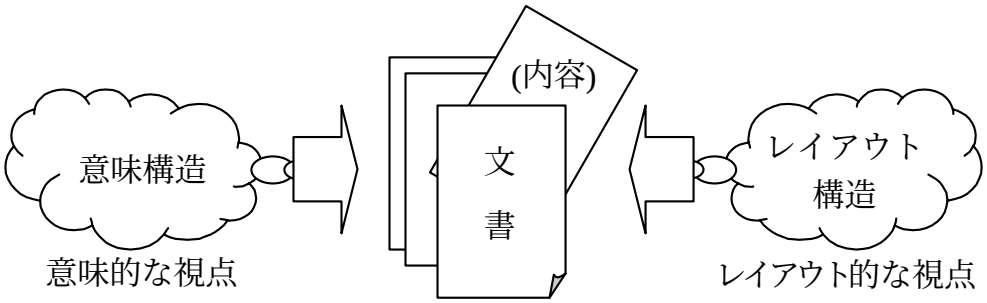


図 2-1　文書画像の意味構造とレイアウト構造
Fig. 2-1　Logical and layout structure of documents.

近年，オフィスにおいてはワードプロセッサ等を用いた電子文書の作成が一般化し，テキストのみならず，図表やイメージを含んだマルチメディア文書の普及が進んでいる．これらの文書を効率的に作成し，またお互いに交換可能とするために，文書記述の規格化が進められている．比較的最初に検討された ODA (Open Document Architecture：開放型文書体系)[2-1)]は，文書の表現方法に依存するレイアウト構造 (物理構造) と意味構造 (論理構造) をそれぞれ保持するアーキテクチャとして普及した．

その後，文書は交換の対象としてはもちろんであるが，データベース化の対象として注目されるようになった．文書に記載される内容を検索する場合などは文書のレイアウト構造は不要となるため，意味構造を中心として扱う形態が

着目されるようになった．文書記述言語であるSGML (Standard Generalized Markup Language) [2-2]は，もともと出版業界用に開発されたもので，意味構造を指定する指令 (タグ) を著者が文書中に挿入するだけで，印刷出力時のレイアウトの問題に著者が関わる必要がなくなった．

ここで，図 2-2 は文書の意味構造の例，図 2-3 はレイアウト構造の例をそれぞれ示したものである．

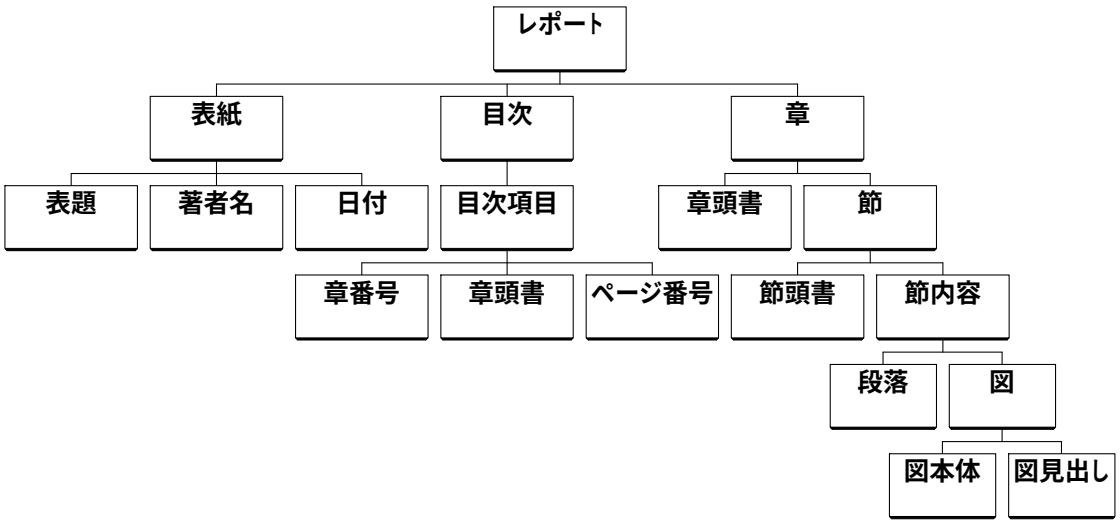


図 2-2　文書の意味構造の例

Fig. 2-2　Example of logical structure.

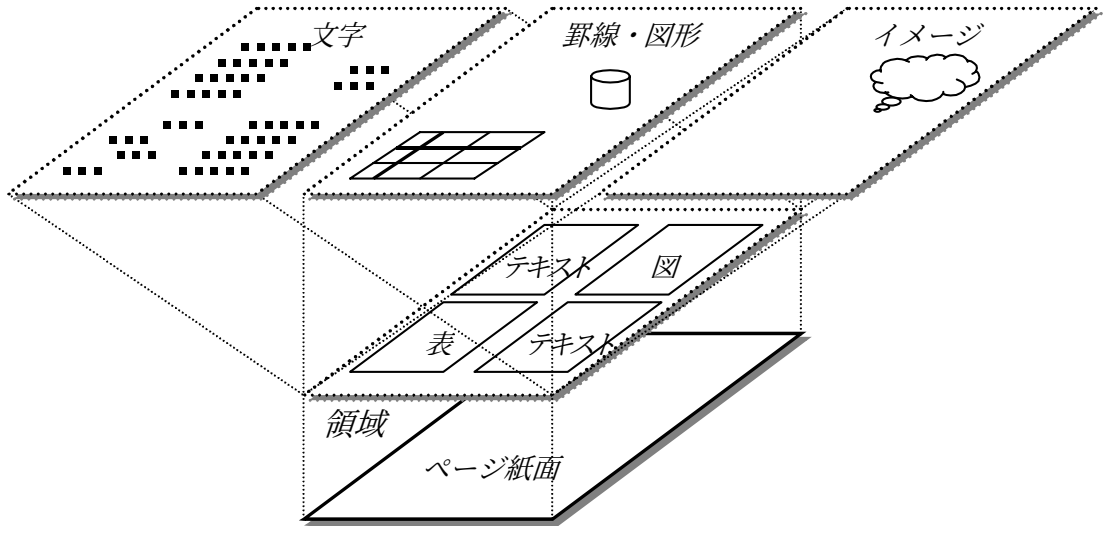


図 2-3　文書のレイアウト構造の例

Fig. 2-3　Example of layout structure.

2.2 従来の技術

2.2.1 処理対象文書画像

オフィスで扱われる文書の種類はきわめて多く，厳密に分類することは困難であるが，本研究ではワードプロセッサ等で作成・印刷出力される文書や雑誌等の印刷物を処理対象とする．一般的に，これらの文書は図 2-4 のような構造をもつものとする．

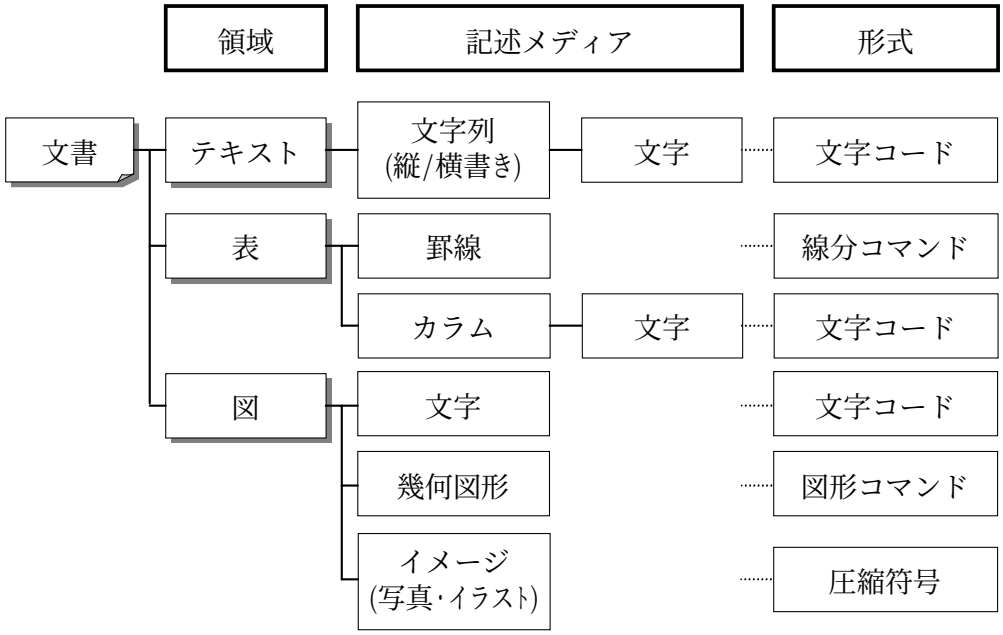


図 2-4　処理対象文書の一般的な構成

Fig. 2-4　Structure of typical document.

2.2.2 文書画像認識の処理フロー

紙で扱われる文書を電子文書に変換することにより，レイアウト構造と意味構造の 2 つの構造をそれぞれ独立に扱うことが可能である．すなわち，文書がもつ表題や著者名，章や節，図表などの意味づけされた構造と，紙への印刷，あるいはディスプレイ表示のための割り付け体裁や表示体裁などのレイアウト構造を分離し，それぞれをコンピュータで処理できる形式に変換，出力する．

従来の文書画像認識研究の多くは，未だレイアウト構造の分離抽出に留まっており，また処理対象文書に関する制約も多い．さらに，意味構造の抽出までを実現した研究はきわめて少ない．本節では，文書画像認識処理の基本となる

従来技術を俯瞰し，レイアウト構造や意味構造認識の前提となる課題を整理し，本研究の位置づけを明らかにする．

図 2-5 は，文書画像認識処理の基本的なフローを示したものである．以下，このフロー図に基づいて，従来技術の現状について述べる．

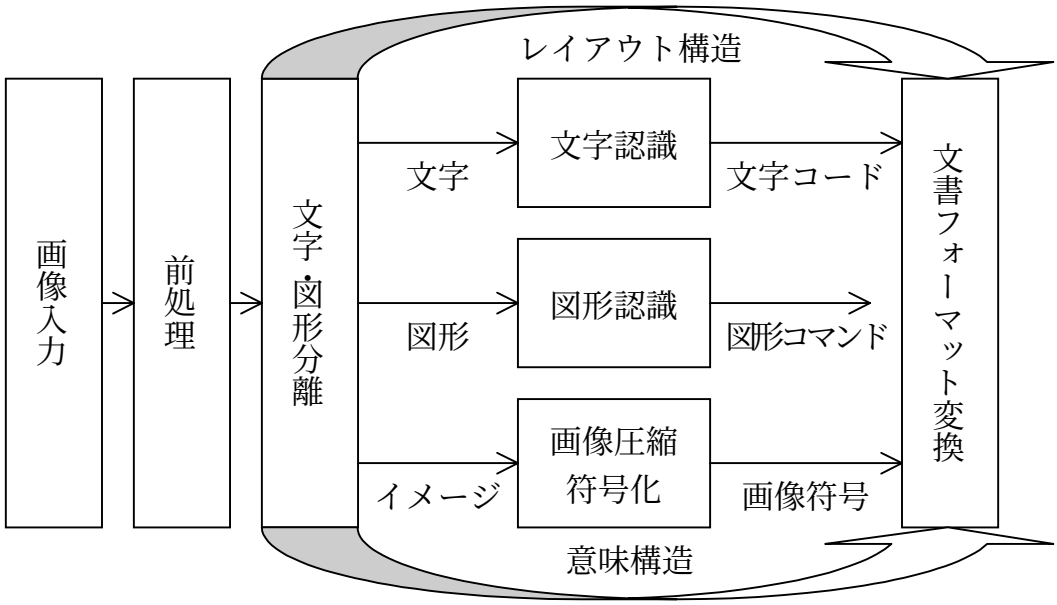


図 2-5　文書画像認識の基本処理フロー

Fig. 2-5　Document recognition processing flow.

2.2.3　画像入力

文書の入力装置としては，ディジタイザ，ライトペン等の 1 次元情報を基本として入力するものと，イメージスキャナ，ファクシミリ等のラスタ走査で 2 次元情報を基本として入力するものに大別される．本研究では，印刷された既存文書の入力を目的とするため，前者のディジタイザ等の会話型入力装置は対象としない．また，色彩情報等の高精細入力については，処理対象文書の拡張性を考慮すると今後必須と考えられるが，現状のオフィス文書はモノクロ印刷されたものが大半であることから，白黒 2 値および階調入力を基本とするイメージスキャナ，ファクシミリ装置を活用することとする．解像度については，低解像度では画像品質が低下し保存性に欠ける，あるいは認識処理において十分な認識精度が得られないなど課題が想定され，一方高解像度では画像サイズが大きくなり，処理・蓄積系の負荷増が想定されるため，現在市販される装置

の解像度を基準に 16 dots/mm (400 dots/inch) からファクシミリの標準モードである 8 dots/mm (200 dots/inch) の範囲を想定する．また，既存文書の多くが製本されていたり，ファイルとして綴じ込まれている現状を考慮すると，入力帳票を一枚ずつ挿入する原稿自動給紙機構付き装置よりは，ブック型 (原稿静止型/フラットベッド型) 装置が入力時の操作性において有利と考えられる．このようなブック型の入力装置を用いて文書を走査する場合，入力画像の回転，反射光のむら等の問題を考慮することが課題として残る．

2.2.4 前処理

領域抽出や文字・図形分離，文字認識，図形認識等の画像の内容を扱う処理の前段に位置し，入力画像を適正な形態に整える処理が前処理である．主な前処理としては，画像の回転補正，次数変換 (拡大・縮小)，雑音除去，二値化などがある．

(1) 回転補正

ブック型の入力装置では，しばしば用紙が傾いて入力されることがあり，傾いた画像は文字や図形の認識精度の低下を引き起こすため，回転補正が必要となる．一般に，原稿自動給紙機構を用いた装置における傾きは 1〜2 度以下とされており，この程度の傾きは許容するのが妥当である．ここでは，ブック型装置で起きる大きな傾きもある程度想定し，一般的な傾斜補正のための画像回転アルゴリズムとして，アフィン変換，ヘルマート変換，2 次等角写像変換，射影変換などを適用することができる．

(2) 次数変換 (拡大・縮小)

文字認識において，特に母型あるいは画素型の印刷文字に対してパタンマッチングを基本とするような認識系においては，標準文字として設定した文字サイズと入力した文字サイズを，予め拡大・縮小により整合させることが望ましい．これまでに，任意倍率の次数変換法として多くの検討がなされてきたが，代表的なものとして，投影法[2-3]，ＳＰＣ法[2-4]，論理和法[2-5]，9 分割法[2-6]，ランレングス法[2-7]，領域判別法[2-8]などを適用することができる．

(3) 雑音除去

入力画像は多くの異なった種類の雑音の影響を受けているが，デジタル画像では量子化雑音の影響[2-9]が最も大きい．その顕著な例としてごま塩雑音があげられるが，その除去には内挿法が用いられる．

雑音とその近傍がはっきり区別できるような孤立点となっているときは，各点の濃度値と近傍の濃度値を比較することによって雑音を検出することができる．この時，その点を近傍の濃度値の平均値で置き換えることによって雑音を

除去するのが内挿法である．これを簡単にしたものとして，画素の値を近傍濃度の平均値と比較する方法がとられ，特に白黒 2 値の画像では，その点と異なっている近傍の点の数を数えることによって雑音を検出できる．一般には，雑音除去フィルタとして 3×3，あるいは 4×4 のウインドを用いる．

(4) 二値化

画像入力はモノクロを基本とするが，筆記具や印刷の状態，用紙の地色によっては濃度むらが発生するため，二値化のためのしきい値の選択が悪いと文字線分のかすれやつぶれの原因となる．二値化のためのしきい値の選定には，画像全体の濃度のヒストグラムから統計的に最適な値を求める方法[2-10]が一般的であるが，細い線のかすれや線と線との間の隙間のつぶれの問題に対して，入力された濃淡画像の特徴点 (尾根点・谷点) の情報をもとに，最適なしきい値を決定する尾根点・谷点法[2-11]が提案されている．尾根点 [谷点]は 3×3 のウインドを用いて，その中央画素が他の点の濃度より高い [低い]場合に与えられる．尾根点は黒，谷点は白になるべき点である．

2.2.5 文字・図形分離

通常オフィスで扱われる文書は，テキスト領域のみならず，図表領域が混在している．テキスト領域では文字および文字列が，図表領域では図表中の文字や表罫線等の線分，定義図形が認識の対象と考えられる．また，図面等においてはテキスト領域といったような領域の区別は存在せず，図面中に書かれた文字や図形要素が認識の対象となる．

入力された文書画像から文字要素，図形要素を分離抽出するためには，まず文字・図形に関する特徴を抽出しなければならない．これらの特徴には，書式等に制限を加えることにより，文字・図形について先見的に得られる特徴 (文字の大きさやピッチ，線分の長さなど) と，文字・図形の間に存在する統計的性質の違いによる特徴 (ランレングス，ストローク密度など) がある．一般的に，図 2-4 で示した文書の構成を仮定すると，テキスト領域・図表領域を分離する場合と，文字・図形要素を分離する場合では，着目する特徴が異なることが考えられる．例えば，テキスト領域には文字列あるいは行間に周期性があり，テキスト領域を他と分離するのに有効な特徴となる．表 2-1 に，領域抽出および文字・図形分離に有効と考えられる特徴をそれぞれ示し，また表 2-2 にこれまで提案されている主な手法を，用いる特徴ごとに分類して示す．

表 2-1　文字・図形分離に用いる特徴

Table 2-1　Features for character - graphic segmentation.

	画像上の特徴	先見的に得られる可能性のある特徴
テキスト･図表領域抽出	･テキスト領域には文字列と行間の周期的な変化がある. ･テキスト領域の連結黒画素の大きさが一定している. ･テキスト領域は図形領域より黒画素密度が高い. ･写真領域は階調情報がある.	･テキスト領域中の行ピッチ,文字ピッチ ･文字の大きさ
文字･図形分離	･文字を構成する連結黒画素は図形より小さい. ･文字ストロークは図形より短い. ･文字ストロークは図形より密である. ･文字要素の黒ランは図形要素より短い.	･文字の大きさ ･定義図形の構造

表 2-2　文字・図形分離技術と適用領域

Table 2-2　Character - graphic segmentation techniques.

適用処理	用いる特徴		研究事例	条件
テキスト･図表領域抽出	･行ピッチ ･文字ピッチ	ランレングス	･RLSA[2-12)] ･ランレングス法[2-13)]	傾き補正要
		文字線分のつぶれ	･ボカシ法[2-14)] ･弛緩法[2-15)] ･拡大･縮退法[2-16)]	
		文字･文字列の投影	･周辺分布法[2-17)]	傾き補正要
		空間周波数	･フーリエ変換法[2-18)]	テキスト領域
	･濃度		･濃度分布法[2-19)]	
文字･図形分離	･連結黒画素群の大きさ		･連結黒画素法[2-20)]	文字･図形非接触
	･ランレングス		･ランレングス法[2-21)]	
	･線分長		･往復走査法[2-22)]	直線図形
	･水平･垂直性		･線順次法[2-23)]	水平･垂直線分

(1) ランレングス法

文書画像中の全ての線分を太らせていくと，文字線は図形中の線分より互いに密なので，早くつぶれが生じる．つぶれてできた連結黒画素群からは，文字列や行間の周期的な特徴 (行ピッチ) や，文字列の高さあるいは幅 (文字サイズ) といった特徴が抽出できる．そこで，これらの特徴が存在する領域をテキスト領域として分離し，さらにこれらの特徴を用いて文字列や一文字切り出しを行う手法が提案されている．

例えば，文字列中の白ランレングスが短くなることに着目し，あるしきい値より短い白ランを塗りつぶすRLSA (Run-Length Smoothing Algorithm)[2-12]，文字列を連結黒画素群として抽出するため，3×3 のマスクフィルタを用いたボカシ法[2-14]，あるいは弛緩法を適用した手法[2-15]などがある．

(2) 周辺分布法

テキスト領域中の文字列や行間は，周辺分布 (黒画素の投影加算) を用いることにより，周期的な山と谷として抽出可能である．周辺分布を用いた手法[2-17]としては，まず紙面全体の周辺分布を紙面の横方向・縦方向にそれぞれ求め，文字列方向に一致する方向に求めた周辺分布には，急峻な山・谷が現れることに着目し，文字列の抽出が可能である．また，個々の文字列について文字列方向と直角方向に周辺分布を求めることにより，文字と文字間に現れる周期的な山と谷に着目して一文字切り出しも可能である．

(3) 空間周波数法

テキスト領域の行ピッチや文字ピッチは，文書画像の濃度分布関数の空間周波数成分として捉えることができる．そこで，文書画像を 2 次元フーリエ変換することにより，文字列の位置や文字列数をフーリエスペクトルのピークから求めることが試みられた[2-18]．

(4) 連結黒画素法

一般に，文字は図形要素に比べて小さい．例えば，回路図やフローチャートに現れる図形要素は文字に比べ，長い線分が連結してできている．そこで，紙面内の連結した黒画素を単位としてラベル付けし，それぞれの大きさや形状によって分離することが考えられる[2-20]．特に，連結黒画素群に外接する矩形枠を設定し，この矩形枠の幅および高さがあるしきい値より小さいものを文字枠として抽出することが可能である．こうした手法は，フローチャートのように文字より小さい図形要素がほとんどない場合や，文字と図形の接触がない場合に有効である．

(5) 往復走査法

図形要素となる線分は，文字ストロークより長く，直線的であるといった性

質を用いた手法が提案されている[2-22]．紙面を水平方向，垂直方向それぞれ独立に往復走査することにより，紙面内の各黒画素に線要素の一部か，塊状要素の一部かを表す特徴量を付与し，しきい値によって文字 (塊状要素) と図形 (線要素) を分離する往復走査法が提案されている．

(6) 線順次法

また，論理回路図面等のように，処理すべき図形要素が水平線分および垂直線分からなる対象に対し，画像入力走査と並行して線分を整形化，コマンド化する手法として，線順次アルゴリズムが提案されている[2-23]．

2.2.6 文字認識

文字認識は，画像として入力された文字パタンを認識して文字コードに変換する処理である．紙面上個々の文字がどこにあるかを検出し，1 字ずつ分離する処理は文字切り出しと呼ばれ，文字・図形分離の説明でも触れた．紙に書かれた，あるいは印刷された文字の認識は，歴史的にも手書き文字認識と印刷文字認識に大別され，読み取り対象字種によってさらに英数字・カタカナ，ひらがな，漢字認識技術に分類される場合がある．また，印刷文字認識では単一書体文字認識と複数書体文字認識に分類される場合もある．本研究で対象とする文字認識技術は，先に述べた文書の性質上，複数書体の印刷漢字認識技術が中心であるが，現在多くのアルゴリズムが研究され実用化も進んでいるため，本研究においては適宜これらの技術を適用することとする．

2.2.7 図形認識

一方，図形認識は対象図形の特定が困難な場合が多く，これまで各種記号や設計図面等で用いられる定義図形を対象としたものが殆どである．基本的には線分の認識を基本としていることから，オフィスで扱われる文書で頻出する表罫線等の認識に適用できるものと考えられる．

2.2.8 画像圧縮符号化

画像の圧縮符号化については，現在多くの標準的な符号化方式が用いられている．文書画像に関してもっとも馴染み深い符号化方式として，ファクシミリ通信の国際標準として用いられた MMR 符号化方式が想定される．

2.2.9 文書フォーマット変換

文書フォーマット変換は，文書画像認識処理の過程で得られた文字情報，図形情報，イメージ情報を先に述べた ODA や SGML 等の文書記述形式に変換することにより，文書のレイアウト構造や意味構造と合わせて記述することであ

る．これまで，文書画像認識により得られた認識結果をもとにこれらの標準形式に変換した研究事例は少なく，文書画像認識により得られる電子化文書と予め電子的に作成された文書を統一的に扱う際の課題等についての検討は未着手である．

2.3 本研究の位置づけ

前節までで文書画像認識処理の基本となる従来技術を俯瞰した．これまで提案されている文字・図形分離技術は，処理対象に依存したケース・スタディが多く，処理対象を拡張し，オフィスの種々の文書に対処することがさらに必要である．また，印刷文書を自動読み取りし，電子的に作成された文書と同様に扱い得る形態に変換するためには，レイアウト構造や意味構造の認識も未だ検討が不十分である．したがって，本研究においては第 3 章における文字・図形分離処理の高度化，第 4 章における一般文書を対象としたレイアウト構造の認識手法の実現，第 6 章における文書画像からの意味構造の認識手法の実現をそれぞれ試み，また第 5 章と第 7 章において第 4 章，第 6 章でそれぞれ述べた手法に基づく認識システムを構成し，実用性の評価を行うものとする．

第 3 章　文字・図形の分離

3.1　はじめに

多種・多様な書式の文書・図面を認識するためには，認識の前処理として，文字と図形を分離抽出することが重要である．本章では，着目する点の近傍の線密度を近接線密度として定義し，近接線密度の値によって文字と図形を分離抽出する手法を提案し，実験を行った結果について述べる．

具体的には，3.2 で本研究の処理対象および従来技術についで述べる．3.3 では新たに定義した近接線密度特徴について述べ，3.4 で文字・図形が混在した図表領域に対し，近接線密度を用いて文字と図形を分離抽出する手法を提案する．3.5 では文字・図形分離抽出実験，および分離抽出した文字・図形に対し文字認識, 図形認識を行った結果について考察し, 3.6 で今後の課題について言及する．

3.2　従来の技術

3.2.1 処理対象

オフィス内の文書・図面には，手書きやワードプロセッサによる報告書，決裁文書，伝票等の帳票，図表類，また活版印刷された書籍，雑誌等の出版物などがある．認識入力の対象は，最終的にはこれらオフィス内の全ての文書・図面であるが，その実現のためのステップとして，現在文字認識技術の確立している印刷漢字を含み，またオフィスの既存文書・図面の大半を占める印刷物を対象とすることが急務である．以下に本章で前提とする文書・図面の特徴を述べる．

① 文書・図面には，文字のみの文章領域，文字・図形の混在した図表領域，また写真等のイメージ領域があり，そのレイアウトは様々である．
② 文章領域の文字 (漢字) のサイズ，字体，印字ピッチは様々である．
③ 図表領域には，表の罫線と文字など，文字・図形が混在している．
④ イメージ領域には，写真，イラスト，サイン，印影，社章等がある．

3.2.2 従来手法

文書・図面中の文字と図形をそれぞれ文字認識，図形認識するためには，文字と図形を個々に分離抽出する必要がある．特に多種・多様な書式の文書・図面に対しては，書式に関する情報を用いずに文字・図形を分離抽出することが重要である．このための手法として，これまで，文書・図面の文章領域，図表領域，イメージ領域の分離抽出に文書・図面画像の周辺分布を用いたもの[3-1)]，黒/白画素のランレングスを用いたもの[3-2)]，画素濃度を用いたもの[3-3)]などが報告

されている．また文字認識のための文字抽出にも周辺分布を用いたり[3-1]，黒画素の連結成分を用いたもの[3-4]などが報告されている．一方，オフィスの文書・図面には図表領域が多く存在し，これらの領域では文字サイズ，文字ピッチが様々で，文字列の規則性がない．このため，これら報告されている文章領域の文字抽出を目的とした手法では，精度良く分離抽出することができない．

そこで，本研究では文字・図形が混在した図表領域を取り上げ，文字列の規則性や文字サイズ，文字ピッチ等の特徴に関わる情報を用いずに，文字・図形を分離抽出する手法について検討を行った．

3.3 近接線密度特徴

3.3.1 基本的な考え方

文字と図形を分離抽出するため，文字と図形の特徴の差異を抽出する．図表領域には，一般的に次のような特徴がある．

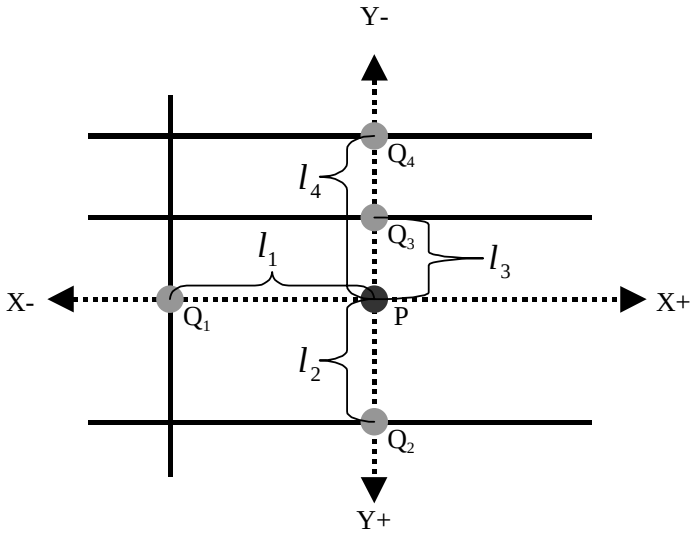
① 文字を構成する線分は互いに接近している．
② 文字同士はまた接近して文字列を成している．
③ 図形を構成する線分は文字を構成する線分に比べ互いに離れている．

これらを文字・図形の複雑さの特徴として着目し，その差異を抽出することが考えられる．

まず，文字・図形の複雑さを文字・図形を構成する線分の密度と定義し，入力文書・図面画像の各黒画素について，近傍の領域に存在する線分数を求める．具体的には，黒画素一点一点に着目し，着目画素を中心とする半径 r の領域を着目画素から周囲に向けて走査し，画素が白から黒に変わったところで線分を横切ったとして，着目画素の値に 1 を加算していく．この時，文書・図面中の文字列は通常横または縦に書かれていること，またメモリアクセスがし易いことを考慮し，図 3-1 に示すように，着目画素から走査する方向を入力画像の上下左右 (X+, X-, Y+, Y-) 4 方向とし，着目画素の線密度は 4 方向それぞれに求めた和と定義する．

図 3-2 に示す文書・図面を入力に用い，矩形で囲んだ領域の黒画素に対して線密度を求め，3 次元の高さで表した結果を図 3-3 に示す．但し入力は 8 dots/mm で行い，半径 r=512 とした．

図 3-3 から，線密度は文字領域上で高くなることが分かる．しかし，図形領域上においても文字列の上下左右に当たる領域で，文字列の影響を受けて高くなり，文字領域と図形領域の線密度の差異は必ずしも明確でない．



P : **文字または図形上の着目画素**
Q_1〜Q_4 : **近傍線分との交点(白から黒への変化点)**
l_1〜l_4 : P**点から**Q_1〜Q_4**点までの距離**

図 3-1　走査の方向
Fig. 3-1　Scanning directions.

そこで，線分が着目画素から遠くなればなる程，着目画素の複雑さは減少すると定義し，線分数の加算に際して重み付けを行う．重み付けに用いる関数には，着目画素から線分までの距離の単調減少関数の中から，着目画素の近傍の影響が大きい逆比例関数を用いる．また，求めた着目画素の複雑さを表す特徴量は，近傍の線密度を表すことから，近接線密度 (Neighborhood Line Density: NLD) と呼ぶ．図 3-1 における近接線密度の計算例を示す．

$$C_{X^+} = 0 \tag{1}$$

$$C_{X^-} = \frac{1}{l_1^{\alpha}} \tag{2}$$

$$C_{Y^+} = \frac{1}{l_2^{\alpha}} \tag{3}$$

$$C_{Y^-} = \frac{1}{l_3^{\alpha}} + \frac{1}{l_4^{\alpha}} \tag{4}$$

$$C = \sum \frac{1}{l_i^{\alpha}} \quad (i = 1, \cdots, 4) \tag{5}$$

$C_{X^+}, C_{X^-}, C_{Y^+}, C_{Y^-}$: **各方向に見た近接線密度**
α：**視野パラメータ**
C : P**点の近接線密度**

ここで視野パラメータαは，α≧0 と定義し，α=0 のとき C は線分数の和となり，図 3-3 に示した線密度と等価である．視野パラメータの性質については次節で詳しく述べることとし，α=1 の場合について，図 3-3 と同じ領域で近接線密度を求め，3 次元の高さで表した結果を図 3-4 に示す．

図 3-4 は図 3-3 に比べ, 文字領域上と図形領域上で近接線密度の差が顕著となっており，文字列の上下左右に当たる領域でも，図形領域上の近接線密度は低い．また, 文字領域上には近接線密度の高いピークが現れていることも分かる．このことから，文字領域と図形領域の近接線密度の差を抽出して，それぞれを分離することが可能である．

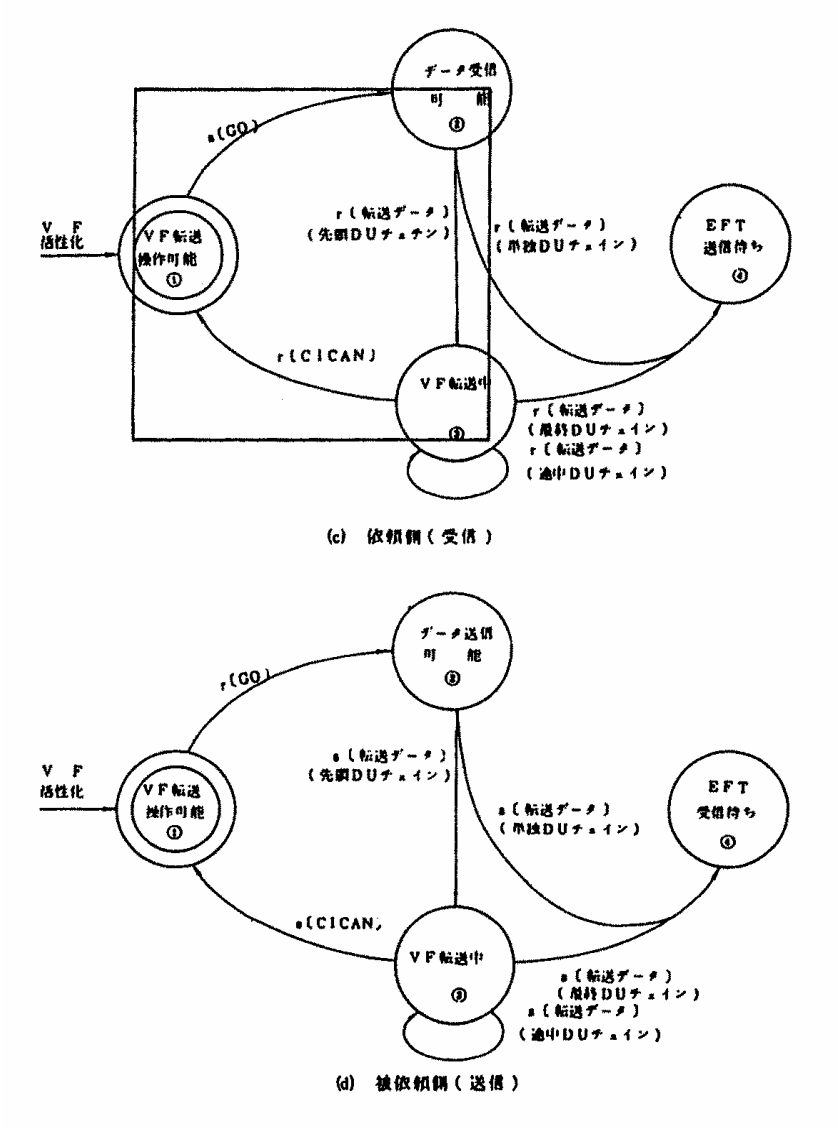


図 3-2　入力に用いた文書・図面の例

Fig. 3-2　Example of inputted document.

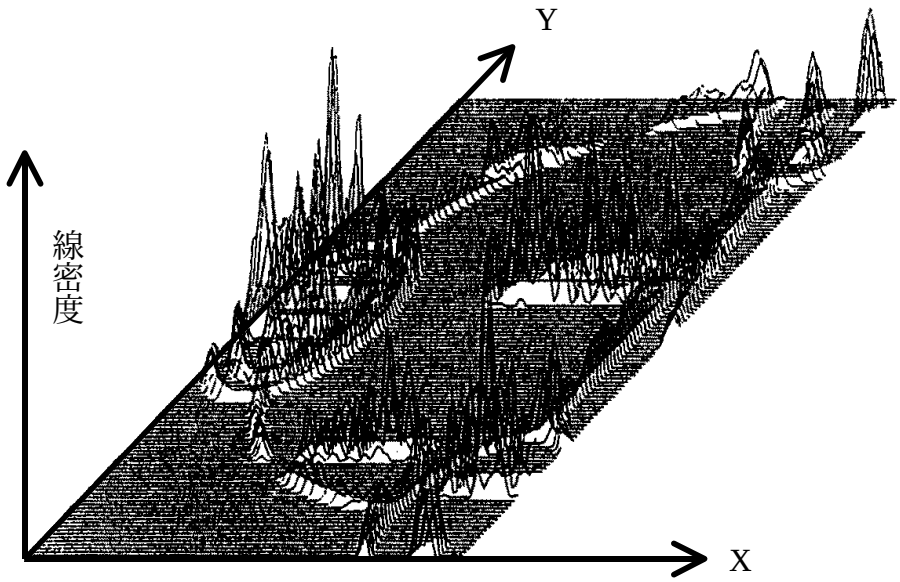


図 3-3　線密度の 3 次元表示

Fig. 3-3　3-dimentional presentation of line density.

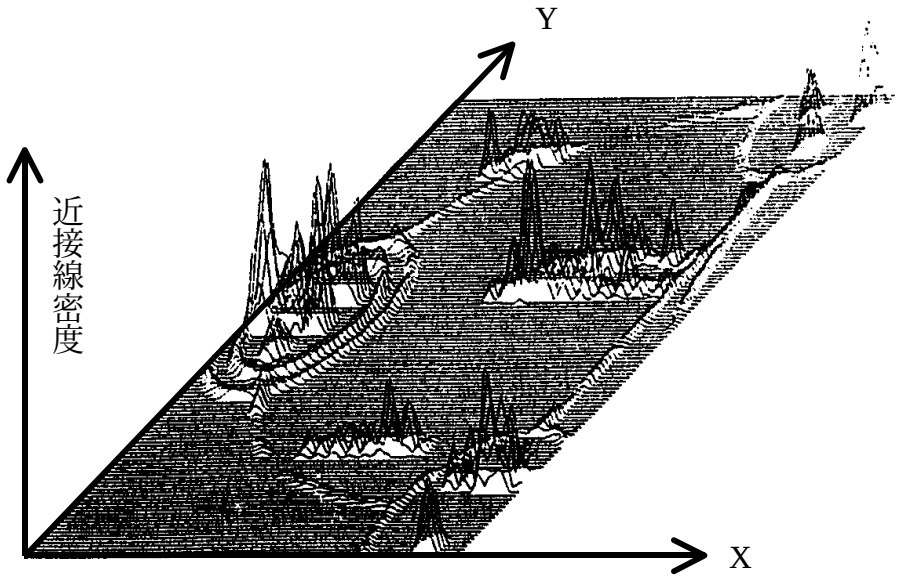


図 3-4　近接線密度の 3 次元表示 (α=1)

Fig. 3-4　3-dimensional presentation of NLD (α=1).

3.3.2　近接線密度特徴の性質

前節で定義した近接線密度が，実際に文字・図形分離抽出処理に有効であることを示すため，次の観点からシミュレーションを行い，近接線密度の性質を明らかにした.

(a) 視野パラメータと近接線密度分布の関係

重み付け関数の視野パラメータαの振る舞いを明らかにするため，αを変化させて近接線密度のヒストグラムを求めた．図 3-2 に示した入力画像を用い，αを 0.5，1.0，2.0 としてそれぞれ近接線密度を求め，そのヒストグラムを図 3-5 に示す．ここで，横軸はαのそれぞれについて近接線密度の最大値が 1，最小値が 0 となるよう正規化し, 縦軸はヒストグラムの総和が 100%となるよう正規化した.

その結果，αを大きくすると近接線密度の分布は小さい方に偏り，小さくすると平坦化することが分かる.

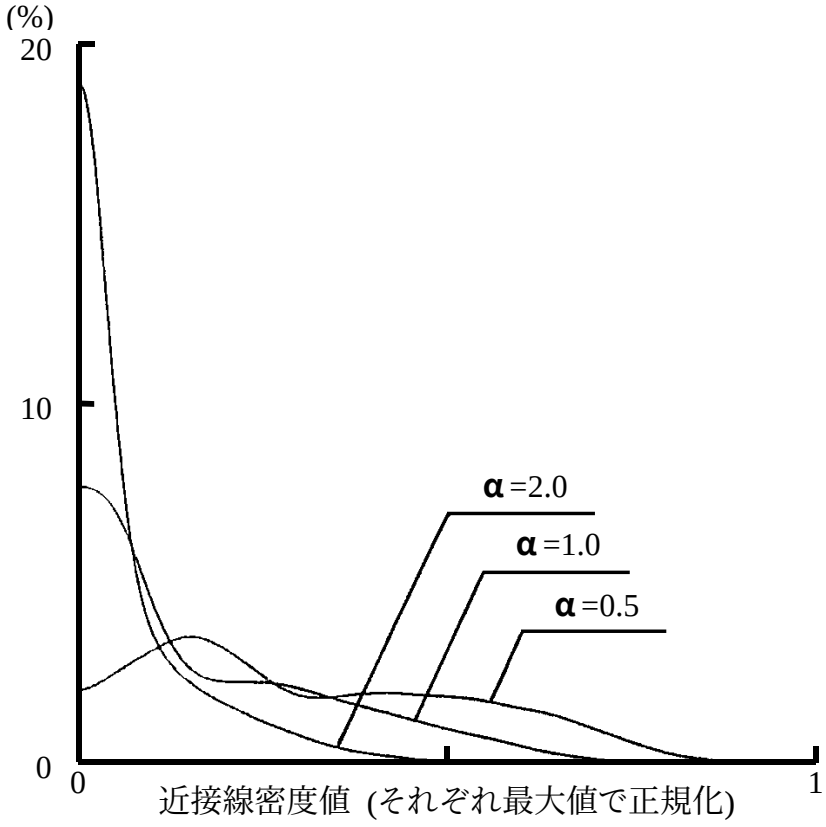


図 3-5　近接線密度のヒストグラム

Fig. 3-5　Histogram of NLD.

視野パラメータを小さくすると重み付けが平坦化し，遠くの線分まで近接線密度に影響するため，視野を拡大することになる．α→0 では，図 3-3 のように文字領域に近い図形領域上の画素においても近接線密度が高まり，分布が平坦化したものと考えられる．一方，視野パラメータを大きくすると，近傍の線分のみが近接線密度に影響するため，視野を狭めることになる．この時，2 値入力画像の量子化誤差によるノッチの影響が顕著となり，ノッチによって一部の画素の近接線密度が高まり，分布は逆に小さい方に偏ったものと考えられる．したがって，近接線密度を文字・図形分離抽出に用いるためには，入力画像中の文字と図形の接近の程度により，視野パラメータを最適化する必要がある．

(b) 文字位置と近接線密度のピーク点の関係

前節で述べたように，文字領域上には近接線密度の高いピークが現れている．このピーク点の性質を調べた．まず近接線密度を求めた領域に対し，3×3 のマスクを用い，中心画素の近接線密度が周囲の 8 画素より高いピーク点を抽出する．さらに求めたピーク点に対してしきい値処理を行い，しきい値より高いピーク点を抽出する．その結果，しきい値を高い方から順に下げていくと，ピーク点は最初文字領域上に現れ，次第に図形領域上にも現れるようになることが分かった．このことにより，しきい値によって，文字領域上のピーク点と図形領域上のピーク点に分離することが可能である．しきい値に領域内の近傍線密度の平均値を用いた場合について，ピーク点を抽出し，ピーク点を中心とする小円を描くことによりピーク点の位置を表した結果を図 3-6 に示す．

この結果，文字領域上でピーク点が抽出されることが分かる．また文字列の端でピーク点が抽出されないところが残り，しきい値をさらに下げると，これら文字領域のほかに，図形領域上でもピーク点が抽出され，分離誤りが生じることが分かった．

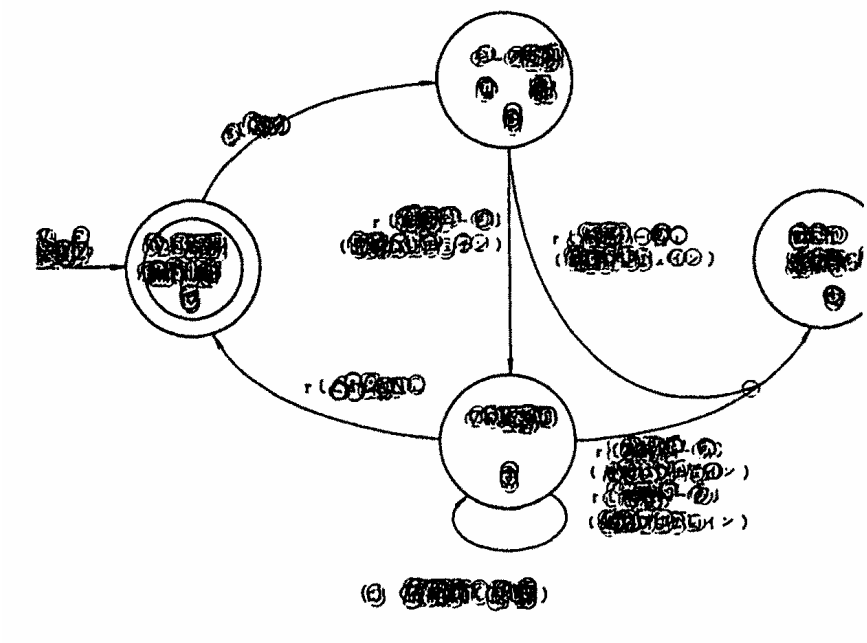


図 3-6　近接線密度のピーク点
Fig. 3-6　NLD high peaks.

(c) 文字サイズと近接線密度の関係

文書・図面に含まれる文字サイズが小さくなれば，文字を構成する線分の間隔が狭まり，近接線密度はより高くなると考えられる．そこで，文字サイズの例として図 3-7 に示すファクシミリテストチャート No.2 を用い，各サイズの文字領域について近接線密度を求め，その平均値を図 3-8 にプロットした．

その結果，視野パラメータが同じときは，文字サイズが小さくなれば近接線密度が高くなることが分かった．視野パラメータ α =1 のとき，文字サイズと近接線密度の平均値の間で，次のような実験式を得た．

$$S = \exp(1.9 - \frac{\ln(N)}{1.33}) \qquad (6)$$

S：文字サイズ(ポイント)
N：近接線密度の平均値

このため，文字サイズが一様な領域で，近接線密度の値から文字サイズを推定することができる．

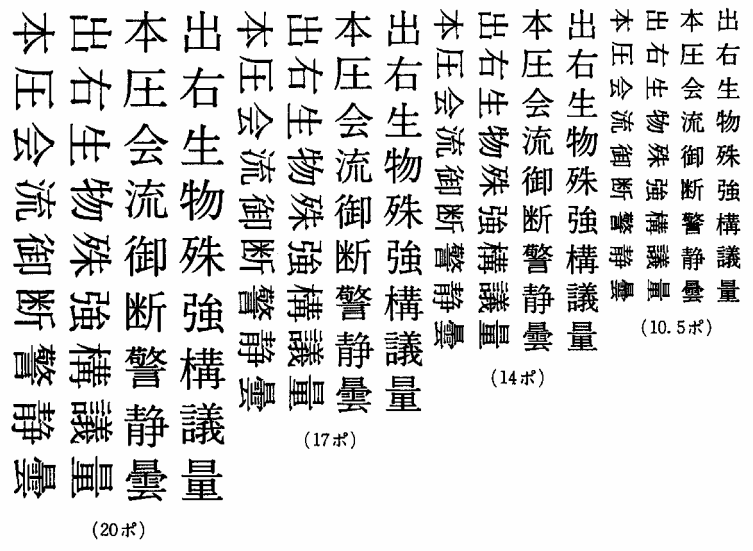


図 3-7　文字サイズの例

Fig. 3-7　Examples of character size.

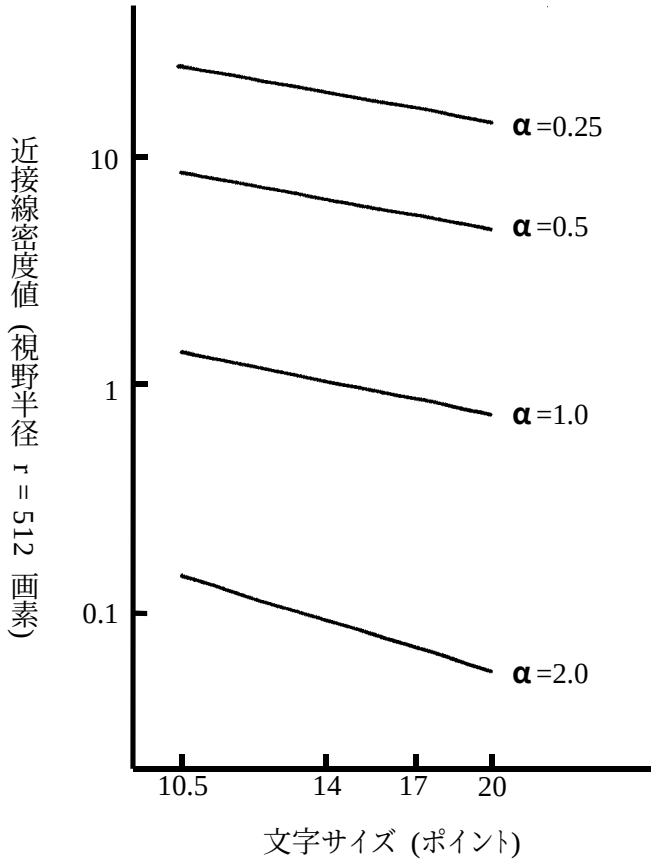


図 3-8　文字サイズと近接線密度の関係

Fig. 3-8　Relation between character size and NLD's value.

3.4 文字・図形分離処理

本節では，これまで示した近接線密度を用い，文字・図形を分離抽出する処理の流れについて述べる．処理対象として，図表領域のみからなる文書・図面，あるいは周辺分布法[3-1)]などにより分離抽出された図表領域を用いる．以下に処理の流れに沿って，その概要を述べる．

(1) 近接線密度の算出とピーク点抽出

入力画像の全黒画素について近接線密度を求める．次に3×3のマスクを用い，近接線密度のピーク点を抽出する．さらに近接線密度の平均値をしきい値として用い，平均値より高いピーク点のみを抽出する．

(2) ピーク点のグルーピング

文字は複数個集まって文字列を成し，同じ文字列内ではサイズが一様と仮定する．そこでピーク点を文字列ごとにグルーピングする．グルーピングにはピーク点の座標を用いる．例えば，横書きの文字列上のピーク点は，入力画像のY座標 (縦方向) の差がしきい値β (文字の高さに相当) より小さいグループとして，グルーピングする．

(3) 文字サイズの推定と文字領域の分離抽出

グループごとにピーク点の近接線密度の平均値を求め，当該文字列の文字サイズを推定する．文字サイズの推定には式 (6) を用いた．各ピーク点に対し，推定した大きさの文字に外接する矩形枠を求め，その領域内を文字領域として分離抽出する．

(4) 一文字切り出し

OCR 等の文字認識技術を活用して文字認識を行うため，文字・図形分離抽出結果から，一文字単位に切り出す必要がある．そこで，先に抽出した文字領域の外接矩形枠を用いる．一文字を正確に切り出すため，矩形領域内の黒画素の連結性に着目し，矩形枠の各辺が黒連結に外接するようにして切り出す．

(5) 後処理

後処理として，次の処理を行う．

① 一文字上に複数のピーク点が存在する場合は，一文字切り出しの際，一つに絞る．

② 図 3-6 に示したように，文字列の端でピーク点の現れない文字領域を抽出するため，文字列の延長上で，当該文字列の文字サイズと同程度の大きさの黒連結を，文字として切り出す．

③ 図形領域上に現れたピーク点からは，図形の一部が文字として誤って抽出される．これらの領域は文字列から孤立しており，黒画素が切り出し枠の外へ連結しているため，この条件を満たせば除去する．

3.5 実験と考察

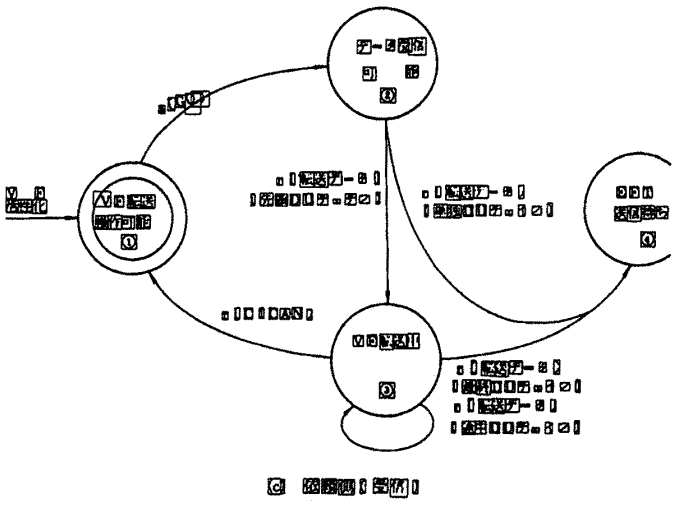
3.5.1 文字・図形分離抽出実験

文字・図形分離抽出実験に，表，グラフ，フローチャート，ブロック図等を含む印刷物を用いた．これらをファクシミリ装置より，16 dots/mmの解像度で入力し，3.4 で述べた処理を行う．また分離抽出結果を評価するため，一文字切り出しした文字，および残りの図形領域を，それぞれ文字認識，図形認識する．文字認識にはマルチフォント印刷漢字認識装置[3-5]を用い，図形認識には線順次アルゴリズム[3-6]を用いて，水平，垂直線分を抽出した．これにより，文字・図形分離抽出エラーは，文字認識エラー，線分抽出エラーとして評価することとした．

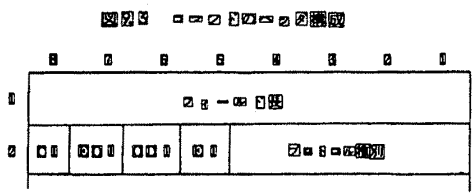
以下，文字・図形分離抽出実験に

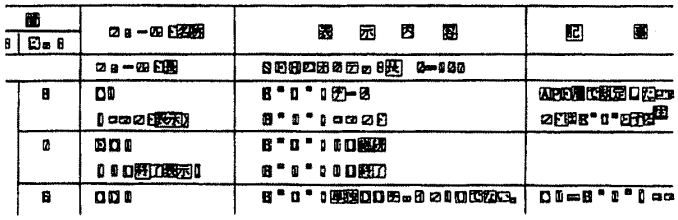
(a) 自由線分を含む図 (状態遷移図)
(b) 和文タイプされた表

を用いた結果を図 3-9 に示す. さらに図 3-9(b) について, 線分抽出結果を図 3-10 に，文字認識結果を図 3-11 に示す．但し，図 3-9，3-10 については入力画像を 8 dots/mm に間引いて処理した結果を示し，図 3-11 については 16 dots/mm の解像度のまま文字認識した結果を示している．



(a)





(b)

 : extracted characters

図 3-9　文字・図形分離抽出の結果

Fig. 3-9　Result of character/graphic segmentation.

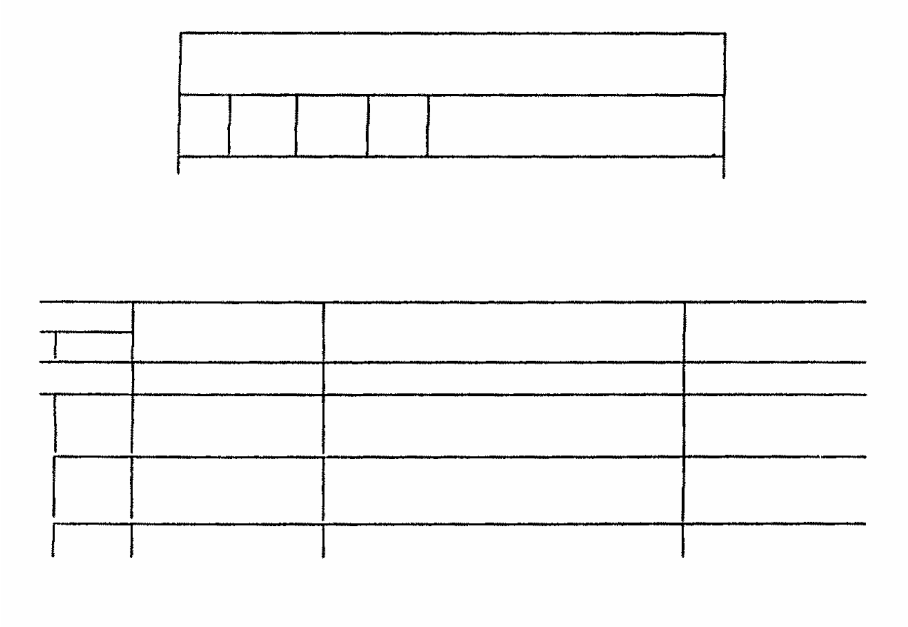


図 3-10　線分抽出結果の例
Fig. 3-10　Result of line extraction.

図9.3コマンドのヘッダ構成
図？５コマンドのヘッダ構成

図9.4システム機能ヘッダ共通部の
図？４システム機能ヘッダ共通部の

フィールド名称表示内容記事
フ？ールド名称表示内容記事

コマンド表示
コマンド？泳

図 3-11　文字認識結果の例
Fig. 3-11　Result of character recognition.

3.5.2 実験結果の考察

(1) 文字の分離抽出

分離抽出した文字を文字認識した結果，良好な認識結果が得られた．一方，リジェクトや誤読等の文字認識エラーも生じ，その原因には文字・図形分離抽出エラーに起因する場合と，文字認識単体による場合があることが分かった．ここでは，分離抽出エラーに起因する場合について考察する．

① 一文字切り出しで，句読点等を含んで切り出した．
② 文字ピッチの狭いところで，隣接する文字の偏や旁を含んで切り出した．
③ 文字と文字の接触，文字と図形の接触により，黒連結の外接枠が抽出できず，正確に一文字が切り出せなかった．
④ 文字に近接している図形の一部を，誤って切り出した．

これらについては，例えば黒連結の大きさ，形状，連続性等にも着目すれば，正確に一文字切り出しが可能と考える．

(2) 図形の分離抽出

本研究では，文字を分離抽出した残りの領域に対し，水平・垂直線分抽出を行い，表等の罫線を抽出した．その結果，

① 図形の一部が誤って文字として抽出された場合，抽出した線分に切れが生じる，
② 文字が誤って残った場合，短い線分が多数抽出される，

などが分かった．これらについては，線分の始終点間の距離が近ければ接続する，また短い線分を除去するなどにより対処することが考えられる．

3.5.3 今後の課題

近接線密度は，文字・図形が混在した図表領域の文字・図形分離抽出に有効な特徴であることが分かった．一方，入力画像中の文字と図形の接近の程度と視野パラメータの関係を明らかにし，分離抽出精度を向上させること，量子化ノイズの影響を明確にすること，また，近接線密度の抽出処理を高速化することなどが，今後の課題と考える．以下，これらの解決策について述べる．

(1) 文字認識結果，図形認識結果のフィードバック

文字認識の結果リジェクトとなったり，線分抽出の結果短い線分が多数抽出される場合，文字・図形分離抽出エラーが考えられる．そこで，リジェクト文字等の位置情報をフィードバックし，再抽出することが考えられる．例えば，近接線密度を視野パラメータを変えて求めるとか，すでに認識した文字のサイ

ズを切り出し枠に反映するといった処理が，分離抽出精度向上に有効と考えられる．

(2) 量子化ノイズの除去

視野パラメータを大きくすると，量子化ノイズによるノッチが近接線密度に強く影響する．そこで入力画像のスムージングを行い，ノッチを予め除去する方法や，近接線密度の算出の際，ノッチ成分をカウントしないなどの方法が考えられる．

(3) 近接線密度抽出の高速化

近接線密度の抽出に，入力画像の全黒画素について周囲 4 方向を走査するため，処理時間を多く要する，そこで，入力画像を横および縦にそれぞれ往復走査することにより近接線密度を算出する．また重み付け関数をテーブル化するなどして，処理時間を短縮することが考えられる．

3.6 まとめ

本章では，オフィスの多種・多様な書式の印刷文書・図面を認識入力するため，文字・図形の混在した図表領域を対象として，文字と図形の複雑さの違いに着目し，それぞれを分離抽出する手法を提案し，実験を行った結果について報告した．その結果，以下のことが明らかとなった．

① 文字・図形が混在した図表領域では，文字領域上に近接線密度の高いピークが現れ，このことが，文字と図形を分離抽出する特徴パラメータとして有効である．
② 文字領域の近接線密度の値から，文字サイズが推定できる．
③ 視野パラメータを変えると，視野を拡大したり，狭めたりする．大きくすると，近接線密度は入力画像の量子化ノイズの影響を強く受ける．
④ 文字・図形分離抽出結果を，文字認識，図形認識した結果，良好な認識結果が得られた．一方，分離抽出エラーに起因する認識エラーが生じた．

オフィスで扱われる文書・図面を認識入力するため，本章で述べたの図表領域の文字・図形分離抽出処理と，文章領域やイメージ領域の処理，および文字認識，図形認識を結合する必要がある．さらに認識入力の対象を，印刷物から手書きへと拡張することも必須である．手書きの文書・図面の入力に対しては，文字サイズ，文字ピッチが変動する対象として，本手法の有効性を確認することが考えられる．

第 4 章　文書画像のレイアウト構造認識

4.1　はじめに

オフィスで扱われる文書は文字情報のみならず，表や図等の図形情報，写真等のイメージ情報を含んでおり，文書中の文字・図形の認識に先立ってこれらを分離抽出する必要があることは前章で述べた．これまで，OCRに代表される文字認識技術や，さらに図形認識技術を活用し，文書の書式や文字・図形情報を認識して，文書画像を電子ファイルに適したコード情報に変換する文書認識システムの検討[4-1), 4-2)]が進められる中，処理対象を新聞や論理回路図面，帳票等に限定し，入力文書・図面に固有の性質に着目して文字・図形を分離抽出し，文字認識，図形認識する試みもなされている[4-3)~4-6)]．しかし，オフィス文書の性質は多様であり，特定の性質に着目したこれまでの手法をそのまま適用しても，十分な認識結果が得られないという問題があった．

そこで，本章では，オフィス文書中の文字・図形に関わる本質的な特徴として周辺分布特徴，黒連結特徴，近接線密度特徴を抽出し，これら複数の特徴を併用して文字・図形を分離抽出することにより，オフィス文書を認識する手法を提案する．以下，4.2 で処理対象とする文書の性質と従来技術について述べ，4.3，4.4 で処理アルゴリズムについて述べる．また，4.5 では提案したアルゴリズムの評価実験を行った結果について考察する．

4.2　文書の性質と従来の技術

オフィスには，新聞，雑誌，マニュアルといった印刷物や手書きの原稿，伝票等が大量に存在し，これらを電子ファイル化する際の入力の省力化が問題となっている．これらの文書は，一般に，本文・見出し等のテキスト情報，表罫線等の幾何図形情報，写真・イラスト等のイメージ情報から成るマルチメディア文書であり，しかもレイアウト等の書式もバラエティに富んでいる．文書認識処理の目的は，こうしたマルチメディアの文書を編集処理に適したコード情報に変換して電子ファイル化することにある．すなわち，種々の書式の文書から書式情報を抽出し，さらに文字，図形情報を分離抽出してそれぞれ文字認識，図形認識し，文字コードや図形コマンドに変換する．

オフィス文書の一般的な構成を図 4-1 に示す．文書中には，文字のみから成るテキスト領域，文字と表罫線から成る表領域，文字と種々の図形から成る図領域，文字・図形以外のイメージ領域がある．これらの領域に含まれる文字・図形をそれぞれ分離抽出するには，まず各領域の性質に着目して領域抽出を行った後，文字・図形の性質に着目してそれぞれを分離抽出する必要がある．

これまで，これら領域の分離抽出に文書画像の周辺分布を用いたもの[4-7]，黒/白画像のランレングスを用いたもの[4-8]，画素濃度を用いたもの[4-9]などが報告されている．また，文字・図形の分離抽出に黒画素の連結成分を用いたもの[4-10), 4-11)]，近接線密度を用いたもの[4-12)]などが報告されている．文書を構成するテキスト，表，図，イメージ領域は，一般的には紙面上で独立した領域であるが，例えば，表領域の内部に図領域やイメージ領域が混在する場合もあり，文書の構成はより複雑である．このことにより，これまで報告されている領域抽出手法や，文字・図形の分離抽出手法を個別に適用しても十分な処理結果が得られなかった．そこで，4.3 では，こうしたオフィス文書を認識するための処理アルゴリズムとして，文書画像中の文字・図形の本質的な特徴である周辺分布特徴，黒連結特徴，近接線密度特徴を組み合わせて用いる手法を提案する．ここで，図 4-1 に示した文書の性質をより具体化するため，本アルゴリズムで処理対象とする文書について，その入力条件，各領域の性質，文字および図形の認識対象を表 4-1 のように設定する．

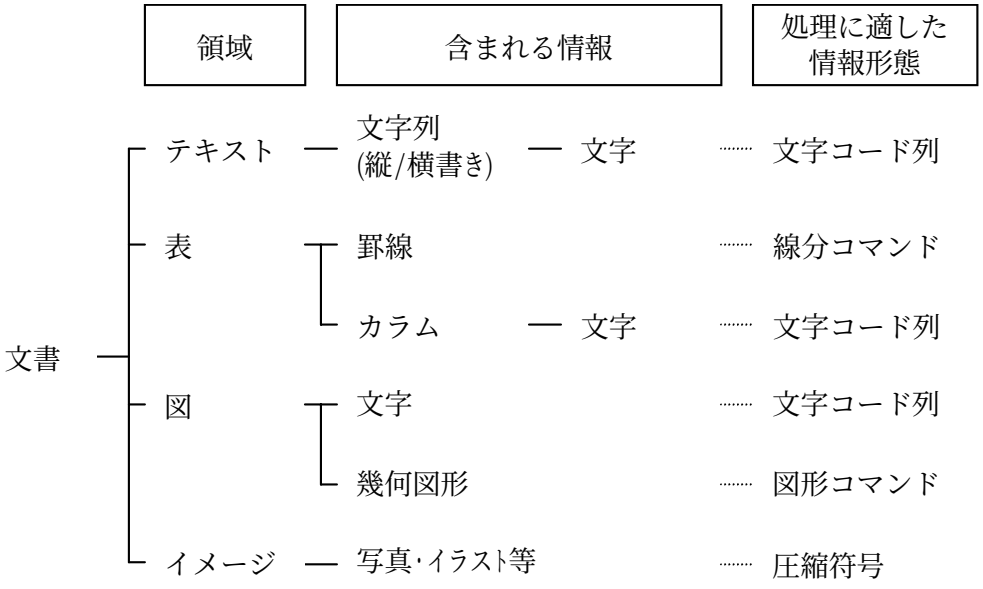


図 4-1　文書の構成

Fig. 4-1　Regions of typical document.

表 4-1　処理対象とする文書の性質

Table 4-1　Features of documents.

入力条件	階調，カラー	白黒 2 値
	印字品質	比較的良いもの (新聞は困難な場合がある)
文書の構成	テキスト領域・表領域・図領域・イメージ領域混在	
	枠付き/枠無し領域混在	
領域の性質	テキスト領域	縦書き・横書き，1～2 段構成
	表領域	表罫線と文字より構成
	図領域	文字・図形混在，文字と図形の接触なし
	イメージ領域	2 値画像
文字認識対象	文字種	英字・カナ・漢字，印刷文字・手書き文字
	文字サイズ	7～24 ポイント
	文字ピッチ等	可変
図形認識対象	図形要素	線図形 (直線)

4.3　レイアウト構造認識の処理フロー

文書認識処理は，前処理，領域抽出処理，テキスト領域・表領域・図領域の認識処理，および後処理に大別される．処理フローを図 4-2 に示す．前処理では，入力された文書画像の雑音除去，傾斜検出，縦書き/横書き判別を行う．領域抽出処理では，画像としてまとまりのある矩形領域を抽出するとともに，内部に図形を含まない領域をテキスト領域と判定する．テキスト領域については，テキスト領域の認識処理で文字列抽出，一文字切り出し，文字認識を行い，文字コード列に変換する．内部に図形を含む領域については，まず，表領域の認識処理を行って表領域かどうかを判別し，表領域でない場合は図領域の認識処理を適用する．図領域の認識処理で認識不能な領域は，認識対象とする文字および図形を含まない領域であり，イメージ領域として処理する，なお，縁取りのある文書や，表の中にさらに表や図を含む文書についても，表領域の認識処理以降を再帰的に適用することで認識可能である．後処理では，各領域の認識結果を文書の編集等に用いるため，各々の処理で抽出した書式情報に基づいて認識結果の編集を行う．以下に，処理の順に従ってアルゴリズムの詳細を述べ，実際の処理例についても言及する．

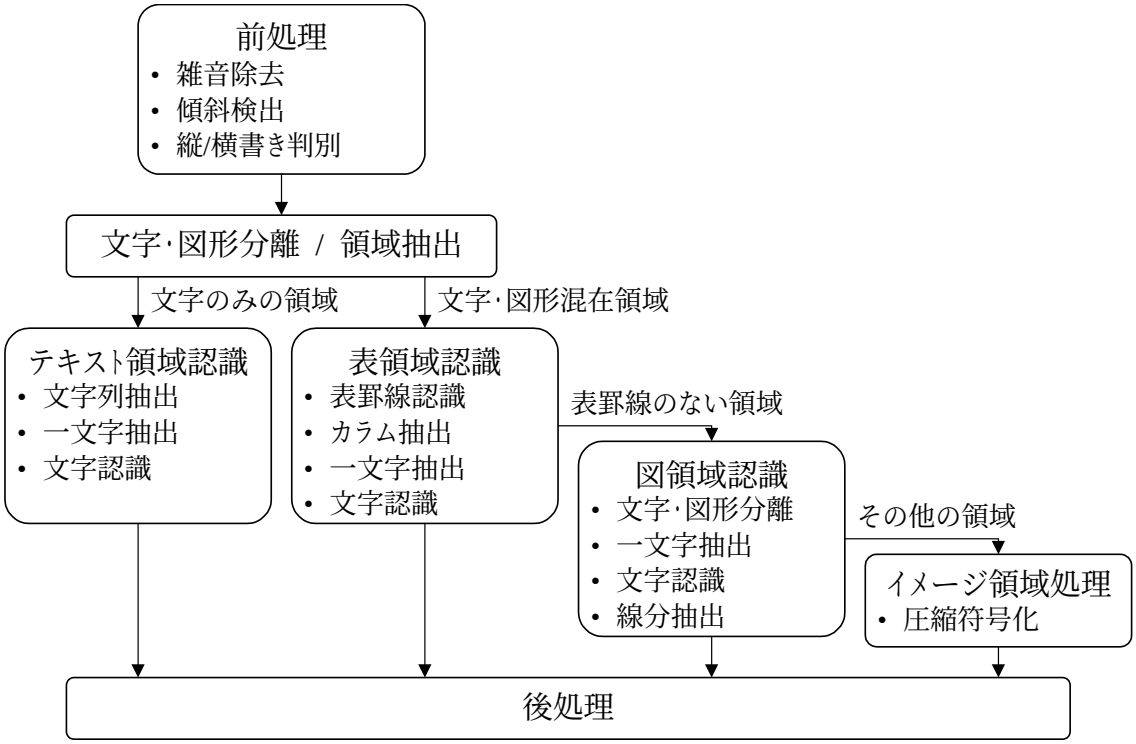


図 4-2　文書認識処理のフロー

Fig. 4-2　Flow diagram of document recognition.

4.4　アルゴリズムの詳細

4.4.1　前処理

(1) 雑音除去

2 値画像として入力する際混入した孤立点等の雑音や量子化誤差によるノッチを除去する.

(2) 傾斜検出と縦/横書き判別

文書が傾いて入力されると，各領域の性質が正しく抽出されないばかりでなく，文字認識率も低下するため，傾斜検出を行って一定角以上傾いた文書をリジェクトする．また，文書には縦書き，横書きがあり，それぞれ文字を読み取る方向など性質が異なるため，これを判別しておく必要がある.

文書画像を文字列の方向に投影した周辺分布には，文字列と行間の特徴が周期的に現れる. そこで入力文書の縦方向, 横方向の周辺分布をそれぞれ抽出し, 周期性を求めて比較することにより, 文字列の方向を知ることができる. 一方, 文書が傾斜すると，周辺分布上に文字列の特徴が正しく現れなくなる．この場

合，文書を縦または横に短冊状に分割して周辺分布を抽出し，文書の傾斜によって周辺分布上の文字列の位置がずれることに着目して，傾斜量を求める．図 4-3 は入力文書画像の例で，図 4-4 に縦および横にそれぞれ 2 分割して求めた周辺分布の様子を示す．次に図 4-4 を用いて，文字列の周期性と傾斜量を求める方法について述べる．

文字列方向に求めた周辺分布上には，文字列部分で値が大きくなるところと，行間部分で小さくなるところが周期的に現れるため，(1) 式あるいは (2) 式に示すような短冊状に分割した周辺分布の積和 FH，FV を求めると，文書の傾斜の度合いに応じて変化する．

横方向に見た周期性を示す特徴量

$$FH = \sum_i (\text{左領域のi番目の周辺分布値}) \times (\text{右領域のi番目の周辺分布値}) \quad (1)$$

縦方向に見た周期性を示す特徴量

$$FV = \sum_i (\text{上領域のi番目の周辺分布値}) \times (\text{下領域のi番目の周辺分布値}) \quad (2)$$

すなわち，文書の傾斜がないとき，文字列の方向に見た特徴量の方が大となる．一方，文書が傾斜すると FH，FV は傾斜のないときに比べ小さくなるため，FH および FV がそれぞれ最大となるよう，文字列の傾斜によるずれ量 Δi だけ一方の周辺分布をシフトして積和を求める．この FH，FV の最大値を用い，

FH＞FV のとき，横書き
FH＜FV のとき，縦書き

とする．横書きの場合は FH を，縦書きの場合は FV を最大とする Δi は文書の傾斜の度合いを示しており，あるしきい値より大きいとき，入力文書の処理を中止する．図 4-4 を用いて FH，FV を計算した結果を，FH，FV のそれぞれの最大値で正規化して図 4-5 にプロットした．図 4-5 では，FH(0) ＞ FV(-2) となり，FH を最大とする Δi は 0 であることから，図 4-3 は傾斜のない横書き文書であることが分かる．図 4-5 から，文字列方向の周辺分布について求めた特徴量は Δi の変化によって大きく変化することが分かり，このダイナミックレンジを比較することによっても文字列方向を知ることができる．

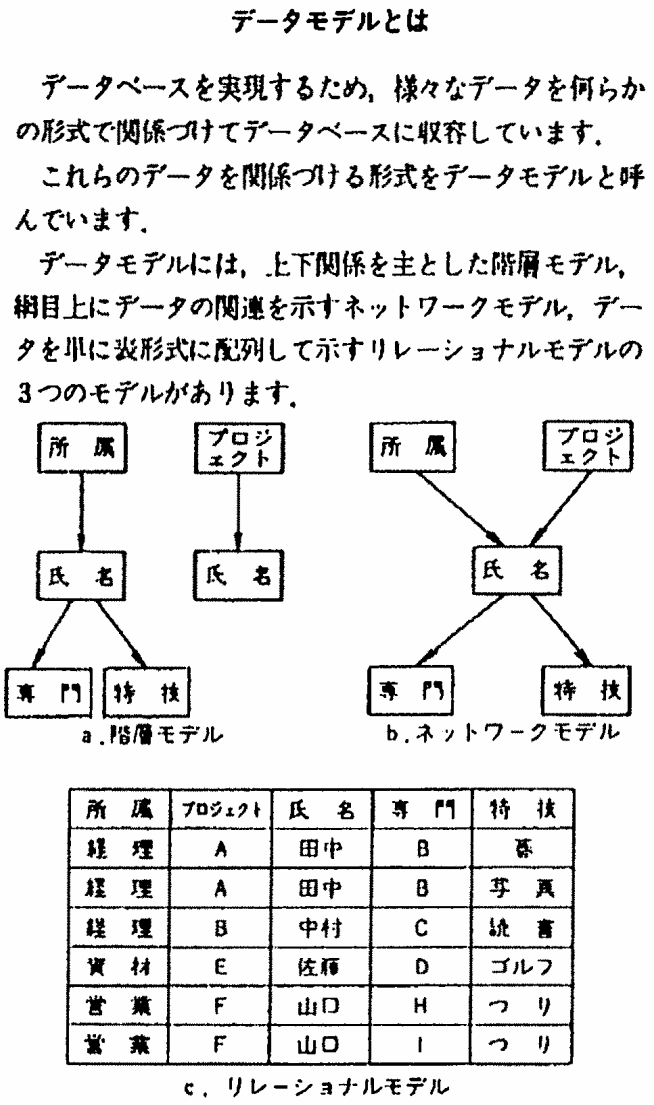
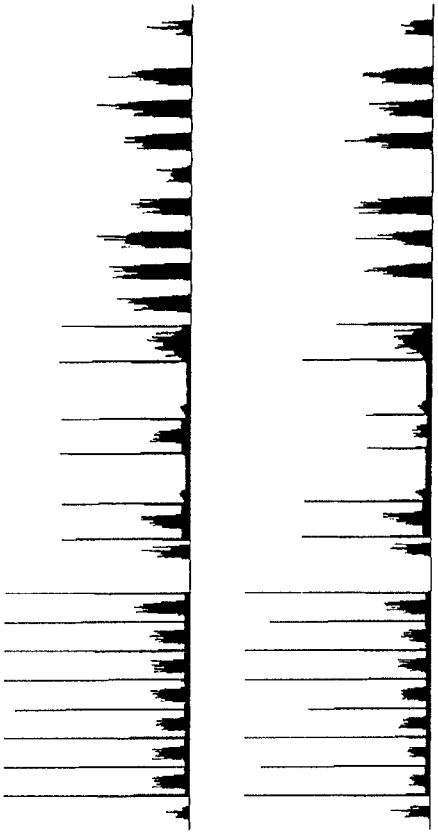


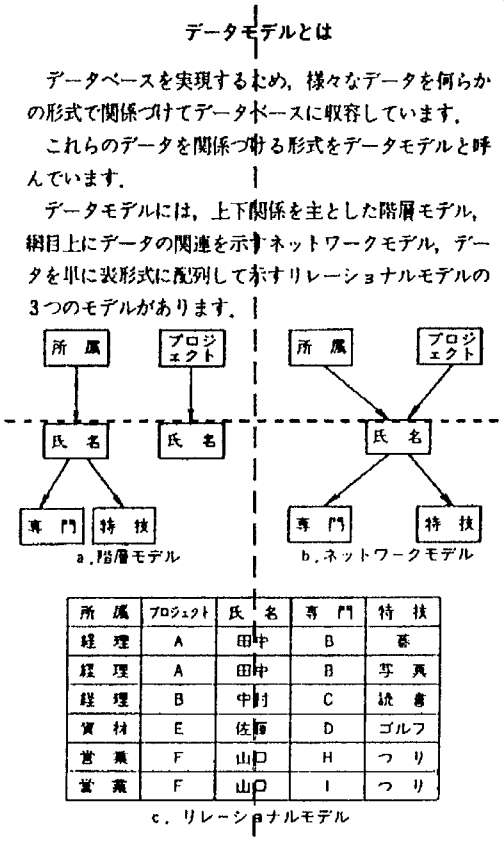
図 4-3　入力文書画像の例

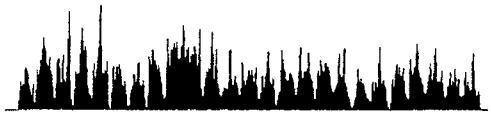
Fig. 4-3　Example of inputted document image.



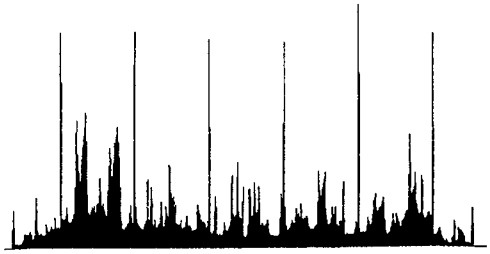
(a) 左領域･横方向の周辺分布

(b) 右領域･横方向の周辺分布





(c) 上領域･縦方向の周辺分布



(d) 下領域･縦方向の周辺分布

図 4-4　周辺分布の例

Fig. 4-4　Examples of projection profiles.

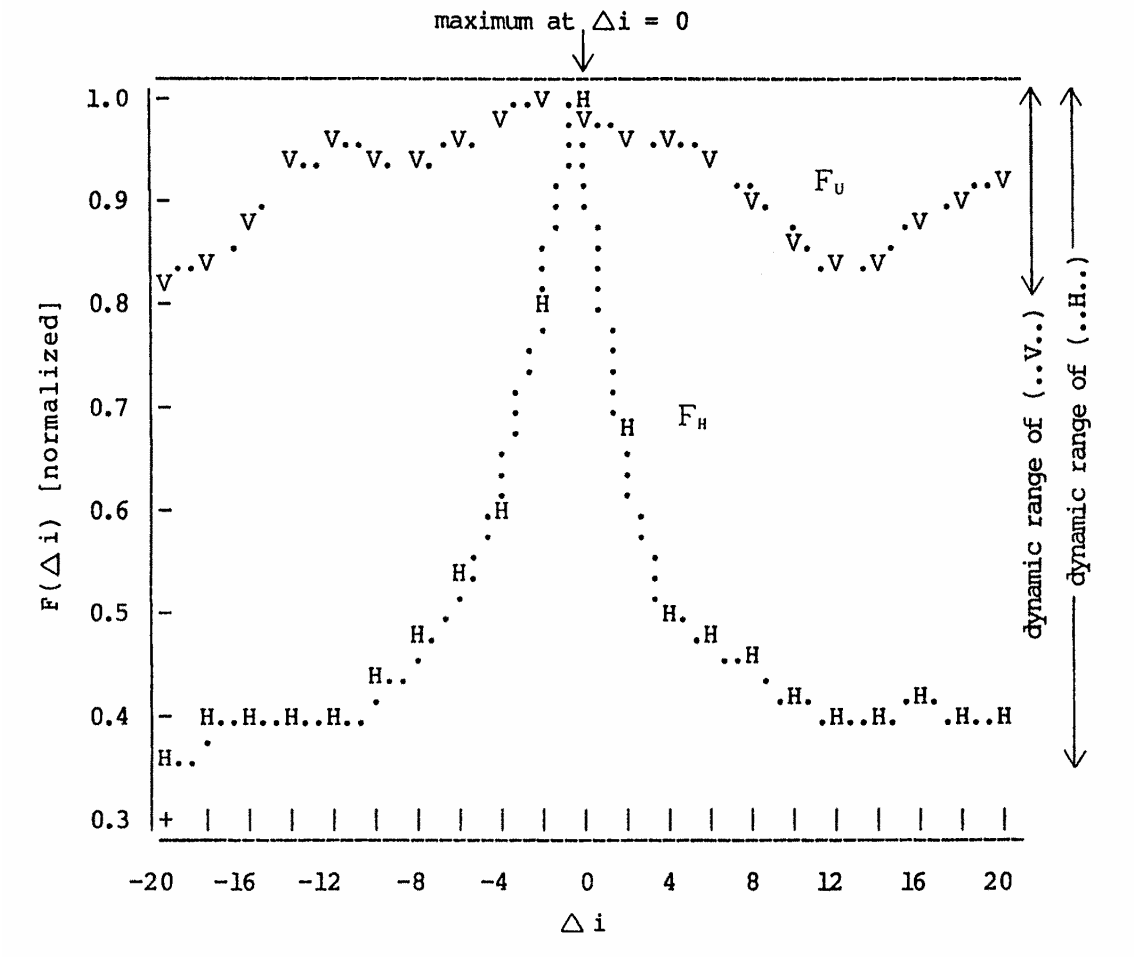


図 4-5　図 4-4 の周辺分布を用いた計算結果

Fig. 4-5　Sketch of calculation results obtained from projection profiles in Fig. 4-4.

4.4.2　領域抽出処理

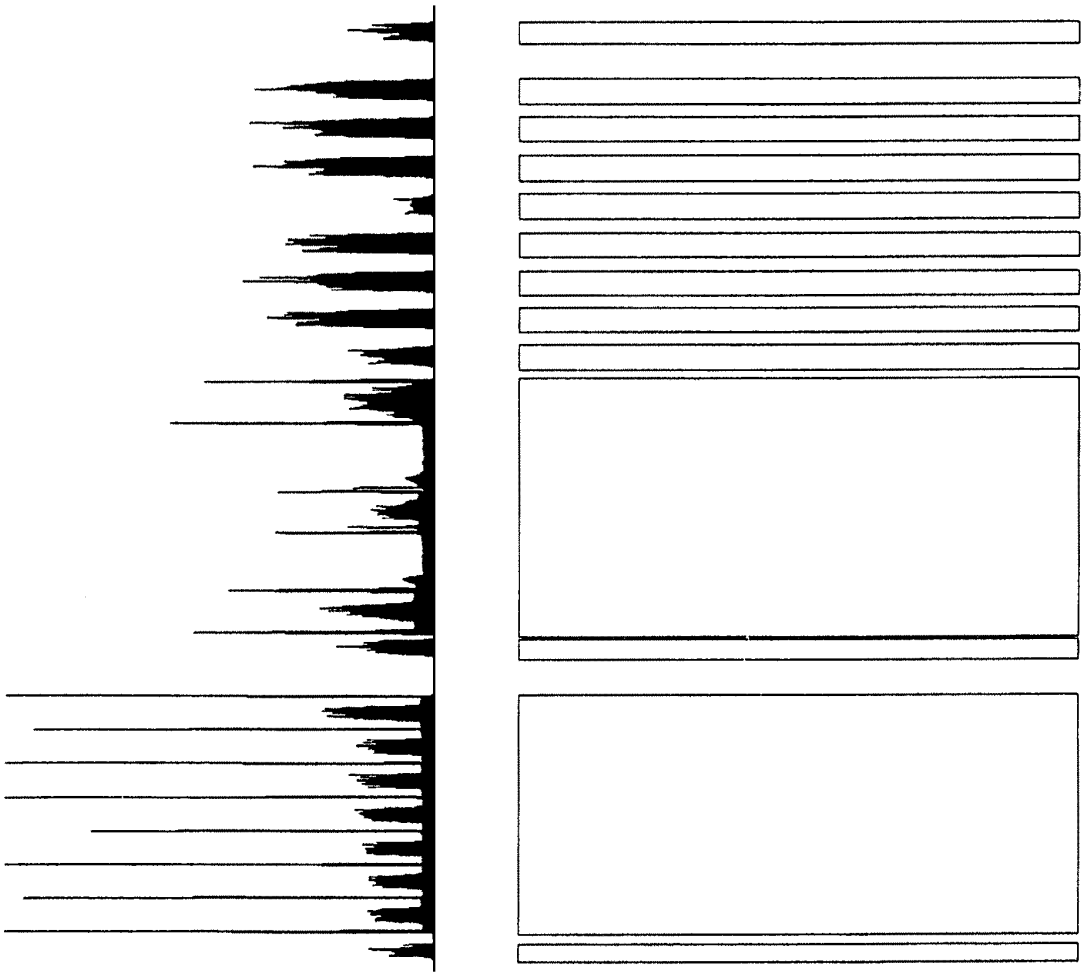
領域抽出処理では，文書画像中のテキスト領域，表領域，図領域といった画像としてまとまりのある矩形領域を抽出する．領域抽出法として，従来文字や文字列の間隙に着目してそれぞれを抽出する周辺分布法や[4-7]，文字・図形の大きさの違いに着目してそれぞれを抽出する黒連結法[4-10,4-11]が提案されている．周辺分布特徴はテキスト領域の抽出に有効であり，また，黒連結特徴は図表領域の抽出に有効である．一方，本章の処理対象のようにこれらの領域が混在する場合，周辺分布特徴のみでは図表領域の抽出が困難となり，黒連結特徴のみではテキスト領域の抽出が困難と考えられる．そこで，以下に述べるように，第一段階で周辺分布特徴を用いて粗い領域分割を行い，第二段階で黒連結特徴を用いて詳細に領域を抽出するといった 2 つの特徴を組み合わせて用いる手法

を提案する．本提案手法により，テキスト，表，図領域が混在した文書に対しても，各領域が安定して抽出可能と考えられる．

一般に，文書に含まれる各領域は，空白領域やフィールドセパレータと呼ばれる線分で区切られている．空白領域は，特に文書を粗く分割するのに有効であり，また，周辺分布特徴から容易に抽出することができる．そこで第一段階として，周辺分布特徴を用いた文書画像の分割を行う．空白領域は周辺分布上で値が低くなるため，その位置を求めることができる，図 4-6(a) は図 4-3 に示した入力文書画像の横方向の周辺分布である．この周辺分布が 0 となる領域を空白領域として，文書を横短冊状に分割した結果を同図 (b) に示す．図に示した例のように横書きの文書では，横方向の周辺分布を用いることにより，文字列や図表領域を粗く抽出できることがわかる．

第二段階では，第一段階で抽出した領域ごとに，さらにその内部を分折し，テキスト領域，表領域，図領域を分離抽出する．

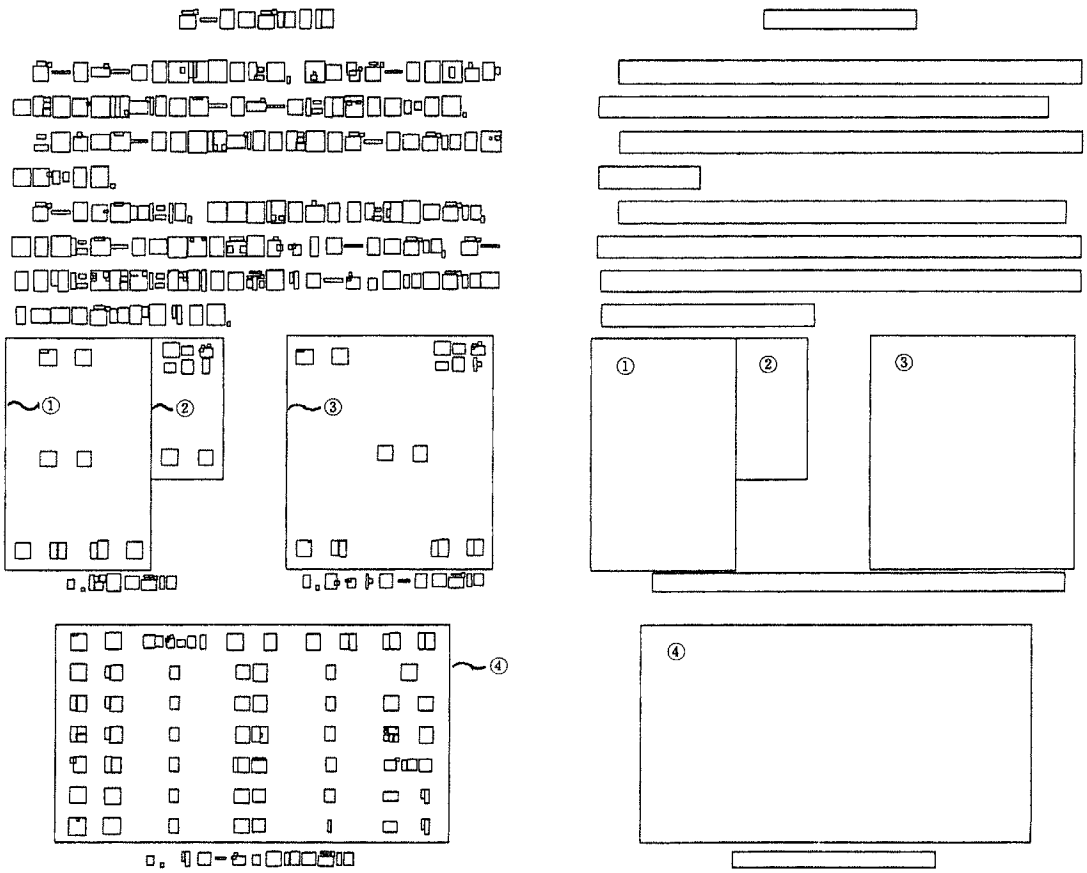
テキスト領域は文字のみからなる領域で，表や図領域は文字と図形が混在した領域である．表や図領域中に含まれる表罫線や種々の線図形の中には，文字に比べ大きいものが必ず存在するため，これらを抽出すればテキスト領域と表や図領域を区別することができる．文字・図形の大きさは黒連結特徴を抽出して求める．すなわち，各黒画素の 8 連結領域に外接する矩形枠の周囲長を求め，あるしきい値より大きい場合は図形，小さい場合は文字とする．しきい値として，周囲長のヒストグラムを求め，文字・図形分離に適当な値を設定する．図 4-7(a) は，図 4-3 の入力文書画像から抽出した黒連結の外接枠を示したもので，図中の①〜④が図形と判断された．同図 (b) は第一段階で抽出した領域について図形の有無を調べ，テキスト領域と，表または図の領域を抽出した結果を示している．ここでテキスト領域は，次節で述べるテキスト領域の認識処理を行う．表または図の領域については，図 4-2 に示した処理の流れに従い，まず表領域の認識処理を行う．処理の過程で表領域の特徴が抽出されない場合は，図領域の認識処理を行う．



(a) 文字列方向の周辺分布　　(b) 領域抽出結果

図 4-6　周辺分布を用いた領域抽出結果

Fig. 4-6　Experimental result of segmentation using projection profile.



(a) 黒連結の外接矩形枠　　(b) 領域抽出結果

図 4-7　黒連結特徴を用いた領域抽出結果

Fig. 4-7　Experimental result of segmentation using connected components.

4.4.3　テキスト領域の認識

テキスト領域中の個々の文字を切り出して文字認識する．まず，縦書き/横書き情報を基に文字列を抽出し，次に，文字列ごとに一文字切り出しを行う．行ピッチや文字ピッチは，文字列抽出や一文字切り出しに有効な情報である．しかしながら，これらの情報は文書によって，また同一文書でも場所によって異なるため，予め特定することができない，そこで，処理するテキスト領域中の文字の大きさを推定し，この大きさ情報を用いて文字列抽出および一文字切り出しを行う．文字列抽出処理は，文字列と文字列の間隔に着目して行う．具体的には，横書きの場合，黒連結を紙面の上から順に，外接矩形枠の左上座標を用いて並べ，推定した文字高さより間隔の広いところを抽出して文字列の区切りとする．縦書きの場合，同様に紙面の右から順に，文字幅を推定して文字列

を抽出する．

文字列が抽出されると，文字列ごとに一文字切り出しを行う．横書きの場合は，黒連結を外接枠の左上座標を用いて左から右へ，縦書きの場合は上から下へ並べ，隣接する外接枠が次のような条件を満たしたとき，それらは同一文字を構成していると見なし，それぞれを包含する新しい矩形枠を抽出する．

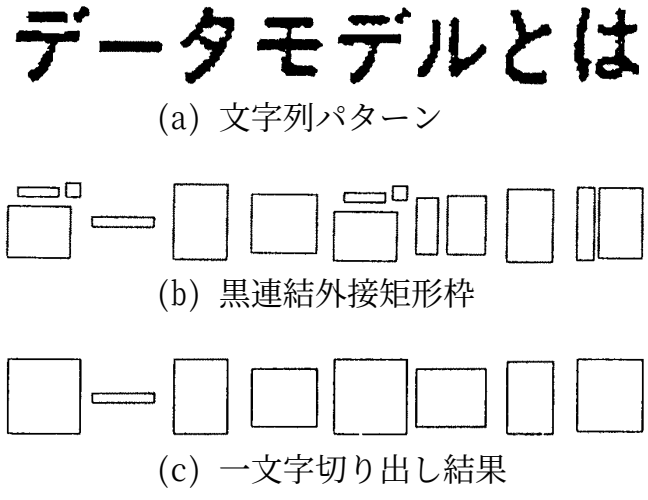
(1) 包含関係にある

(2) 共通領域を持つ

(3) 中心座標間の距離が推定した文字の高さ，幅より近い．

ここで (3) は，文字が偏や旁など複数の離れた黒連結から構成されるとき，一文字を切り出すのに必要な条件である．(3) の条件を用いた一文字切り出し結果の例を図 4-8 に示す．

一文字切り出しの結果に基づき文字認識を行う．OCRに代表される文字認識技術を活用して文字認識を行う場合，処理に適した文字パタンサイズが設定されている．テキスト領域中の文字の大きさは様々であるため，文字パタンサイズの正規化を行う．拡大/縮小は文字パタンの幾何学的特徴を保存するため，領域判別法[4-13]を用いた．また，印刷文字，手書き文字に対処するため，文字認識には，ストローク構造集積法[4-14]を用いた．文字認識の結果が一意に定まらない文字については，候補文字が抽出されるため，文書認識処理全体が終了した後，オペレータ介在により候補文字の選択を行う．



(a) 文字列パターン

(b) 黒連結外接矩形枠

(c) 一文字切り出し結果

図 4-8　一文字切り出し結果の例

Fig. 4-8　Examples of character extraction.

4.4.4 表領域の認識

領域抽出処理で抽出された表または図領域について，まず表領域かどうかの判別を行う．表は水平，垂直線分で囲まれ，表の内部もまた罫線で区切られた矩形領域であることから，まず領域内の線分を抽出する[4-15]．次に，抽出された線分の始点，終点座標に基づいて線分の接続関係を調べ，線分の中で表罫線を成しているものを抽出する．抽出されない場合は図領域と判断し，次節で述べる図領域の認識を行う．

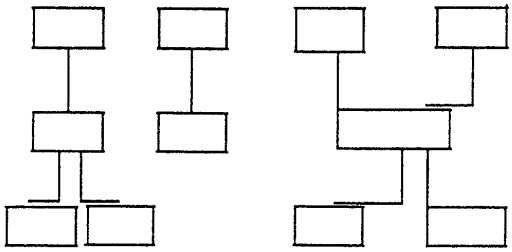
表を認識することにより，帳票等に記入されたデータの自動入力が可能となる．この時，表のどのフィールドにどんなデータが書かれているかを抽出することが重要である．そこで，表の構造を解析して各フィールドの位置と大きさを抽出する．図 4-9 は，図 4-7(b) で抽出された①～④の領域の線分抽出結果を示したもので，このうち，④の領域が表領域と判断され，各フィールドの左上点と右下点の座標を抽出した結果を示している．

表領域内の文字の認識は，各フィールドごとに行う．各フィールドの位置と大きさを基に，その内部に含まれる黒連結を抽出し，4.4.3 で述べたテキスト領域の認識処理と同様に，文字列抽出および一文字切り出しを行って文字認識する．

4.4.5 図領域の認識

表の特徴が抽出できなかった領域は，図領域として処理する．図領域中にはしばしば図のタイトルなど，図形に混ざって文字が書かれている．図領域の認識処理では，まず文字と図形を分離抽出し，文字については文字認識し，図形については線分を抽出する．一方，オフィス文書中の図領域には線分以外の様々な図形も混在することから，これらの図形はイメージデータのまま扱うこととした．

文字・図形を分離抽出するため，文字と図形の複雑さの違いに関する特徴を導入する．これは，図形のなかには点線のように小さなものも存在し，黒連結等の大きさ情報のみでは文字と図形を十分区別できないため，文字は図形より複雑であるという性質に着目するためである．文字・図形の複雑さは，第 3 章で述べた近接線密度特徴により表すことができる．図 4-10 は図 4-7(b) の③の領域について近接線密度を抽出した結果で，近接線密度の高いところが複雑な領域であることを示している．この近接線密度があるしきい値より高く，極大となる点を抽出し，その周囲の矩形領域を抽出した結果を同図 (c) に示す．この矩形領域は図領域中の文字部分の領域に対応しており，この領域をテキスト領域とみて, 4.4.3 で述べた文字列抽出, 一文字切り出しおよび文字認識を行う．



1	2	・		
・				
				・
			・	35

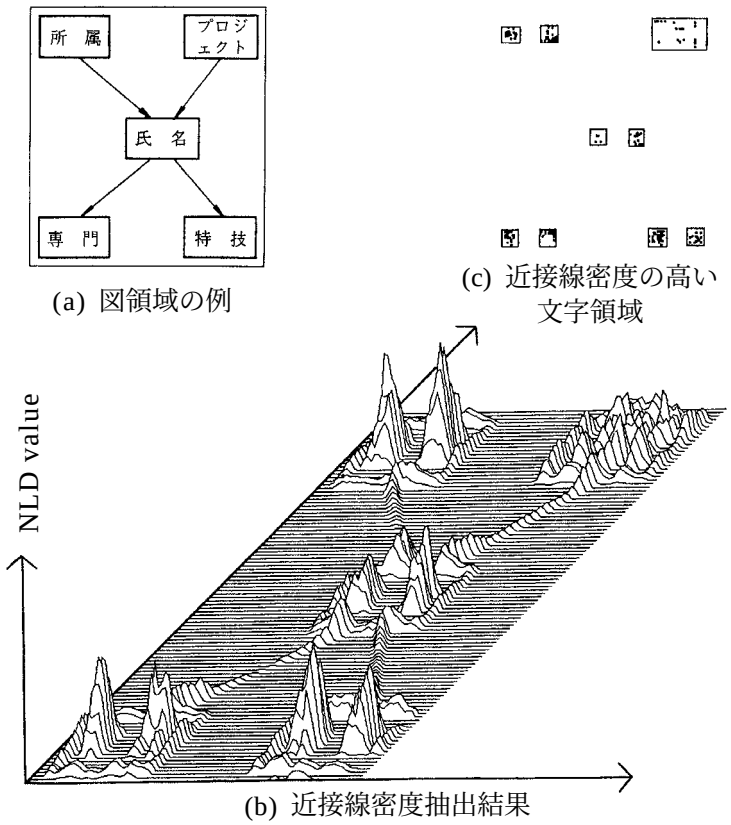
(a) 線分抽出結果

フィールド番号	左上点 x座標	左上点 y座標	右下点 x座標	右下点 y座標
1	109	1095	235	1140
2	243	1095	370	1140
3	378	1095	505	1140
4	513	1095	640	1140
5	648	1095	776	1140
6	109	1148	235	1193
7	243	1148	370	1193
8	378	1148	505	1193
9	513	1148	640	1193
10	648	1148	776	1193
11	109	1201	235	1246
12	243	1201	370	1246
13	378	1201	505	1246
14	513	1201	640	1246
15	648	1201	776	1246
16	109	1254	235	1300
17	243	1254	370	1300
18	378	1254	505	1300
19	513	1254	640	1300
20	648	1254	776	1300
21	109	1308	235	1353
22	243	1308	370	1353
23	378	1308	505	1353
24	513	1308	640	1353
25	648	1308	776	1353
26	109	1361	235	1406
27	243	1361	370	1406
28	378	1361	505	1406
29	513	1361	640	1406
30	648	1361	776	1406
31	109	1414	235	1459
32	243	1414	370	1459
33	378	1414	505	1459
34	513	1414	640	1459
35	648	1414	776	1459

(b) フィールドの位置情報

図 4-9　表罫線とフィールド抽出結果の例

Fig. 4-9　Experimental results of line extraction and resulting table structure.



(a) 図領域の例

(c) 近接線密度の高い文字領域

(b) 近接線密度抽出結果

図 4-10　近接線密度の例

Fig. 4-10　Example of NLD profile.

4.4.6　後処理

イメージスキャナから入力した文書の認識結果は，蓄積や検索のみならず，編集にも用いられる．認識した文書がキーボード等から入力した文書と同様に処理できるよう，認識結果に基づいて文書ファイルを作成する．文書ファイルは，各領域の位置や大きさなどの書式情報，認識結果の文字コード情報，図形コマンド情報，およびイメージ情報から構成される．ここで，イメージ情報は，蓄積する際のファイル量削減のため画像符号化する．

4.5　実験と考察

4.5.1　実験システムの構成

4.4 で述べたアルゴリズムを評価するため，認識実験を行った．実験に用いた文書認識システムの構成を図 4-11 に示す．文書をイメージスキャナより解像度 400 dots/inch の 2 値画像として入力し，一旦イメージメモリに蓄積する．雑音

除去や各特徴抽出，線分抽出は，処理を高速化するため，専用ハードウェアにより行う．また，文字認識は漢字 OCR で行う．CPU は処理全体の流れを制御するとともに，領域抽出や一文字切り出し処理を行う．また，後処理を行って認識結果をビットマップディスプレイに表示する．

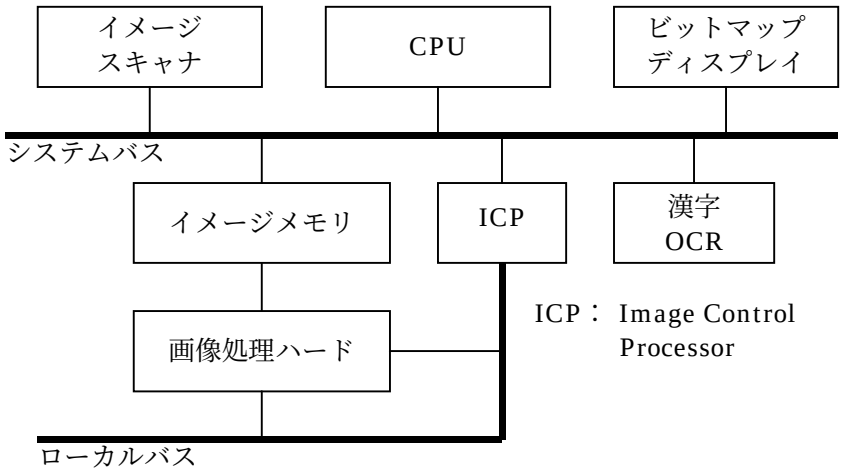


図 4-11　文書認識実験システムの構成
Fig. 4-11　Experimental document recognition system.

4.5.2　実験結果の考察

実験は，4.4 の各節で述べたアルゴリズムの機能を個別に評価するため，以下に述べる (1) 〜 (5) の 5 項目について行った．各実験の入力には，4.2 で述べた文書の中から，印刷文書で，アルゴリズムが評価し易い内容の文書をそれぞれ選択して用いた．さらに，(6) では文書全体の処理結果について考察した．以下，実験結果とその考察について述べる．

(1) 傾斜検出，縦/横書き判別実験

4.4.1 で述べた前処理の機能を確認するため，入力した文書の傾斜検出，および縦/横書き判別を行った．入力には，縦書きおよび横書きの文書を用い，それぞれについて 0°，±1°，±2° 傾斜させて入力した．

実験の結果，それぞれの入力文書の傾斜量に応じた Δi が抽出され，

傾斜角 $\Theta=\tan^{-1}(\Delta i$／周辺分布を抽出する際分割した領域の幅)

で求められることが分かった．また，縦/横書きも正しく判別された．傾斜した文書に対して文字列抽出を行ったところ，±2° 傾斜した文書では，文字列の間

隔が狭い場合，文字列抽出誤りが生じたため，±2°以上傾斜した文書をリジェクトすることとした．

(2) テキスト領域，表領域，図領域の抽出実験

4.4.2 で述べた領域抽出機能を確認するため，入力した文書に対して領域抽出を行った．入力には，表や図が混在した文書 (日本文/英文)，および縁取りのある文書を用いた．

実験の結果，周辺分布特徴と黒連結特徴を併用することにより，表領域の中に図領域が混在するなど，領域の重なりがある場合でも，各領域が正しく抽出できることが分かった．特に，縁取りのある文書は，一旦表領域が抽出され，その内部からさらに各領域が正しく抽出された．

(3) 文字切り出しおよび文字認識実験

4.4.3 で述べたテキスト領域の文字切り出しおよび文字認識機能を確認する実験を行った．入力にはテキスト領域のみの縦書き文書，横書き文書を用いた．

実験の結果，縦書き/横書きに依らず，文字列が正しく抽出され，さらに，文字並びの順に一文字が切り出されて文字認識された．しかし，漢字と英数字が混在するなど，文字ピッチが急に変化する場合があり，一文字切り出し誤りが生じて，文字認識結果がリジェクトとなった．これは，文字の大きさに着目しただけでは，文字ピッチの変動に十分対処できないためで，文字認識結果をフィードバックして再切り出しを行うなどの手法を導入する必要がある．

(4) 表領域のフィールド抽出実験

4.4.4 で述べた表領域の認識機能を確認するため，表領域のフィールド抽出実験を行った．入力には，一般的な事務処理等で用いられている帳票を用いた．

実験の結果，4 本の罫線で囲まれる領域という条件で，表の大きさやフィールド数の違いに依らず，正しくフィールドの位置が抽出された．しかし，表の中には，外枠罫線が省略されたものがあり，これらの表のフィールドが正しく抽出されないため，省略された罫線を補って処理するなど，アルゴリズムの改善が必要である．

(5) 図領域の文字・図形分離抽出実験

4.4.5 で述べた図領域の認識機能を確認するため，図領域の文字・図形の分離抽出実験を行った．入力には，図領域のみの文書で，ブロック図やフローチャートを用いた．

実験の結果，図の中のタイトル，ブロック図やフローチャートの中の文字列が正しく抽出された．しかし，文字に接近した図形で，かすれなどにより切断

しているところでは，図形部分も複雑となるため，近接線密度特徴に着目しただけでは抽出誤りが生じることが分かった．この問題を解決するため，文字認識結果と図形認識結果をフィードバックして文字・図形を分離するなどの手法の導入が必要である．

(6) 全体の評価

文書全体の処理結果について考察する．文書認識アルゴリズムを全体的に評価するため，次のような評価項目が考えられる．

① 領域抽出率

テキスト，表，図，イメージの各領域が正しく抽出されたか否かを示す．具体的には，抽出された各矩形領域がそれぞれ所望の領域に対応しているか否かで判断し，次のような場合は抽出誤りとする．

- 表罫線の抽出誤りにより，表領域が図領域となる
- 文字・図形分離抽出誤りにより，図領域がイメージ領域となる

② 文字認識率

文書中の文字の認識率を示す．具体的には，テキスト領域，表領域については，一文字切り出し率と正しく切り出した文字に対する文字認識処理での文字認識率の積で表すことができる．また，図領域については，文字の分離抽出率と前記の文字認識率の積で表すことができる．

以上について評価を行ったところ，表 4-1 に示した文書に対して，外枠罫線のない表が抽出誤りとなった点を除けばほぼ100 %の抽出率を達成した. また，文字認識率については，テキスト領域，表領域で 90 %以上，図領域では 85 %以上を達成することができた.

領域抽出率に関しては，(4) で述べたような外枠罫線のない表や，また，線分の切れなどに起因する罫線の抽出誤りがある表を正しく抽出するため，線分コマンドのみならず文字認識結果に基づいた領域抽出処理も必要である，また，文字認識率に関しては，文字切り出し誤りに起因する文字認識結果のリジェクトをフィードバックし，再切り出しを行うとか，単語処理等の言語レベルの処理を行って文字認識率の向上を図る必要がある．

以上の実験結果より，4.4 で述べたアルゴリズムによってオフィス文書の認識入力が可能となることが分かった．また，認識結果は文書の編集に有効に活用でき，文書を蓄積する際のファイル量を 2 桁程度削減できることも分かった．

今後の課題としと，各領域の認識処理において実験の考察で述べた問題点を解決するため，アルゴリズムの改善を図るほか，文字認識結果や図形認識結果を活用した処理も必要である．また，種々の図形の認識や，オフィス文書特有

の印影，サイン等，イメージ情報として扱っている領域の認識など，認識対象の拡大を図ることも必要と考えられる．

4.6 まとめ

本章では，オフィス内に存在する文書の書式や文字情報，図形情報を認識する処理アルゴリズムについて述べた．また，実験システムを用い，認識実験を行った結果，文書中のテキスト領域，表領域，図領域の抽出が可能で，各領域に含まれる文字・図形についても文字認識，図形認識が可能であることを示した．

文書認識処理アルゴリズムは，処理対象とする文書の性質に密接に関係する．処理対象の拡大や認識性能の向上は必須であり，処理アルゴリズムの改善も必要と思われる．その際，パターンレベルの処理アルゴリズムのみならず，プロダクション・システム[4-16), 4-17)]に代表されるような，認識結果を活用した言語レベルの処理アルゴリズムの開発も必要である．また，文書認識システムの構築に当たっては，処理速度性能の向上も重要である．

第 5 章　レイアウト構造認識に基づく文書認識システムの構成法

5.1　はじめに

オフィスにおける文書の作成，編集，蓄積，検索，通信等の業務の合理化に対する要請に対し，大量のオフィス文書の電子ファイル化が急務となっている[5-1)]．オフィス文書を電子ファイル化する手段としては，現在，文書をイメージスキャナから入力してイメージ情報としてファイル化する方法と，ワードプロセッサ等からのキー入力による方法とがある．前者は入力が簡便である反面，ファイル容量が膨大になる，文書編集に対する柔軟性に欠ける，等の問題があり，後者は文書編集に対する柔軟性はあるものの，入力に多大な労力と時間を要するという問題がある．これらの問題に対し，イメージのままで切り貼り等の編集を行うイメージ編集システム[5-2), 5-3)]も提案されているが，編集機能に自ずと限界が生ずる点，ファイル容量の問題が解決されていない点を考慮すると十分とは言い難い．

そこで，本章では，イメージ入力された文書に対し認識技術，画像処理技術を適用して，これをワードプロセッサ等で処理可能な文書ファイルに変換する文書認識システムの構成法について論述する．具体的には，5.2 で文書認識の概要および文書認識の実現技術について述べる．さらに，5.3 では文書認識の処理モデルを提示し，モデルに基づくサービス性の評価を行い，システム構成に向けての幾つかの指針を導出する．最後に，5.4 でシステム実現例として LAN を用いたシステム構成を例示し，5.5 で本章のまとめと今後の課題を述べる．

5.2　文書認識の実現技術

オフィスにおける文書には，本文・見出し等のテキスト情報，図表等に含まれる幾何図形情報，印影・写真等のイメージ情報が混在し，しかもその書式は一般には不定形である．文書認識では，イメージ入力されたこれらの文書から文字情報，図形情報，イメージ情報を分離抽出し，それぞれ文字コード，図形コマンド，圧縮イメージに変換する処理を行う[5-4)]．この文書認識の概念を図 5-1 に示す．文書認識により以下の効果が期待できる．

a) 文字コード，図形コマンドレベルでの文書の修正，編集が可能となる．(文書編集への柔軟性)
b) イメージのまま蓄積するより少ないファイル量で文書の電子ファイル化が可能となる．(情報圧縮効果)

c) キー入力と比較して遥かに少ない労力と時間で文書入力が可能となる．(文書入力の高速化・省力化)

文書認識の実現には，以下の処理技術が必要である．

① 文書の書式構造を解析し，テキスト領域，図表領域，イメージ領域を分離抽出する技術
② テキスト領域に対し，個々の文字を一文字ずつ切り出し文字認識により文字コードに変換する技術
③ 図表領域に対し，文字と図形要素を分離抽出し，文字については文字認識，図形要素については図形認識し，それぞれ文字コード，図形コマンドに変換する技術
④ イメージ領域に対し画像圧縮符号化を行う技術

本章では，既存アルゴリズムの適用性という観点から上記処理技術について概説する．

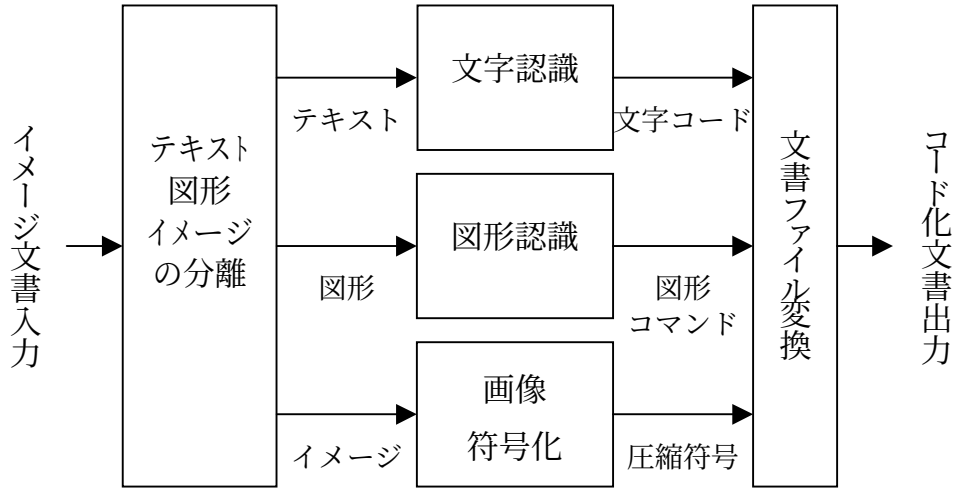


図 5-1　文書認識の概念

Fig. 5-1　Concept of document recognition.

5.2.1　領域の分離抽出

文書認識では，文書中の領域を表 5-1 のように捉え，各領域のもつ性質の違いに着目して領域の分離抽出を行う．領域の分離抽出技術として幾つかの手法が提案されているが，ここでは周辺分布法[5-5]と黒連結法[5-6]を併用したアルゴリズムについて述べる．

周辺分布法は，文書画像の水平方向，垂直方向の投影をとる手法である．テキスト領域の場合，文字列方向の投影値が文字列部分で高い値を示し，行間部分で低い値となり，これらが周期的に出現する性質がある．この周辺分布の周期性に着目してテキスト領域とそれ以外の領域の切り分けが可能となる．一方，黒連結法では，文書画像中の互いに連結する黒画素の集合 (黒連結) を抽出し，図形を構成する黒連結が文字を構成する黒連結よりも大きいことに着目して領域の切り分けを行う．

周辺分布法は，行ピッチ，文字ピッチ等の文書の構造をマクロ的に捉えるのに有効な手法であるが，文書画像のミクロな特徴は捉えにくい．一方，黒連結法では文書画像の細かな特徴を反映できるものの，個々の黒連結特徴を組織化して領域判定までもってゆく手順が複雑となる．そこで，周辺分布法で文書の大まかな領域分割を行った後，各領域について黒連結を調べるアルゴリズムが両手法の利点を生かす上で有効と考える．

表 5-1　文書に含まれる領域

Table 5-1　Domains included in document.

領　域	定　　義
テキスト領域	文字情報のみから成る領域.各文字は,ほぼ一定の文字ピッチ,行ピッチをもって規則的に配置する.
図　表　領　域	線図形と文字より構成される領域.
イメージ領域	テキスト領域,図表領域以外で,写真・印影等の情報が存在する領域.

5.2.2　テキスト領域の処理

(1) 一文字切り出し

テキスト領域中から一文字切り出しを行うアルゴリズムとして，前節で述べた周辺分布特徴と黒連結特徴を活用する方法がある．すなわち，周辺分布より得られる文字ピッチ，行ピッチ情報を手掛かりに近接する黒連結を統合し，これら黒連結を含む外接矩形として文字を切り出す．

(2) 文字認識

文字認識については，OCR技術が適用可能である．例えば，手書き漢字を認識するストローク構造集積法[5-7]，マルチフォント印刷漢字認識アルゴリズム[5-8]

等が文書認識のための文字認識技術として考えられる.

5.2.3 図表領域の処理

(1) 文字・図形分離抽出

図表中の文字を分離抽出するアルゴリズムとして，文字と図形の複雑さの違いに注目した近接線密度法[5-9]がある．近接線密度法では，文書画像中の各黒画素について複雑さの指標となる近接線密度値を求め，文字部分の近接線密度が図形部分の近接線密度より高くなることを利用して，文字と図形の分離を行っている．また，5.2.1 で述べた黒連結法による文字・図形分離も可能である．すなわち，文字と図形の大きさの違いに着目して文字・図形分離を行う方法である[5-10].

黒連結法では，図形に接触した文字の抽出ができず，近接線密度法の場合，単純なストローク構造をもつ文字 (例えば数字など) が孤立して存在するときこれを文字として抽出できない等の問題がある．したがって，両アルゴリズムを併用して両者の弱点をカバーする方法が，精度の高い文字・図形分離を行う上で望ましい.

(2) 図形認識

文字認識では認識すべきカテゴリセットが明確に定義されるが，図形認識の場合，認識対象に依存してカテゴリセットが大きく変動する．そのため，これまで開発されている技術も論理回路図の認識[5-11]，フローチャートの認識[5-12]等，対象を特定した技術となっている．現時点で共通性，実用性の高い技術は線分抽出技術であり，線順次法[5-13]等のアルゴリズムが適用可能である．したがって，図表領域の処理については，① 処理対象を限定して対象に依存したアルゴリズムを用いる，② 線分のみをコマンド化し，その他はマンマシンインタフェースで補う，③ 線分以外の図形要素を含む図表領域はイメージ領域と同様に扱う，等の形態を採らざるを得ない.

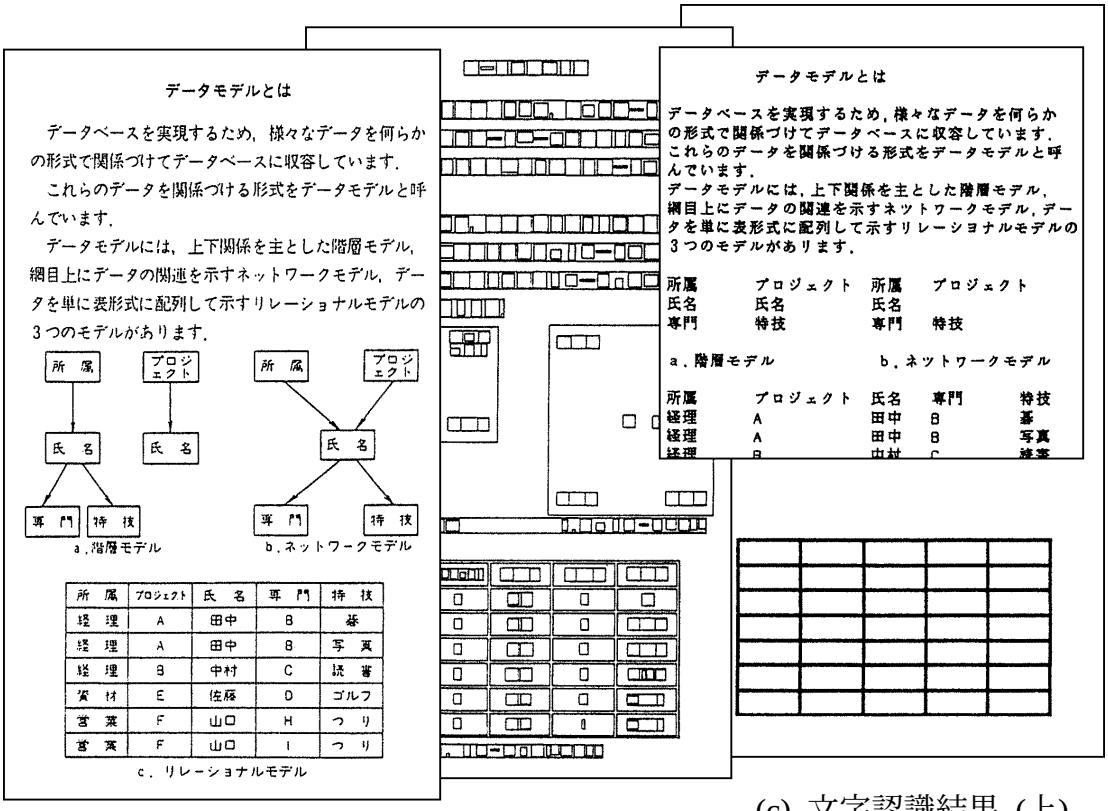
5.2.4 イメージ領域の処理

イメージ領域については，通信または蓄積におけるデータ量削減のため画像圧縮符号化を行う．符号化形式には，通信の互換性を保つ意味から，国際標準のMMR符号化方式[5-14]が望ましいと考える.

5.2.5 処理実験例

5.2.1～5.2.4 に示した処理技術に関し，その実現性をシミュレーション実験により確認した．領域の分離抽出，一文字切り出しには周辺分布法と黒連結法を

併用したアルゴリズムを，文字・図形分離抽出には近接線密度法と黒連結法を併用したアルゴリズムを，文字認識にはマルチフォント印刷漢字認識アルゴリズムを，図形認識には線順次法による線分抽出アルゴリズムを，それぞれ汎用計算機上のプログラムで実現し，シミュレーション実験を行った．実験結果を図 5-2 に示す．図 5-2 では，入力文書例およびそれに対する文字，図形，イメージの分離結果，文字認識結果，線分抽出結果を示している．



(a) 入力文書　(b) 文字・図形分離結果　(c) 文字認識結果 (上)　(d) 線分抽出結果

図 5-2　処理実験例

Fig. 5-2　Example of recognition experiment.

5.3　文書認識の処理モデルと評価

前節まで，文書認識における所要機能と各機能の実現技術について述べた．本章では，これらの機能を組み合わせて文書認識を実現する具体的処理モデルを提示し，このモデルに基づき，文書認識による情報圧縮効果，認識処理速度，文書入力の省力化効果を考察する．

5.3.1 文書認識モデル

(1) 処理モデル

文書認識の処理モデルを図 5-3 に示す．モデルの構成要素は，認識すべきイメージ文書がキューイングされる入力文書キュー，認識結果のコード化文書を出力する出力文書キュー，文書認識を実行する文書認識モジュール，文書認識処理の流れを制御する処理制御部，および文書認識モジュールのワークデータエリアとなるイメージメモリである．文書認識モジュールは，さらに特徴抽出，領域分割，一文字切り出し，文字認識，文字・図形分離，図形認識，イメージ圧縮の 7 つの内部モジュールより構成し，それぞれが独立に動作可能とする．本処理モデルでは，複数ページを文書認識モジュール内で扱うことを前提に，各内部モジュールに処理要求キューを設けた．また，イメージデータ，特徴抽出結果等の実データはイメージメモリ上に配置するものとする．各内部モジュールの機能概要を以下に示す．

① 特徴抽出モジュール
イメージ文書より周辺分布，黒連結の各特徴を抽出する．処理結果は領域分割モジュールにキューイングする．

② 領域分割モジュール
周辺分布，黒連結の各特徴を用いてテキスト領域，図表領域，イメージ領域を分離抽出する．テキスト，図表，イメージの各領域は，それぞれ，一文字切り出し，文字・図形分離，イメージ圧縮の各モジュールにキューイングされる．

③ 一文字切り出しモジュール
テキスト領域中の文字を周辺分布，黒連結の各特徴により切り出す．切り出し結果は文字認識モジュールにキューイングされる．

④ 文字認識モジュール
文字パタンを文字コードに変換する．

⑤ 文字・図形分離モジュール
図形領域に対し近接線密度特徴を抽出し，これと黒連結特徴により文字と図形を分離する．分離された文字と図形は，それぞれ，文字認識，図形認識モジュールにキューイングされる．

⑥ 図形認識モジュール
図形パタンを図形コマンドに変換する．

⑦ イメージ圧縮モジュール

イメージ領域に対し MMR 符号化を行う.

本処理モデルでは，ページ内およびページ間の並列処理を考慮し，次のアルゴリズムで処理フローを制御する.

a) 各内部モジュールは, 処理制御部からの起動により作動し, 処理要求キュー先頭の処理を行う. 処理終了後, 処理結果を次モジュールにキューイングし, 処理制御部に割込通知して停止する．なお，一文字切り出し，文字認識，文字・図形分離の各モジュールは, 要求された処理を一定の処理単位に分割して行い，残存した処理分は自処理要求キューの末尾に登録される.
b) 処理制御部はこの割込通知により動作し，次の i) 〜 iii) の処理を行った後停止する.
   i) 認識を完了した文書があれば，これを出力文書キューにはき出す.
   ii) 文書認識モジュールが同時処理するページ数 (マルチ度) が一定となるよう，入力文書キューからの取り込みを行う.
   iii) アイドルとなっている内部モジュールをサーチし, 処理要求がキューイングされていればその内部モジュールを起動する.

上記処理フロー制御により，ページ内においてはテキスト領域の処理，図表領域の処理，イメージ領域の処理が並列的に実行可能であり，さらにテキスト領域内では一文字切り出しと文字認識とが，図表領域内では文字・図形分離と図形認識・文字認識とがそれぞれパイプライン的に並列処理可能である. 一方, ページ内でシーケンシャルにならざるを得ない処理については，ページ間での並列処理が可能である．例えば，領域分割までの処理とそれ以降の処理とはページ内で並列処理できないが, 1 ページ目の領域分割以降の処理実行中に 2 ページ目の領域分割を行う等の並列処理は可能である.

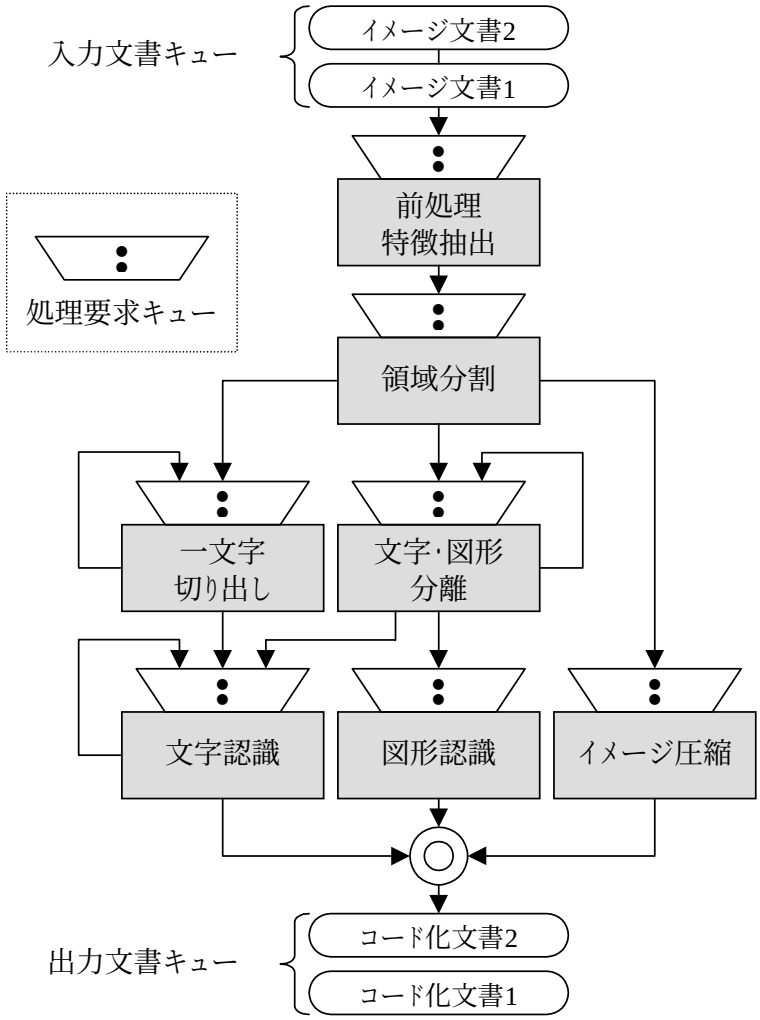


図 5-3　文書認識の処理モデル

Fig. 5-3　Model for document recognition process.

(2) 文書モデル

文書のモデルについては，これまでミクストモード通信を対象とした文書構造モデル[5-15]が報告されているが，文書の統計的性質が明確にされていないため，文書認識の評価には適用できない．ここでは，文書認識評価のために，① 文書中に占めるテキスト，図表，イメージの割合，② テキスト領域中の文字数，③ 図表領域中の文字数，図形要素数をモデル化項目とし，任意に抽出した電気通信関係の論文，記事 100 ページに対する測定結果を基に図 5-4 の文書モデルを設定した．

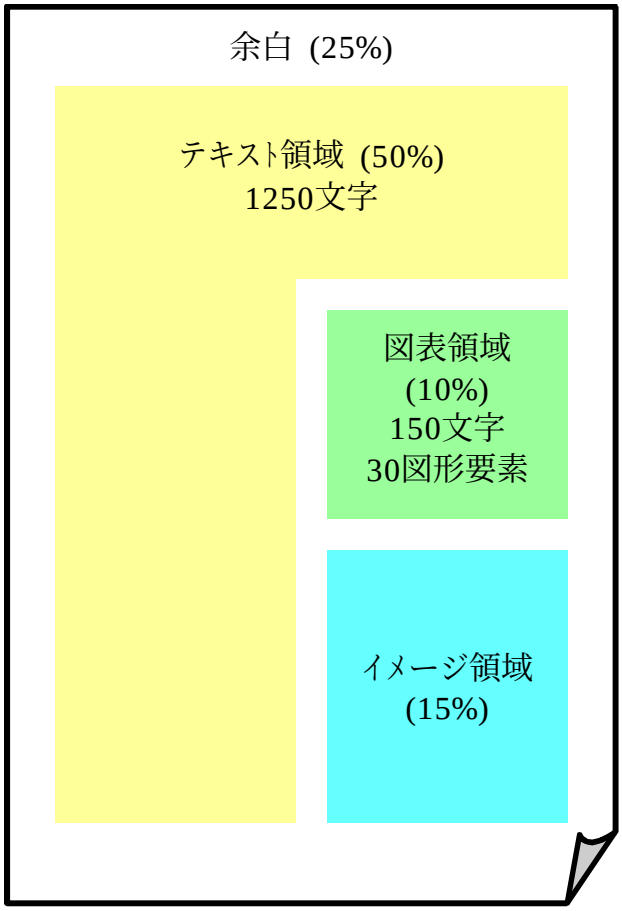


図 5-4　文書モデル

Fig. 5-4　Document model.

5.3.2　文書認識による情報圧縮効果

A4 文書を 400 dots/inch の白黒 2 値画像として表現すると約 2MByte の情報量となる．一方，これを文書認識した場合には，テキスト領域，図表領域がコード化されるため，大幅な情報圧縮効果が期待できる．5.3.1 の文書モデルに基づいて試算した情報圧縮効果を表 5-2 に示す. 表 5-2 の算出においては次の仮定を置いた.

a) 文字については, 文字コード, 文字位置, サイズ, フォント等の情報を含め, 8 バイト / 文字とする.

b) 図形 (線分) については，図形種別，始点・終点座標，線分属性のために 10 バイト / 図形を要する.

c) MMR 符号化による圧縮率は，解像度にほぼ比例することから，300，400 dots/inch の解像度に対し，それぞれ 1 / 15，1 / 20 とする.

表 5-2　文書認識による情報圧縮効果

Table 5-2　Data compression effect by document recognition.

解像度	テキスト	図表	イメージ	文書全体
300 dots/inch	1/ 46	1/ 61	1/15	1/44
400 dots/inch	1/103	1/137	1/20	1/76

5.3.3　文書認識速度

5.3.1 に示した処理モデルに従って文書認識した際の速度性能をシミュレーション実験により確認した.

5.3.3.1　実験の条件

本実験では，認識アルゴリズムのシミュレートは行わず，モデルにおける処理の流れを時間軸上でシミュレートすることにより，速度性能を求めることとした．速度性能を決定する要因として，入力文書の性質，内部モジュールの処理速度，マルチ度 (文書認識モジュールが同時処理するページ数) があり，これらについての実験条件を以下の (1) ～ (3) に示す．なお，実験の単純化のために，① 入力トラヒックは十分密であり，入力キューが空になることはない，② イメージメモリ容量は無限大で，イメージメモリアクセスに対する競合等のオーバヘッドは無視する，という仮定を置いた.

(1) 入力文書の性質

認識対象文書は，5.3.1 の文書モデルに従うものとし，次の 2 つの文書群を確率的に疑似発生させた.

[タイプ 1]：テキスト / 図表 / イメージ混在文書

ページ内にテキスト領域，図表領域，イメージ領域が混在する文書群．ページ毎に各領域の含有率は変えるが，平均値は前述の文書モデルを満足するよう，乱数を用いて疑似発生させる.

[タイプ 2]：テキスト / 図表 / イメージ独立文書

ページ内には領域混在のない文書群で，テキスト文書，図表文書，イメージ文書から成る．テキスト文書，図表文書，イメージ文書のページ数の比が前述の文書モデルを満足するように疑似発生させる．発生の順序は乱数により，ランダム化する.

なお，タイプ 1，タイプ 2 の文書サイズは共に A4 とし，入力解像度は 400 dots/inch を想定した.

(2) 内部モジュールの処理速度

直接画像を扱う処理をハードで，その他をソフトで行うことを想定し，アルゴリズムの複雑さおよび市販部品速度を前提に，表 5-3 の内部モジュール速度を推定した. 実験では, この数値を基に文書認識速度の期待値を求めると共に, 各内部モジュールの速度を変数としても捉え，これを変化させて内部モジュール間の負荷バランス等を調べた.

表 5-3　各内部モジュールの処理速度の推定値

Table 5-3　Estimation on processing speed of inner modules.

モジュール	処理速度
特　徴　抽　出	3　msec/raster
領　域　分　割	1　sec/A4
一文字切り出し	20　msec/character
文　字　認　識	20　msec/character
文字・図形分離	10　μsec/pixel
図　形　認　識	400　nsec/pixel
イメージ圧縮	200　nsec/pixel
処理制御部	10　μsec/interrupt

(3) マルチ度

文書認識における並列処理効果を調べるために，マルチ度をパラメータとした文書認識時間，モジュール稼動率を求めた.

5.3.3.2　実験の結果

(1) 文書認識速度の期待値

前述のタイプ 1 文書，タイプ 2 文書に対するページ当たりの平均文書認識時間を図 5-5 に示す. 但し, 内部モジュールの処理速度は表 5-3 の値を仮定している．マルチ度 1 では，ページ間の並列処理は行われないため，このときのタイプ 1, タイプ 2 文書の認識時間差がページ内並列処理効果の差となって現れている．マルチ度を上げることに伴う認識時間の減少は，ページ間並列処理効果によるものであり，マルチ度 4 以上で飽和傾向を示している，表 5-3 の数値を前提とすれば，並列処理により 30 秒/A4 程度の文書認識速度は十分期待できる．図 5-6 はこのときの内部モジュールの稼動率を示すものである．モジュール毎に稼動率が大きく異なるのは，各モジュールが処理負荷的にバランスしていな

いためであり，文字認識モジュールがボトルネックとなっている．

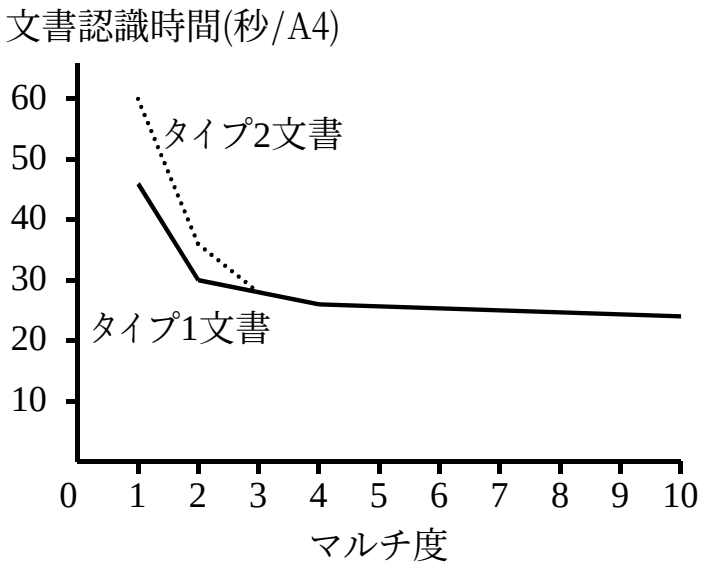


図 5-5　文書認識速度の期待値

Fig. 5-5　Expected speed of document recognition.

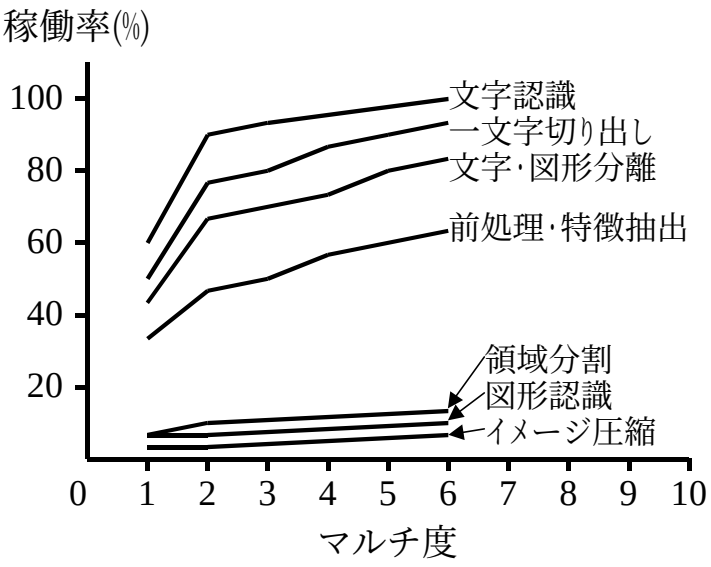


図 5-6　内部のモジュールの稼働率

Fig. 5-6　Working ratio of inner modules.

(2) 内部モジュールの負荷の平準化

各内部モジュールの処理速度をパラメータとして変化させることにより，処理負荷が平準化される内部モジュール速度を実験的に求めた．入力文書にはタイプ 1 文書を用いた．結果を表 5-4 に示す．表 5-4 における T は係数 (処理時間係数と呼ぶ) であり，T の値に依らずモジュール稼動率はほぼ一定で図 5-7 の傾向を示した．このようにモジュール負荷の平準化を図るためには，表 5-4 に示す速度比で内部モジュール速度を設定する必要があるが，この速度比は文書中に含まれる各領域の比率とほぼ一致する．また，図 5-7 に示すように，モジュール負荷の平準化を図った場合には，マルチ度を十分高くとらないと稼動率を 100 %に近づけることができない．図 5-8 には，処理時間係数 T と文書認識時間の関係をマルチ度をパラメータに示す．

表 5-4　モジュール負荷を平準化する処理速度

Table 5-4　Processing speed for equalization of module load.

モジュール	処理時間*	速度比**
特 徴 抽 出	2.9 T (msec/raster)	1.05
領 域 分 割	13.1 T (sec/A4)	1.00
一文字切り出し	11 T (msec/character)	0.48
文 字 認 識	10 T (msec/character)	0.52
文字・図形分離	8 T (μsec/pixel)	0.10
図 形 認 識	8 T (μsec/pixel)	0.10
イメージ圧縮	5 T (μsec/pixel)	0.16

* T：処理時間係数

** 単位時間当たりの処理画素数に換算した速度比で，“領域分割”を1とした

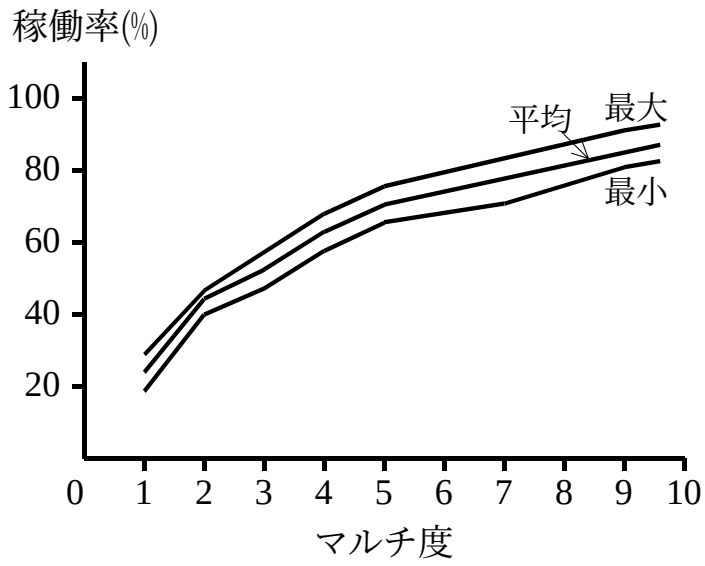


図 5-7　モジュール負荷の平準化による稼働率

Fig. 5-7　Working ratio with equalization of module load.

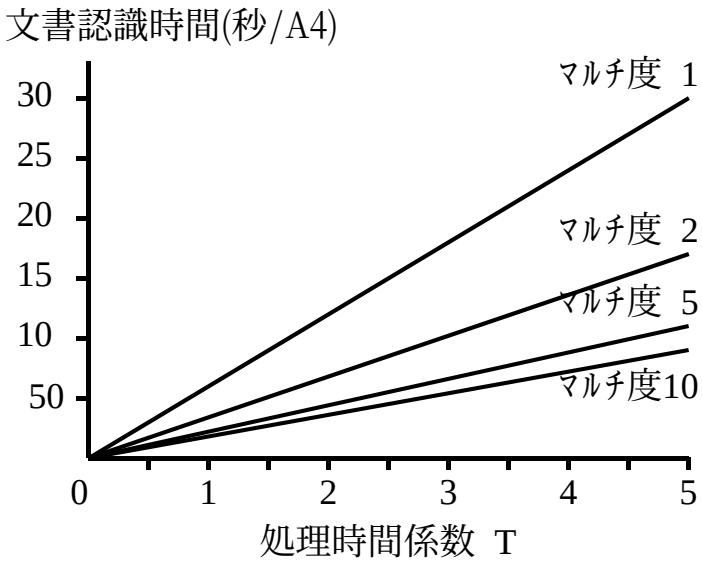


図 5-8　モジュール負荷の平準化による文書認識時間

Fig. 5-8　Document recognition time with equalization of module load.

(3) 装置設計上の指針

実験結果から得た装置設計上の指針を列挙する．

① 並列処理による認識速度向上を図るため，マルチ度を考慮した装置設計が望ましい．しかし，マルチ度を上げるとイメージメモリ量も増加するため，コスト性能比の観点に立った現実の装置設計では，速度改善率の高いマルチ度2〜3で設計するのが妥当と考える．

② 文書認識速度が設計条件として規定された場合，これを達成するために必要な各内部モジュール速度は表5-4および図5-8より推定可能である．

③ 文書中の各領域の比率に応じて，各内部モジュール速度を設定すれば，モジュール負荷の平準化を図ることができる．

④ 但し，マルチ度が制限される環境下では，モジュール負荷の平準化が最適とは限らない．図5-6と図5-7の比較によれば，限られたマルチ度で特定モジュール (例えば文字認識モジュール) の性能をフルに引き出すためには，負荷バランスを崩しても他モジュールの速度を上げることが有効となる．

⑤ 現状において30秒 / A4程度の文書認識速度は十分期待できるが，さらに認識速度を上げるためにはボトルネックとなっている文字認識の速度向上が必要となる．

5.3.4 文書認識による省力効果

本節では，テキスト入力を対象に，キー入力の代表的方式であるカナ漢字変換入力方式と文書認識入力方式とをオペレータ作業量の観点から比較検討し，文書認識による省力化効果を明らかにする．

カナ漢字変換入力におけるオペレータ作業には，カナキー入力を行う初期入力作業と，ディスプレイ画面を見ながら同音異義語の選択を行う校正作業とがあり，表5-5の作業特性を持つ[5-16]．一方，文書認識入力の場合のオペレータ作業は，主として認識できなかった文字 (リジェクト文字) に対する修正作業となる．この作業はリジェクト文字に対する候補文字をディスプレイ等に表示し，その中から正解文字を選択する作業となるため，カナ漢字変換での校正作業と類似する．そこで，[リジェクト修正時間 / 字 = 同音異義語選択時間 / 回]を仮定して，オペレータ作業量を試算した．図5-9は，文書認識入力におけるオペレータ作業量をカナ漢字変換入力の作業量を1として表したもので，パラメータとして認識率をとっている．リジェクト修正における入力速度がカナキー入力速度に比べて遅いことから，認識率が40 %以上でないと省力化効果は得られない．現状の認識技術では，比較的丁寧に手書きされた漢字あるいは印刷漢字に対して95 %以上の認識率を実現しており[5-7), 5-8)]，文字切り出しミス等による

認識率低下を考慮しても 90 %以上の認識率は期待できる．したがって，文書認識によりオペレータ作業量は 1 / 6 以下に軽減される．

表 5-5　カナ漢字変換入力の作業特性

Table 5-5　Characteristic of Kana to Kanji translation.

項　目		記　事
条件	入力者	50 時間訓練後の素人女性
	入力文書	新聞社説.漢字含有率 40%
特性	入力速度	30～40 字/分
	作業比率	初期入力作業：2/3 校正作業　　：1/3
	同音異義語選択速度	2.8 秒 / 回

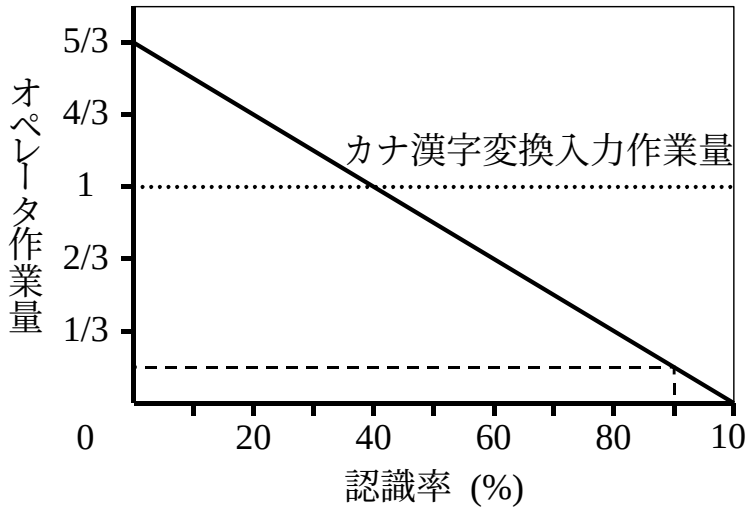


図 5-9　文書認識入力におけるオペレータ作業量

Fig. 5-9　Reduction effect in operator task by document recognition.

5.4　システムの構成例

前節の検討結果により，文書認識速度はオペレータの校正速度の 10～20 倍に達することが予想される．したがって，複数の文書処理端末から文書認識機能を共同利用するシステム形態が，文書認識性能を十分に活かす上で望ましい．このようなシステムの構成例として，図 5-10 に示すLANによる実現形態が考えられる．図中のメディア処理ステーションは，文字・図形・イメージ等の各種

メディアの処理により文書認識を実現するステーションで，複数のワークステーションからイメージ入力された文書を並列的に処理し，認識結果を各ワークステーションに返送する機能を持つ[5-4]．各ワークステーションは，この結果を文書の編集，蓄積，印刷等に活用する．本構成では，複数のワークステーションがメディア処理ステーションを共用することにより，文書認識性能をフルに活かすことができると共に，リソース共用による経済効果も期待できる．さらに，ワークステーションが持つ高度のマンマシンインタフェース機能をリジェクト修正のために活用することが可能となる．

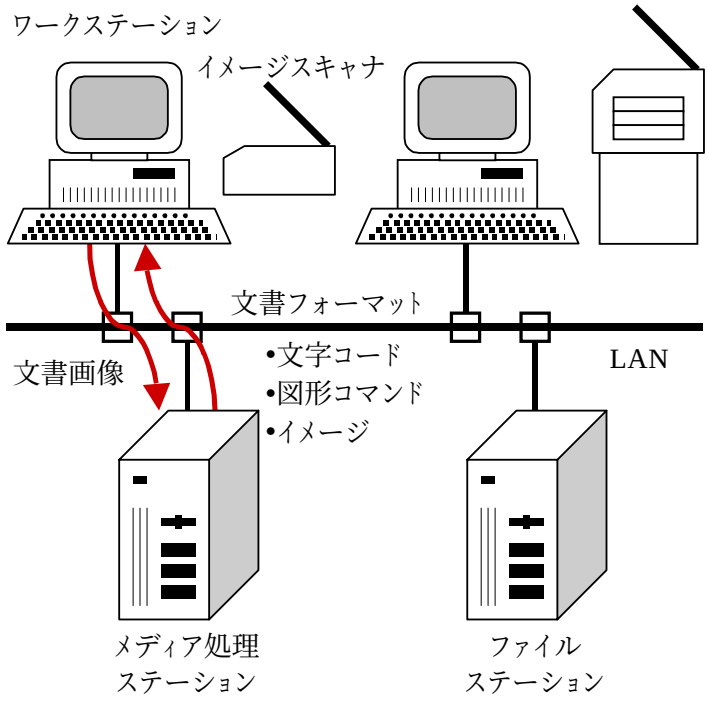


図 5-10　LAN における文書認識システムの実現例
Fig. 5-10　Document recognition system in LAN.

5.5　まとめ

本章では，文字・図形・イメージが混在するマルチメディア文書に対し，文字・図形・イメージの各情報を分離抽出し，それぞれを文字コード，図形コマンド，圧縮イメージに変換する文書認識システムについて論述した．文書認識がもたらす効果として，

a) イメージ入力された文書をワードプロセッサ等で編集可能な文書ファイルに自動的に変換できる

b) 文書をイメージのまま蓄積するのと比較して, 蓄積ファイル量の削減が図れる
c) 文書をキー入力するのと比較して, 初期入力時間の大幅な短縮が図れると共に，校正作業を含めたオペレータ作業の省力化が図れる

などが期待できる．本章では，文書認識の具体的な処理モデルを提示することにより，上記効果についての定量的な評価を試みた．シミュレーションによる評価実験により，文書認識の有効性を確認すると共に，システム構成に向けての指針を導出した.

今後の主な課題としては，① 本章で得た指針を基にシステムの具現化を図り，実システム上での性能評価により実用性を確認すること，② 文書認識のサービス性を左右する認識率の向上を図ること, が挙げられる. 特に後者については，個々の処理アルゴリズムの改良のみならず，文字認識結果を文字切り出しにフィードバックする等のアルゴリズム間のインタラクションを反映した処理制御技術が必要となる．この課題に対し，知識工学の手法であるプロダクションシステムの導入が検討されている[5-17), 5-18)].

第 6 章　文書画像の意味構造認識

6.1　はじめに

文書を紙に代わって電子的に扱うことにより，蓄積・検索の効率化はもとより再利用を図ることも容易となる．近年，文書をデータベース化し，コンピュータネットワークを用いてアクセスするといったシステムの構築が盛んである．一方，依然として文書の保管や配布に紙が用いられることも多く，これら文書の電子化に，前章で述べた文書認識システムの構築が有効である．しかしながら，データベースへの入力に際しては単純にフルテキストを入力するのみならず，検索時のキーとなる意味的構成をも付加した形で入力することが求められ，多大な労力が必要となる．このため，文書画像から意味的構成を自動抽出する構造解析技術が不可欠となる．

文書画像の構造解析に関しては，文書の意味的構成が文書紙面のレイアウトに反映されていることに着目し，構成要素を文書画像中の領域に対応付けて抽出することが考えられる．これまで，構成要素領域の位置関係などをルール化し対象文書に関する知識として用いるモデルベース型の手法が多く提案されている[6-1)~6-4)]．これらの手法は，モデルに構成要素領域の位置関係を記述するため，構成要素の相対位置が変動する場合には適用できないなどの問題があった．

そこで本章では，構成要素領域の位置のみならず，文字の大きさや行の幅などに着目し，これらを画像特徴として捉えることにより，例えば紙面中の図や表など相対位置が固定していない構成要素を含む文書に対しても構造解析可能な手法を提案する．

以下，本章では 6.2 でこれまでの構造解析手法の問題点と提案手法の特徴について述べ，6.3 で提案手法のアルゴリズムの詳細を述べる．6.4 では，提案手法による学術論文誌の論文フロントページを用いた実験の結果と考察について述べ，6.5 でまとめと今後の課題について述べる．

6.2　従来の技術

従来より文書画像の構造解析に関し，構成要素領域の位置関係を記述した文書モデルを用いる手法が提案されている．これらの手法では，構成要素の位置関係が異なる文書ごとにそれぞれモデルを用意する必要があるが，モデルが数種類程度に限定される図書目録カード[6-1)]や縦書き名刺[6-2)]に対しては良い結果が得られている．また，学術論文誌紙面を対象にした例でも，論文フロントページに対してモデルを数種類に限定して効果を上げている[6-3)~6-4)]．しかしながら，中間ページや最終ページにおいて，図表など位置が固定しない構成要素領域を

含む場合には，複数のモデルが必要となる．

そこで，これまでのモデルベース型の解析手法に対し，文書画像から抽出される画像特徴に基づくパターン分類による解析手法[6-5)~6-6)]を提案する．これは，構成要素ごとに文字の大きさや文字間隔，行間隔などが異なることに着目し，これらを画像特徴の違いとして抽出し，総合的に判断することにより構成要素を識別する手法である．提案手法には以下の特徴がある．

(1) 構成要素間の位置関係が固定している必要はない．
(2) 文字列に相当する領域を基本単位とすることにより，構成要素領域が矩形である必要はない．
(3) 画像特徴をベクトル表現することより，距離値に基づく解析結果の確からしさが定義できる．
(4) 文書紙面上の局所並列処理が可能である．

従来このようなパターン分類に基づく手法は，テキスト，表，図領域の分離や，ヘッダ，フッタ，本文領域の分離を目的として研究されてきた[6-7)~6-8)]．しかしながら，データベースへの入力を前提として構成要素を詳細に抽出するまでには至っていない．また，対象文書に依存する処理パラメータを解析手続きの中に一体化しているなど，他文書への適用性も低いものとなっていた．提案手法では，パターン分類に用いる画像特徴をベクトル表現し，学習によって得られた参照ベクトルを対象文書に応じて入れ換えることで，他文書への適用性を確保している．

6.3 パターン分類手法

6.3.1 手法の特徴

提案手法の処理フローを図 6-1 に示す．本手法は，構成要素領域の属性が既知の文書画像を用いた学習過程と，未知の文書画像に対する解析過程に分けることができる．

学習過程では，入力された学習用文書画像に対して傾き補正等の前処理を施し，パターン分類の対象となる基本矩形を抽出する．次に，基本矩形ごとに文字の大きさなどを示す画像特徴を要素とする特徴ベクトルを求める．さらに，基本矩形が属する構成要素名を人間が教示し，特徴ベクトルを統計処理して参照ベクトルを作成する．

解析過程では，学習過程と同様の手法により基本矩形の特徴ベクトルを求める．次に，各特徴ベクトルを学習過程で得られた参照ベクトルと比較することにより，基本矩形を構成要素に分類し，分類結果から構成要素領域を作成する．

以下，処理フローに従って処理の詳細について述べる．

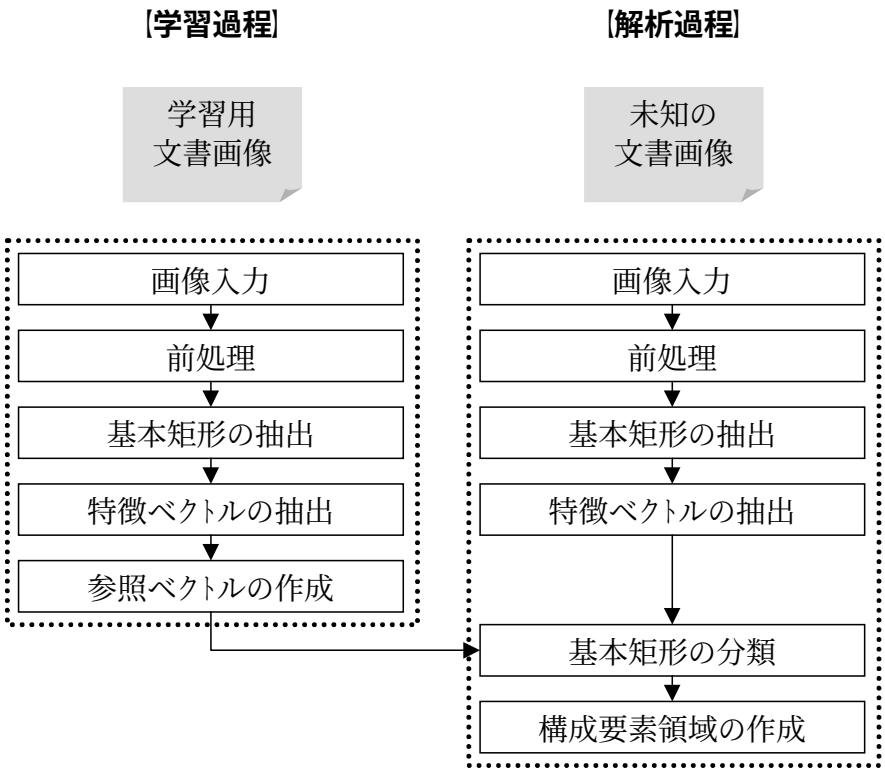


図 6-1　処理フロー

Fig. 6-1　Proposed system overview.

6.3.2　基本矩形の抽出

[画像入力]

文書を 1 ページずつスキャナで読み取り 2 値画像に変換する.

[前処理]

文書画像入力時に混入した周辺ノイズや孤立点ノイズを除去するとともに, 傾きを補正する. 次に, 図 6-2 に示すように紙面中の全ての黒画素を囲む外接矩形を求め, この外接矩形が文書ごとに定めた画像 (サイズは水平方向 X[dot], 垂直方向 Y[dot]) の中心に位置するように画像を移動させる. これは, 後で述べる基本矩形特徴の一つに絶対座標を用いるものがあり, 印字領域の位置を揃える必要があるためである.

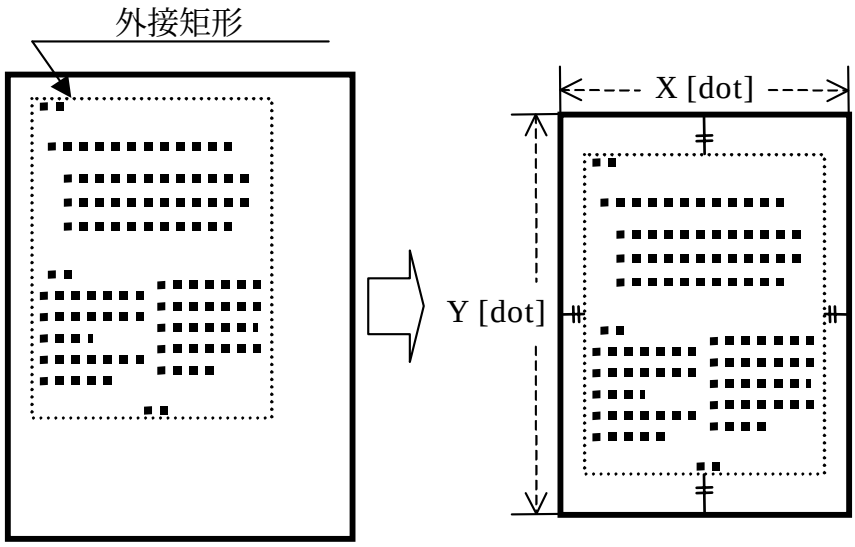


図 6-2　前処理

Fig. 6-2　Preprocessing.

[基本矩形の抽出]

パターン分類の基本単位となる基本矩形を抽出する．まず，8 連結方向で隣接する黒画素領域を求め，これに外接する最小矩形領域を求める．次に，領域の重なりあった矩形をそれらの外接矩形で統合する．この結果得られた矩形のうち，面積がしきい値 (Ac[pixel]) 未満の黒連結矩形をノイズとして除去する．残った黒連結矩形に対し，以下の手続きに従って基本矩形を抽出する．

(1) 黒連結矩形の横方向の射影分布を求め，空白帯の幅がしきい値 (Wv[dot]) より大きい場合，入力画像を横短冊に分割する．

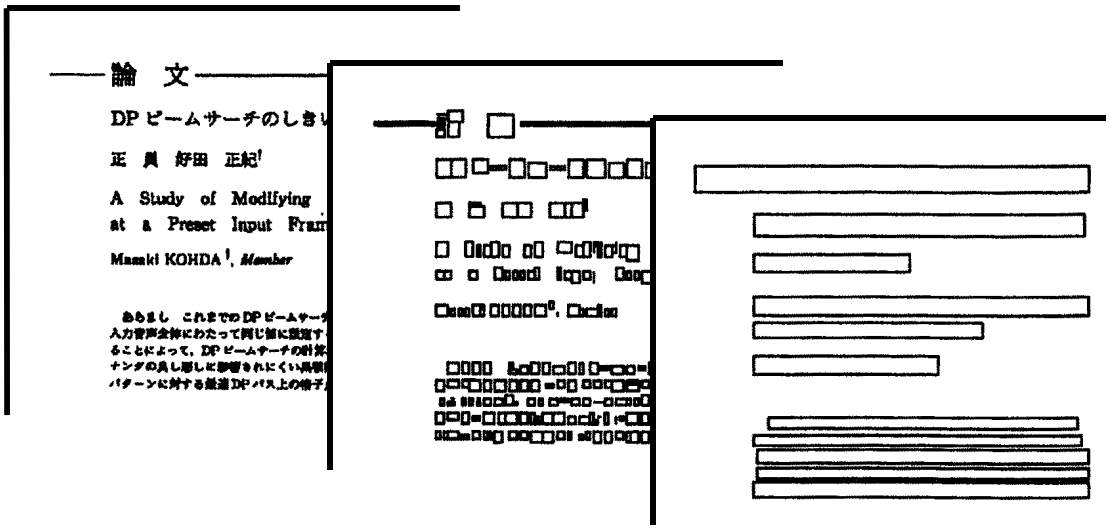
(2) (1) で分割した各矩形領域に対し，今度は矩形内で縦方向の射影分布を求め，同じく空白帯の幅がしきい値 (Wh[dot]) より大きい場合，該矩形を縦短冊に分割する．

(3) (2) で分割した各矩形領域に対し，再度横方向に (1) の処理を行う．

(4) (3) で分割した各矩形領域に対し，再度縦方向に (2) の処理を行う．(3) ～ (4) の処理は，学術論文誌等に多い 2 段組に対応する処理である．

(5) (4) で分割された各矩形領域に対し，面積がしきい値 (Ab[pixel]) より大きい場合，この領域を基本矩形領域とする．

ここで上記の各パラメータは，基本矩形が文字列領域に対応するように値を設定する．図 6-3(a) に示す入力画像 (解像度 400 dots/inch) に対し，Ac=10[pixel]，Wv=1[dot]，Wh=100[dot]，Ab=40[pixel]とした場合の黒連結矩形，基本矩形を同図 (b)，(c) にそれぞれ示す．



(a) 入力画像　　(b) 黒連結矩形　　(c) 基本矩形

図 6-3　黒連結矩形と基本矩形

Fig. 6-3　Connected components and basic rectangles.

6.3.3　参照ベクトルの作成

[特徴ベクトルの抽出]

個々の基本矩形から特徴ベクトルを算出する．特徴ベクトルは表 6-1 に示す 11 個の特徴量から構成される．特徴量の定義を図 6-4 を用いて説明する．ここで, 画像の左上端を原点として水平方向に x 軸, 垂直方向に y 軸とし, 画素数[dot]を基本単位とする座標系を用いている．

行に関する特徴量 lh (行の高さ), ali (上の行との間隔), bli (下の行との間隔)は, 文書が横書きの場合，基本矩形を水平方向に投影した結果得られる矩形を行と見なして算出する．nmb (黒連結矩形数) は基本矩形に含まれている黒連結矩形の数を，sqd (縦横比) は基本矩形の高さ (ey-sy) に対する幅 (ex-sx) の比を，dnst (黒連結矩形密度) は基本矩形の面積 (area) に対する中に含まれている黒連結矩形の総面積の割合を示している．概念的には，sx (左上 x 座標) は基本矩形の左マージンを，lh (行の高さ) は文字の大きさを，nmb (黒連結矩形数) は基本矩形中の文字数を，dnst (黒連結矩形密度) は文字のつまり具合を反映している．

以上の 11 個の特徴量を要素とするベクトルを基本矩形の特徴ベクトル Vb とする．

$$Vb = (sx, sy, ex, ey, lh, ali, bli, nmb, sqd, area, dnst)$$

表 6-1　基本矩形特徴量

Table1　Basic rectangle features.

特徴量	記号
左上 x 座標	sx
左上 y 座標	sy
右下 x 座標	ex
右下 y 座標	ey
行の高さ	lh
上の行との間隔	ali
下の行との間隔	bli
黒連結矩形数	nmb
縦横比	sqd
面積	area
黒連結矩形密度	dnst

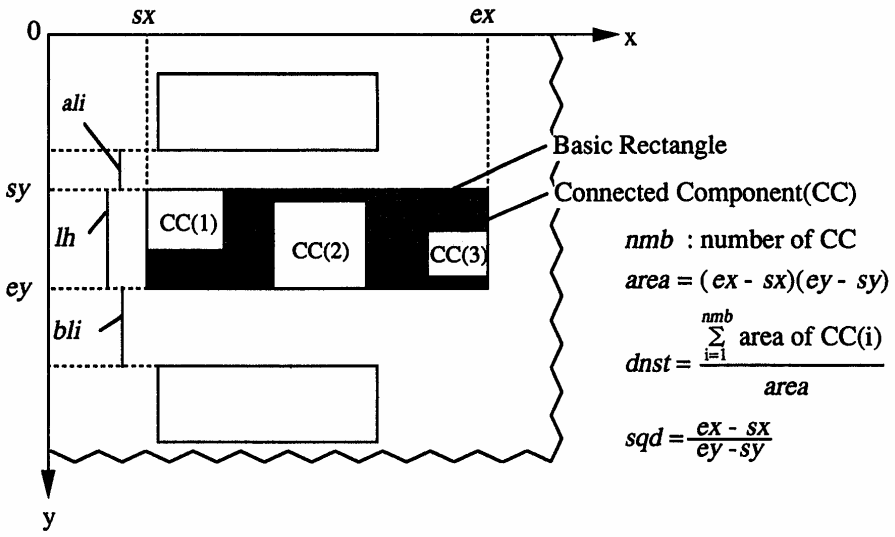


図 6-4　基本矩形特徴

Fig. 6-4　Basic rectangle features.

[参照ベクトルの作成]

学習過程において，判別分析によりパターン分類のための参照ベクトルを求める．

本手法では正準ベクトル法を用いる[6-9]．次のようなp×n行列Vを考える．ここでpは特徴ベクトルの次元数, nは学習に用いる基本矩形の総数である．まず,

あらかじめ手動により学習に用いる参照画像から得た基本矩形を全て正しい構成要素に対応づけておく．構成要素がr種類ある場合，第j構成要素に属する基本矩形の数がK_jであるとする $(K_1+K_2+\cdots+K_r = n)$. Vの最初のK_1行は第一構成要素に属する特徴ベクトルを行としてならべたもの，次のK_2行は第二構成要素の特徴ベクトル，以下，第r構成要素まで同様にまとめられているとする．ここで，行列Vにある線形変換Aを施したV･Aが各構成要素をなるべく明確に区別するようなp×q行列Aを求める．このために以下の様なn×r行列Uを考え，同じ構成要素に属するU・Bの要素が同じ値を持つようにする．

$$U = \begin{bmatrix} K_1^{-1/2} & 0 & 0 & \cdots & 0 \\ K_1^{-1/2} & 0 & 0 & \cdots & 0 \\ : & : & : & : & : \\ 0 & K_2^{-1/2} & 0 & \cdots & 0 \\ 0 & K_2^{-1/2} & 0 & \cdots & 0 \\ : & : & : & : & : \\ 0 & 0 & 0 & \cdots & K_r^{-1/2} \end{bmatrix}$$

ここで，V・A=U・B (B は群対比ベクトルからなる行列となる) と考えると，V と U について正準相関分析を行うことにより，正準判別変量 V・A を求めることができる．この方法では，正準判別変量全体のばらつきを示す全分散のうち，正準判別変量の構成要素ごとの平均値のばらつきを示す群間分散が占める割合が最大となる正準判別変量を求めていることになる．そこで，このようにして求めた正準判別変量の構成要素ごとの平均ベクトルを本手法の参照ベクトルとする．

6.3.4 基本矩形の分類

[基本矩形の分類]

解析過程においては，前述した特徴ベクトルと参照ベクトルを比較することにより基本矩形を構成要素に対応づける．

まず，すでに求められている特徴ベクトルを上述した変換行列 A により変換する．次に，変換の結果得られたベクトルと全ての構成要素の参照ベクトルとのユークリッド距離をそれぞれ求め，それらのうち最も距離の短い構成要素に基本矩形を分類する．

[構成要素領域の作成]

ここまでの処理で，各基本矩形に距離値をもった構成要素が対応付けられた．この結果を基に各構成要素に対応する領域を作成する．本手法では，同じ構成

要素を第一候補にもつ基本矩形を囲む外接矩形をその構成要素の領域とする．他の構成要素領域との関係を考慮せずに構成要素ごとに処理するため，構成要素領域が重なる場合も想定される．

6.4 実験と考察

6.4.1 実験の目的と評価尺度

6-3 で述べた構造解析手法が実際の文書画像に対し有効であることを示すため，評価実験を行った．実験では，入力文書画像より抽出した基本矩形の分類精度と，分類された基本矩形より生成される構成要素領域の生成率をそれぞれ測定した．

測定には，Leaving-one-out法[6-10]を適用した．実験対象となる全文書画像から 1 枚を取り除いた残りの文書画像から生成される全基本矩形を用いて参照ベクトルを作成し，取り除いていた 1 枚の文書画像から生成される全基本矩形を分類する．

評価尺度としては，以下に示す基本矩形分類率と領域生成率を用いた．

$$\text{基本矩形分類率} = \frac{\text{正しく分類された基本矩形数}}{\text{全基本矩形数}} \tag{1}$$

$$\text{領域生成率} = \frac{\text{正しく生成された領域数}}{\text{全領域数}} \tag{2}$$

6.4.2 実験の詳細と結果

[評価実験の詳細]

実験には, 電子情報通信学会論文誌に収録されている論文フロントページ (以下IEICEと呼ぶ) 110 枚と情報処理学会論文誌に収録されている論文フロントページ (以下IPSJと呼ぶ) 118 枚を用いた．論文フロントページは，意味的構成要素を多く含んでいること，またフルテキストデータベース化[6-11]に際し重要度が高いことから実験対象とした．各論文誌の意味構成要素の抽出対象は，IEICEで 13 項目，IPSJで 14 項目である．図 6-5 にIEICEの構成要素領域を示す．但し，右ページ番号 (Right page number) に代わりに左ページ番号 (Left page number) となる場合があり, これらを個別に扱ったため全体で 13 項目となる．

実験は，ワークステーションを用いて行った．また，文書画像の入力にはワークステーションに接続されたイメージスキャナを用いた．

画像は解像度 400[dots/inch]で, 論文誌をスキャナの走査面に直接押し当てて入力した. 前処理後の画像の大きさは, 水平方向 2800[dot], 垂直方向 3800[dot] である．基本矩形を求める際のパラメータは，図 6-3 の場合と同様に

Ac=10[pixel], Wv=1[dot], Wh=100[dot], Ab=40[pixel]とした．Wh は，本文の左右の段にまたがる基本矩形領域が生成されないように値を設定した．

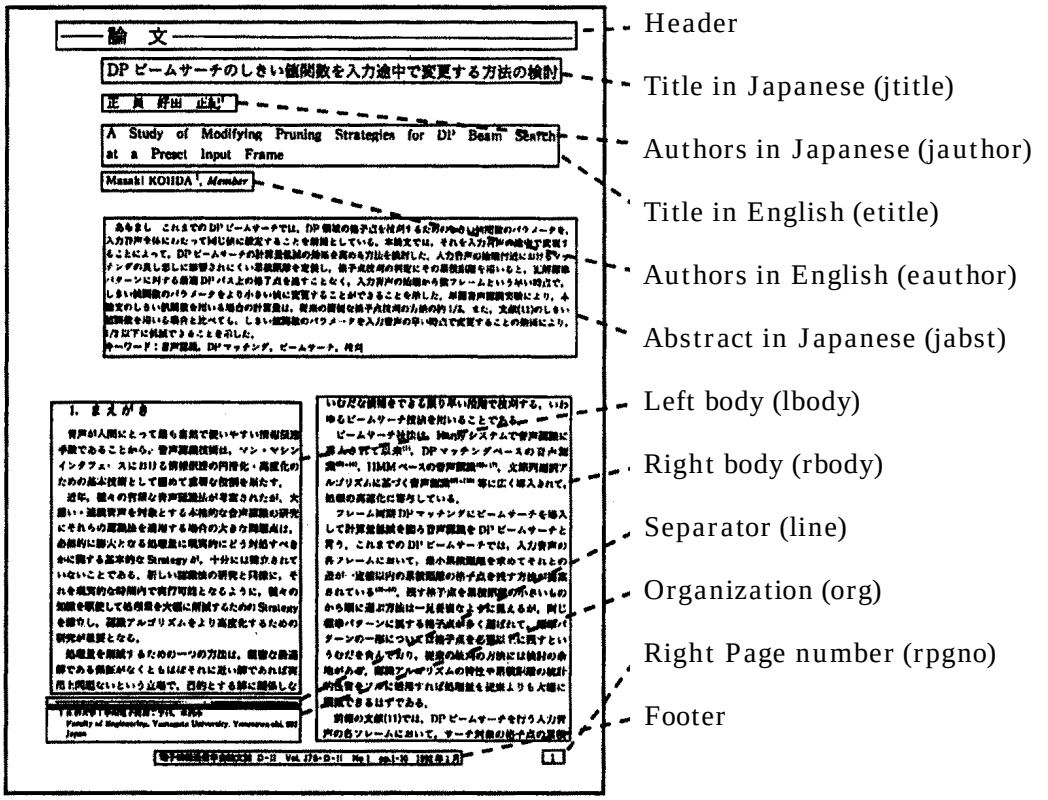


図 6-5　IEICE の構成要素領域

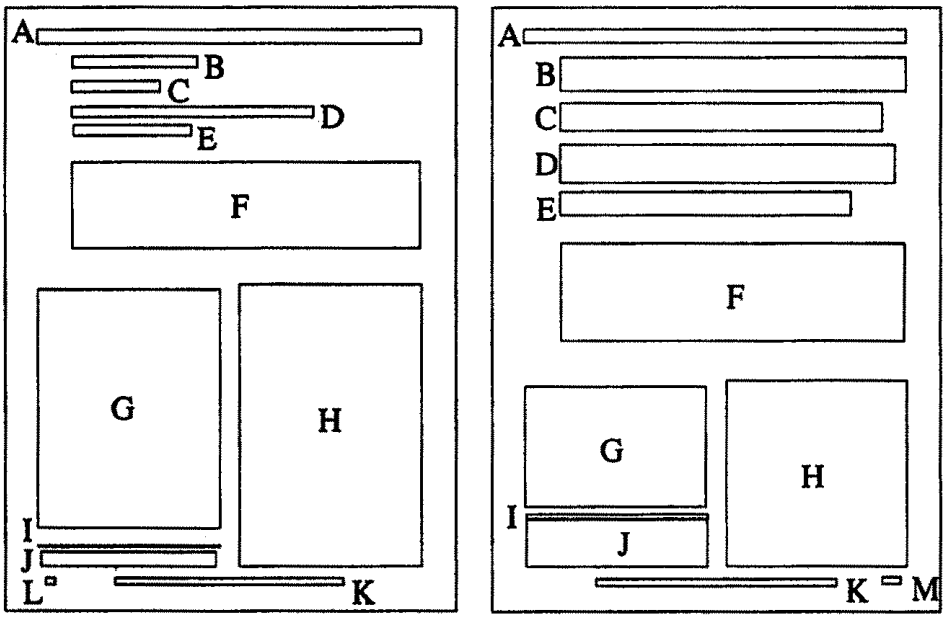
Fig. 6-5　Logical component regions of IEICE title page.

[評価実験の結果]

IEICE，IPSJ 全ページの基本矩形の総数は 14,760 個，基本矩形分類率は 98.5 %，構成要素領域の総数は 2,972 個，領域生成率は 89.8 %であった．構成要素別の基本矩形分類率および領域生成率を表 6-2 に示す．この結果から，提案した特徴ベクトルで基本矩形が精度よく分類できることが分かった．図 6-6 に，構成要素領域が 100%正しく抽出できた例を示す．

誤分類の分析を行った．表 6-3 に基本矩形の累積分類率を示す．IEICE では第 3 候補までに，IPSJ でも第 4 候補までに全ての正解が含まれていた．すなわち，オペレータの確認修正作業を考慮すれば，正しい構成要素を効率的にオペレータに提示できることが分かった．

基本矩形の誤分類パターンを分析した結果を表 6-4 に示す．表 6-4 で，「和文標題 (jtitle)」，「あらまし (jabst, eabst)」，「本文 (lbody)」の中の段落最終行が他行に比べて幅が短いため，他の構成要素に誤分類された例が 133 個あり，実験全体の誤分類総数 (222 個) のうち 59.9 %を占めている．また，IEICE の「右本文 (rbody)」が「ページ番号 (rpgno)」に誤分類されている基本矩形 (14 個) は，全て右本文中の数式の番号が記述されている比較的幅の短い基本矩形であった．さらに，IPSJ の「英文標題 (etitle)」が「英文著者表記 (eauthor)」に誤分類されているもの (13 個) は，そのほとんどが基本矩形の高さが低い副題であった．これらの誤分類例を図 6-7 に示す．このように誤分類の大半をいくつかのパターンに分類できることがわかった．



A : Header
B : Title in Japanese (jtitle)
C : Authors in Japanese (jauthor)
D : Title in English (etitle)
E : Authors in English (eauthor)
F : Abstract in Japanese (jabst)
G : Left body (lbody)
H : Right body (rbody)
I : Separator (line)
J : Organization (org)
K : Footer
L : Left page number (lpgno)
M : Right page number (rpgno)

図 6-6　IEICE 論文フロントページ画像の解析結果

Fig. 6-6　Analysis results of IEICE title pages.

表 6-2　構成要素別基本矩形分類率と領域生成率

Table2　Experimental results.

IEICE			IPSJ		
構成要素名	基本矩形分類率 (%)	領域生成率 (%)	構成要素名	基本矩形分類率 (%)	領域生成率 (%)
header	100	100	vol	100	99.2
jtitle	89.3	82.7	jname	100	100
jatuthor	98.3	82.7	idate	100	100
etitle	96.3	93.6	jtaitle	89.2	86.4
eauthor	99.3	99.1	jauthor	99.8	87.3
jabst	95.3	92.7	jabst	99.0	86.4
lbody	99.5	59.1	etitle	93.8	87.3
rbody	99.4	98.2	eauthor	92.5	53.4
line	100	100	eabst	96.9	65.3
org	100	92.7	lbody	98.7	89.8
lpgno	100	100	rbody	100	100
rpgno	100	96.2	line	100	100
footer	100	100	org	98.5	82.2
			pgno	100	99.2
Total	98.5	91.6	Total	98.4	88.3

表 6-3　基本矩形の累積分類率

Table 3　Cover rates.

文書の種類	IEICE	IPSJ
基本矩形総数	7089	7671
第１候補	98.6 %	98.5 %
第２候補まで	99.8 %	99.7 %
第３候補まで	100　%	99.9 %
第４候補まで	―	100　%

(IEICE では第 3 候補までに，IPSJ では第 4 候補までに全て分類される)

表 6-4　基本矩形の主な誤分類の分析

Table4　Analysis results of wrong classification.

文書	正しい分類先	実際の分類先	個数	全体に占める割合	誤分類された矩形の特徴
IEICE	jabst	lbody	46	44.7 %	最終行の矩形で長さが短い
	jtitle	jauthor	14	13.6 %	最終行の矩形で長さが短い
	rbody	rpgno	14	13.6 %	数式の番号を表している矩形で長さが短い
	IEICE 誤分類合計		103	100　%	
IPSJ	eabst	eauthor	30	25.2 %	最終行の矩形で長さが短い
	jtitle	jauthor	15	12.6 %	最終行の矩形で長さが短い
	lbody	org	15	12.6 %	段落最終行の矩形で長さが短い
	etitle	eauthor	13	10.9 %	副題で矩形の高さが低い
	jabst	eauthor	13	10.9 %	最終行の矩形で長さが短い
	IPSJ 誤分類合計		119	100　%	

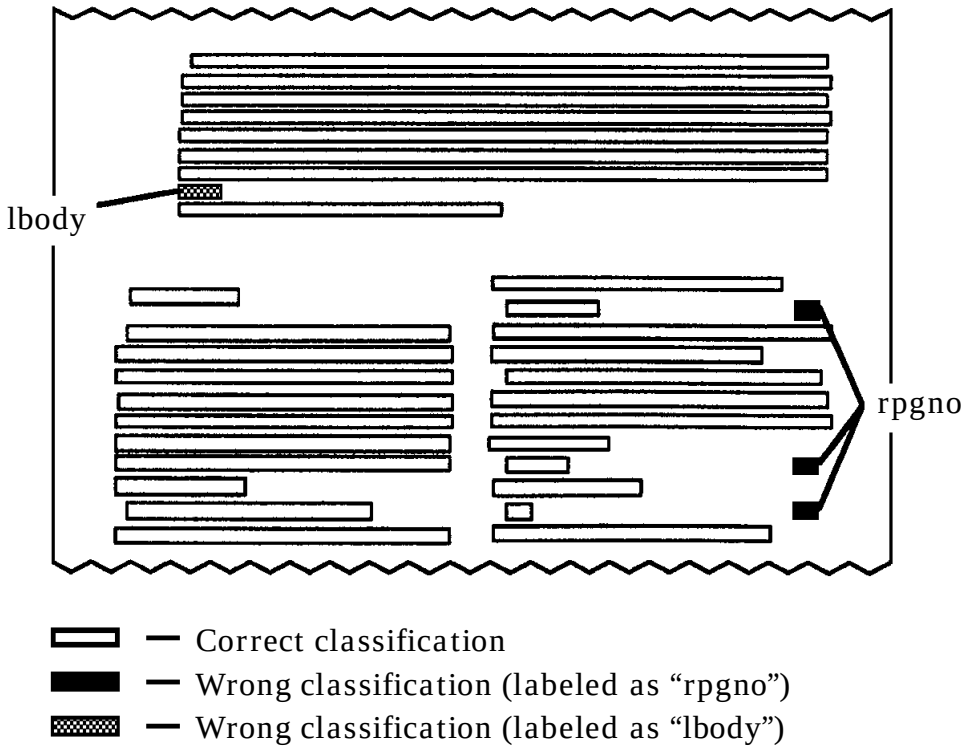


図 6-7　基本矩形の分類例

Fig. 6-7　Example of labeled rectangles.

6.4.3 考察

6.2 で述べた提案手法の特徴について，評価実験の結果と対比させながら考察する．

(1) 対象文書中の構成要素間の相対関係が安定している必要はなく，例えば図 6-8 のように左本文 (lbody) のない論文フロントページに対しても解析可能である．従来手法では，モデルに構成要素間の位置関係を記述しているため，先の例のように特定の構成要素領域が欠如する場合に対しては別にモデルを追加する必要があった．また，提案手法では，着目する基本矩形の識別は他の基本矩形の識別結果に影響されない．したがって，従来の多くの知識ベース型解析手法のように文書画像全体を解析する必要はなく，必要な領域のみを指定して効率よく解析することが可能である．

(2) 文字列に相当する領域を分類の基本単位としているため, 基本矩形領域を統合することにより，矩形では表現できない構成要素領域を抽出することが可能である．

(3) 参照ベクトルとの距離を指標として基本矩形を分類するため, 基本矩形に対し，参照ベクトルとの距離の近いものから順に確からしい構成要素を提示することができる．また，

(第一候補の距離 $\leqq p$) かつ (第一候補と第二候補との距離差 $< q$)

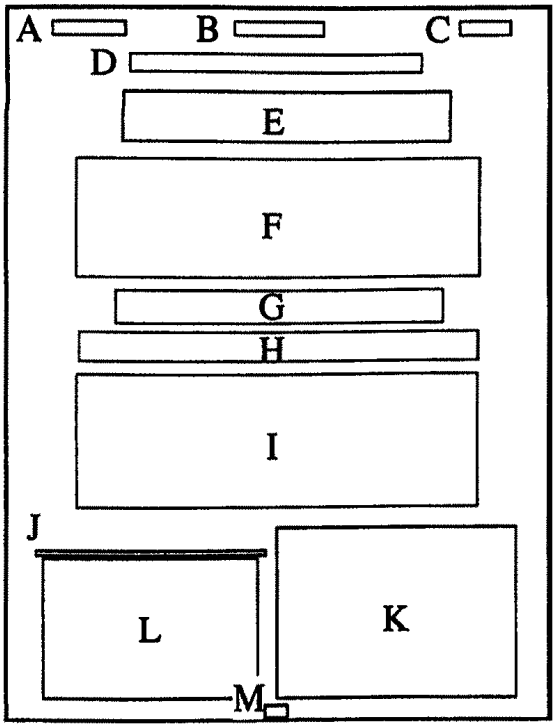
p, q：確からしさのしきい値

となる場合には，第一候補が第二候補に比べて確からしいとは断定できないとして強調表示し，オペレータに注意を促すなど，オペレータによる確認修正作業の操作性を向上させることが考えられる．$p = 3.5$，$q = 3.0$ として基本矩形を抽出した結果を表 6-5 に示す．ここで，再現率とは誤分類のうち上記の条件式で抽出できたものの割合を示し，適合率とは上記条件式で抽出した基本矩形のうち誤分類であったものの割合を示している．

(4) 個々の基本矩形の特徴ベクトルをそれぞれ独立して算出することができるため，並列処理による高速化が可能である．また，基本矩形の識別に際しても，従来手法のように相互関係を用いていないため，並列処理が可能である．これらの点から，処理の高速化が可能な手法である．

提案手法の課題について述べる．一般に知識ベース型の手法では，文書モデルの中に構成要素領域の階層構造が記述されており，解析結果として構造化された構成要素領域が抽出される．一方，提案手法では個々の基本矩形を単に構成要素に対応づけているため，特定の構成要素を抽出したい場合には効率的であるが，フルテキストデータベースの作成において文書の構造化が必要な場合，

構成要素間の関係づけが必要となる．実験に用いた論文フロントページでは，構成要素と意味構造のインスタンスとが一意に対応づけられるため，構造化は容易である．しかし，例えば名刺の住所欄などのように一つの構成要素の中に複数のインスタンスが含まれる場合には，構成要素領域の構造化が困難となる．



A : Volume number (vol)
B : Journal name (jname)
C : Issue date (idate)
D : Title in Japanese (jtitle)
E : Authors in Japanese (jauthor)
F : Abstract in Japanese (jabst)
G : Title in English (etitle)
H : Authors in English (eauthor)
I : Abstract in English (eabst)
J : Separator (line)
K : Right body (rbody)
L : Organization (org)
M : Page number (pgno)

図 6-8　本文 (左) のない IPSJ 論文フロントページ画像の解析結果
Fig. 6-8　Analysis result of IPSJ title page with no left body.

表 6-5　距離値を用いた誤り指摘による再現率と適合率
Table5　Recall and precision rates.

文書	指摘数	再現率	適合率
IEICE	696	90.3%	13.4%
IPSJ	635	92.4%	17.3%

6.5 まとめ

構成要素の位置，行間隔，文字の大きさ，文字数などの特徴を画像特徴として捉え，これらを総合的に判断して文書画像の意味構造を解析する手法を提案した．提案手法は，文字列領域に相当する基本矩形領域から特徴ベクトルを算出し，予め作成した参照ベクトルと比較して構成要素に対応づけるものである．利点としては，構成要素の位置関係が定まらない場合や矩形表現できない場合でも抽出可能となること，構成要素の確からしさを表現できること，基本矩形単位に並列処理が可能となることを述べた．また，論文フロントページを対象とした実験により 98.5 %の基本矩形分類率を得たことを示し，本手法の有効性を確認した．

以下，今後の課題について述べる．第一に，本手法は意味的構成要素領域の抽出を主目的としているが，データベースへの入力に際しては構成要素間の関係記述が求められる場合があり[6-12]，さらにこうした意味構造の抽出が必要となる．第二に，領域生成率をさらに改善する必要がある．これには，基本矩形分類率の改善に加え，基本矩形から構成要素領域を決定する際，構成要素領域自身の画像特徴に着目し，基本矩形の識別誤りを検出して自動訂正することが考えられる．

第 7 章　学術論文誌認識システムの実用化

7.1　はじめに

各種学会から出版される学術論文誌をデータベース化し，一般に提供するサービスとしては，国内では科学技術振興事業団のオンライン情報検索システムJOIS(JICST Online Information System) [7-1)]などを挙げることができる．利用者は，論文タイトルや著者名などをキーとして，所望の論文の書誌情報や抄録を検索することができる．また，論文誌のページ複写を郵送やファクシミリにより受け取ることもできる．

一方，書誌情報のみならず，本文，図表，引用文献，著者所属機関など論文に記載される一次情報を全てデータベース化する試みが行われている．これにより，本文に対するフルテキスト検索[7-2)]に加え，著者名と所属機関名など項目間の関連検索も可能となり，検索の利便性向上に有効である．こうしたねらいのもと，学術論文を国際規格であるSGML (Standard Generalized Markup Language) [7-3), 7-4)]を用いて予め記述する試みも行われている．

しかしながら，年間約67万件もの論文等をデータベース化しているJOISや，国内外で試みられている電子図書館[7-5), 7-6)]の構築に際しては大量の遡及文書を扱う必要があるため，これらのフルテキスト化，SGML文書化については依然として書誌情報に限定される場合が多い[7-7), 7-8)]．そこで，遡及論文誌からのフルテキスト，および項目や項目間のリンクを自動抽出し，データベース構築作業を支援するシステムが望まれる．

本章では，前章までに述べた文書画像構造解析をさらに拡張し，論文誌のページ画像から論文を構成する項目 (以下意味構成要素と呼ぶ) の抽出，ならびに項目間，例えば著者名と著者所属機関名の関係づけを行うことにより，論文単位の構造化文書を生成するシステムを構築した結果について述べる．本システムでは，文書の記述に先に触れたSGMLを用いており，近年普及の著しいWWW (World Wide Web) をはじめとする他のアプリケーションで扱われる文書への変換も可能である．

以下，7.2では学術論文誌からSGML文書を生成する際の課題について述べる．7.3では，これらの課題を解決するための構造解析手法について述べる．7.4では，提案手法を実装した学術論文誌認識システムについて述べ，実際に構成したシステムの評価実験結果について考察を加える．

7.2　システム設計における課題

SGMLで記述される文書には，文書の意味構造を定義する文書型定義 (DTD :

Document Type Definition) が必要となる．和文の学術論文を対象としたDTDの例としては，情報知識学会学会誌に適用されているDTD[7-9]や，科学技術振興事業団が試作した情報管理誌のDTD[7-10]などがある．これらのDTDは，SGML文書データベースの検索を前提として設計されており，著者名や著者所属機関名，本文中の章題，パラグラフ，図表，参考文献などの意味構成要素が定義されているほか，著者名と著者所属機関名のリンク関係などの記述についても定義されている．

ここで，遡及論文誌からOCR (光学文字読み取り) 技術を用いてSGML文書を生成することを考える．これまで，文書画像の文字読み取りに関する研究や，また文書の意味構成要素を抽出する文書画像構造解析に関する研究が盛んに行われている．特に，意味構成要素の抽出に関し，新聞紙面を対象としたルールベースに基づく手法[7-11]，ODA (Open Document Architecture：開放型文書体系) の文書モデルを用い，技術情報誌を対象とした節番号部等の解析による手法[7-12]，論文誌を対象とし，部分領域の相対的な位置関係などのレイアウト規則に着目した手法[7-13), 7-14), 7-15]などが提案されている．しかしながら，先に述べたSGML文書のDTDに対応した意味構成要素を抽出し，要素間の関係づけを含めてSGML文書化するには，未だ以下の課題が残されている．

(a) レイアウトが変動するページ画像からの部分領域の抽出

従来の研究では，論文フロントページなどレイアウトの比較的固定的な文書画像から意味構成要素に対応した部分領域を抽出する場合について論じている[14]．しかしながら，論文の中間ページや最終ページでは，図表，参考文献，著者紹介などのレイアウト位置が変動するため，こうしたページ画像に対しても図や表など非テキスト要素を含むこれら部分領域を抽出することが必要となる．

(b) 部分領域の意味構成要素への対応づけ

従来の研究では，レイアウト解析により抽出された部分領域の意味構成要素への対応づけが，先に述べたDTDで定義されている意味構成要素と合致しない場合があった．例えば，複数の著者が記述されている部分領域では，個々の著者名領域を著者ごとに分けて扱い，それぞれの所属機関名と対応づける必要がある．

(c) 論文単位の構造化

イメージスキャナより入力される文書画像は，一般にページ単位に扱われるため，従来の研究では構造解析結果をページ単位のまま扱っている場合が多い．しかしながら，学術論文は連続する複数のページから構成され，DTDも論文単位に定義されることから，個々のページ解析結果を統合して，論文単位のSGML文書を生成することが必要となる．

以下，次節では，本節で述べた課題を解決し，学術論文誌からSGML文書を生成する手法について述べる．

7.3 システムの実現方式

論文誌のページ画像からSGML文書を生成する場合の基本的な処理フローを図7-1に示す．

① 個々のページ画像に対して，7.3.1で述べる手法によりブロック分けした部分領域を抽出する．

② 部分領域内の書式情報 (例えば，文字の大きさや紙面上の位置，前後の行間隔などの印字規則) がDTDで定義された意味構成要素ごとに異なることに着目し, 部分領域を意味構成要素へ対応づける. 後述する提案手法において, 書式情報を反映させた11種類の画像特徴から成る特徴ベクトルを部分領域ごとに抽出して対応づけを行う．

③ テキスト化すべき部分領域に対しては文字認識処理を施すとともに, その他の図表等の部分領域は画像符号化を行い, ページ単位にSGML文書の素材となる論理タグ付き文書を生成する．

④ 論文誌の各ページに対して処理を行った後, 一論文を構成する複数ページにまたがる文章のつなぎ合わせ処理等を行い, 論文単位のSGML文書を生成する．

本研究で対象とするDTDについては，現在情報知識学会学会誌等に適用されている和文の学術論文を対象としたDTDの構造[7-9]を参考に作成したものを用いる. 作成したDTDを付録に, 定義されている要素や要素間の関係を図7-2に示す. 但し，本DTDは書誌情報として扱われる意味構成要素や本文等検索時に利用価値の高い意味構成要素に限定し，脚注や付録を除外したものとなっている．一方，検索時に論文誌のページを参照できるようにするため，ページ番号を追加している．

以下，上記処理フローに従って処理の詳細を述べる．但し，入力画像は2値画像とし，前処理としてスキャン時に混入した周辺ノイズや孤立点ノイズを除去するとともに, 傾き補正を行う. また, 入力画像の左上点を原点に横方向をx軸, 縦方向をy軸とし，画像中の全ての黒画素を含む外接矩形の左上端点が (x,y) = (SX,SY), (SX, SYは固定値) となるよう平行移動することにより，ページ画像中の印字領域の位置を揃える．

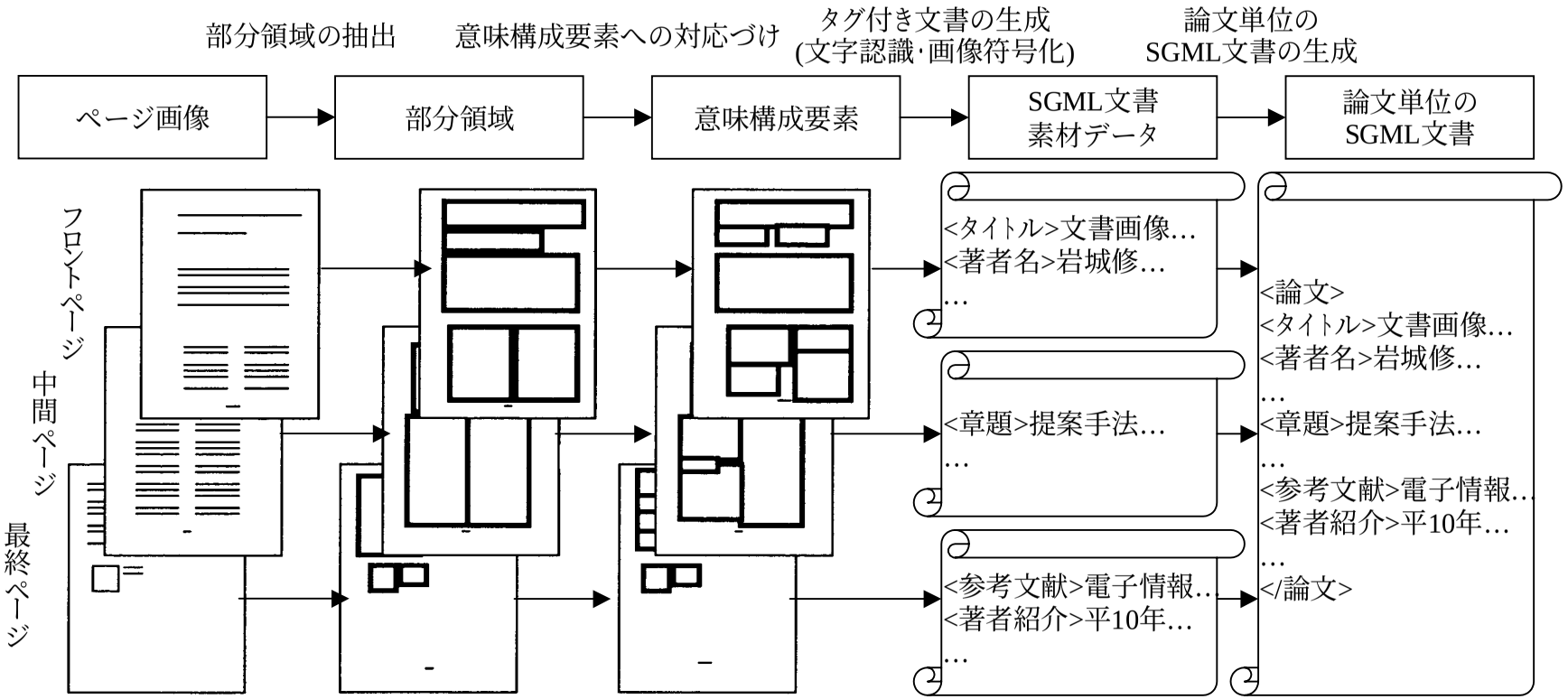
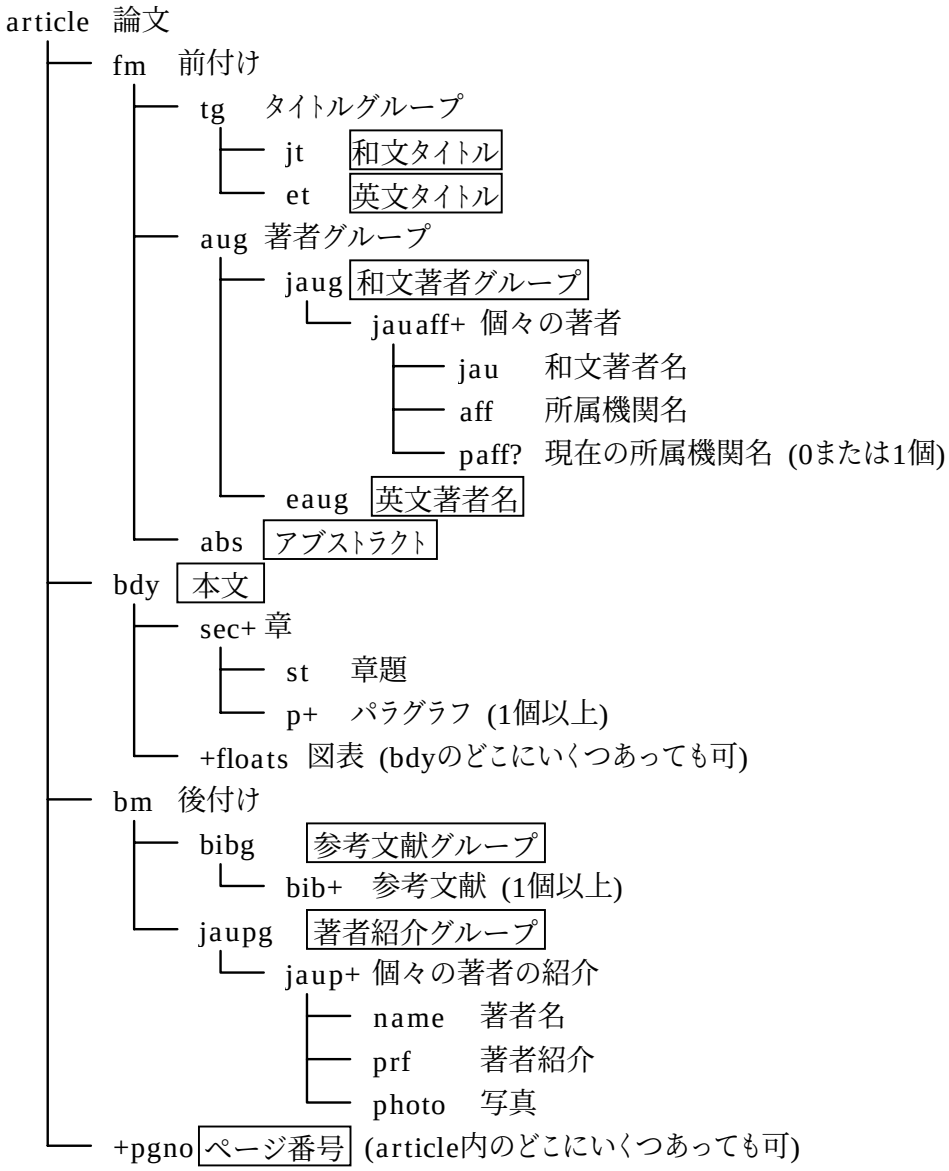


図 7-1　処理フロー

Fig. 7-1　Process overview.



要素名の後の＋はその要素が 1 個以上いくつあっても良いことを, 要素名の後の？はその要素が 0 個または 1 個であることを, 要素名の前の＋はその要素がそれを含む要素中のどこにいくつあっても良いことをそれぞれ示している. また, □で囲まれた要素は, 部分領域の抽出過程で抽出される領域であることを示している.

図 7-2　SGML DTD の構造

Fig. 7-2　Structure of SGML DTD.

7.3.1 部分領域の抽出

部分領域の抽出は，基本矩形の抽出とその分類からなり，パターン分類に基づく構造解析手法[7-16]を用いて実現する．ここで，基本矩形とは8連結黒画素に外接する矩形のうち一定面積以上の近接する黒連結矩形を統合したもので，例えば個々の文字列に相当する領域である．次に，基本矩形の位置座標や大きさ，縦横比など11種類の画像特徴からなる特徴ベクトルが書式情報を反映していることに着目し，タイトル，アブストラクト，本文など予め学習して求めた意味構成要素ごとの参照ベクトルと，入力画像から抽出した基本矩形の特徴ベクトルを比較し，基本矩形を部分領域に分類する．用いる画像特徴が基本矩形の位置のみならず文字サイズなどの書式情報を反映していることから，図や表などレイアウト位置が固定的でない要素が含まれているページ画像でも構造解析可能となる．なお，参照ベクトルは以下のページ種別ごとに学習して作成する．

- フロントページ
  タイトルを含むページ．論文の最初の1ページ．
- 中間ページ
  フロントページ，最終ページ以外のページ．
- 最終ページ
  参考文献または著者紹介を含むページ．通常，論文の最後のページ，あるいは最後から2〜3ページ．

本処理により，表 7-1 に示す部分領域名がそれぞれの基本矩形に付与される．図 7-3 は，同じ部分領域名をもつ隣接する基本矩形を統合することによって最終的に得られる部分領域の例である．

表 7-1　抽出される部分領域

Table 7-1　Partial regions extracted from technical journals.

ページ種別	部分領域名
フロント	和文タイトル，英文タイトル，和文著者グループ (和文著者名と所属機関名)，英文著者名， アブストラクト，本文，ページ番号
中 間	本文，ページ番号
最 終	本文，参考文献グループ，著者紹介グループ (著者名と写真と著者紹介)，ページ番号

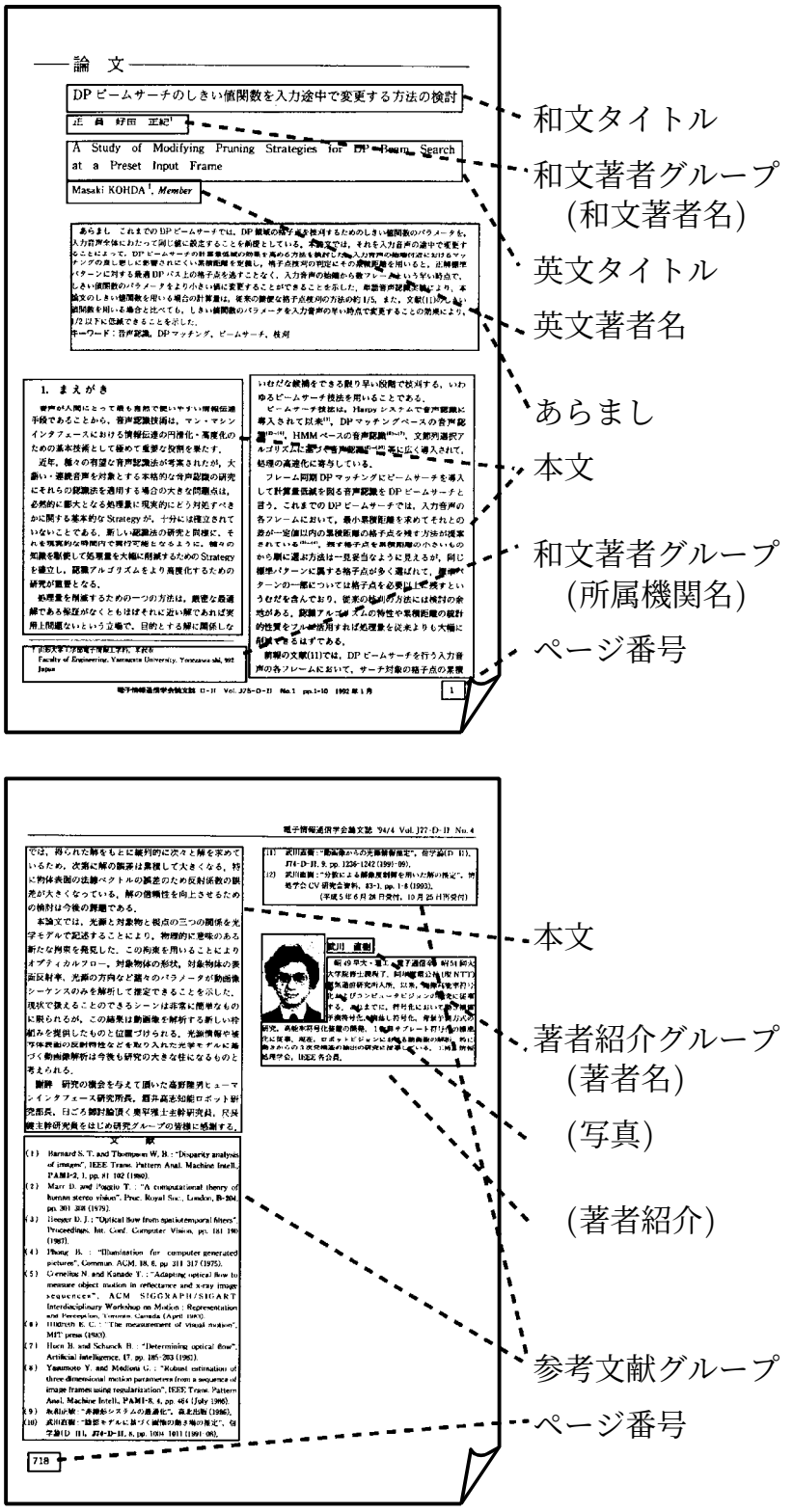


図 7-3　部分領域生成例

Fig. 7-3　Examples of partial regions.

7.3.2 意味構成要素への対応づけ

抽出した部分領域を図 7-2 に示した DTD で定義された意味構成要素に対応づける．表 7-1 に示した部分領域のうち，和文タイトル，英文タイトル，英文著者名，アブストラクト，ページ番号はそのまま意味構成要素に対応させることができる．一方，和文著者グループ，本文，参考文献グループ，著者紹介グループは，図 7-2 に示した DTD で定義される意味構成要素に対応づけるため，各部分領域をさらに細分化する必要がある．具体的には，これら部分領域に対し表 7-2 に示した以下の処理を行う．

表 7-2　部分領域の意味構成要素への対応づけ
Table 7-2　Relations between the partial regions and logical elements.

部分領域名	処理内容
和文著者 グループ	個々の和文著書名領域，所属機関名領域， 現在の所属機関名領域の抽出と対応づけ
本文	章題領域，パラグラフ領域の抽出と対応づけ
参考文献 グループ	個々の参考文献領域の抽出
著者紹介 グループ	個々の著者名領域，写真領域，著者紹介領域の抽出と対応づけ

7.3.2.1 和文著者グループ

和文著者グループに含まれる著者ごとの和文著者名，所属機関名，および現在の所属機関名領域を抽出し，それらの対応づけを行う．和文の学術論文では，和文著者名領域に複数の著者名が記され，所属機関が異なる場合，一般に対応づけのための記号が印字される．例えば，電子情報通信学会論文誌や情報処理学会論文誌の場合には，図7-4に示すように各和文著者名の右肩に†や＊が印字される．

そこで，まず和文著者名領域から以下に述べる方法によりこれら記号の種類と数を抽出する．具体的には，基本矩形に含まれる黒連結矩形の中から†または＊などの記号に対応する矩形を抽出する．これは，和文著者名の記述に使用される文字に比べて記号が小さいことを利用し，黒連結矩形の横幅および高さから記号を特定する．次に，抽出した記号より左側をそれぞれ個々の和文著者名領域とし，この領域の座標値と†，＊などの記号の数を所属機関名への対応づけの情報として出力する．

一方，著者所属機関名領域では，図7-5に示すように複数の機関名が記される場合，機関名ごとに著者名と対応づけるための記号が同様に印字される．そこで，和文著者名領域と同様，記号の種類と数を求め，右側に位置する機関名領域を抽出する．また，所属機関名は，通常複数行にわたって記述されるので，左側に記号のない基本矩形をその上の基本矩形と統合することで個々の機関名を表す領域を抽出する．

これらの処理結果を基に，個々の和文著者名領域に対して，それぞれの記号の数が一致する著者所属機関名領域を求め，対応関係を出力する．対応関係のDTD上の記述に関しては，付録に示したDTDの (1) の部分にあるように，要素の階層構造により定義している．

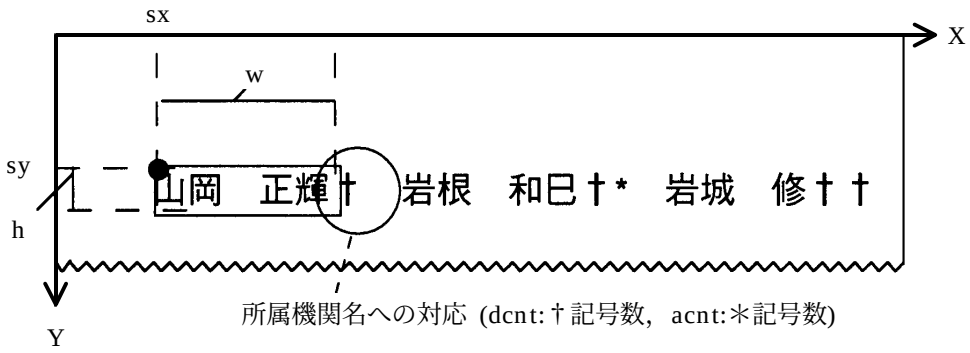


図 7-4　和文著者名が記述される領域の例

Fig. 7-4　Example of regions in which Japanese authors' names are printed.

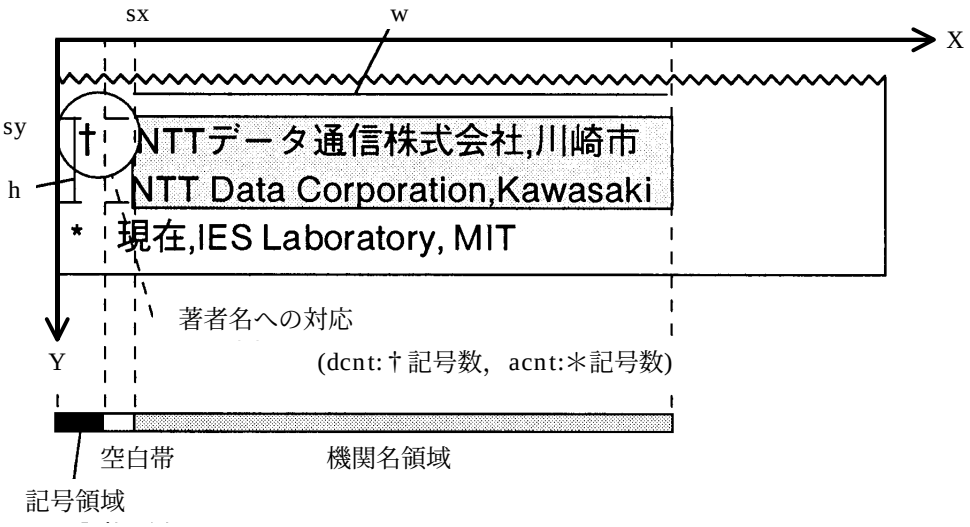


図 7-5　所属機関名が記述されている領域の例

Fig. 7-5　Example of regions in which names of authors' organizations are printed.

7.3.2.2 **本文**

本文領域に分類された基本矩形から，章題とパラグラフを抽出するとともに，章題とパラグラフの対応づけを行う．但し，本稿では箇条書きの各項目や，行を変えて記載される数式をひとつのパラグラフとして扱う．以下，部分領域の抽出では，図7-3に示したような2段組の場合，本文領域が2つ抽出されるので，それぞれの領域を個別に処理する．

まず，基本矩形を章題とパラグラフに分類する．章題は前後の行間隔がパラグラフ内の行間隔よりも広いことに着目する．また，章題が複数行にわたる場合も想定する．図7-6に示すように，行間隔に相当するパラメータh (例えば解像度400 dot / inchの画像で50程度) を用いて (1) から (4) までの条件に該当する基本矩形を章題に分類し，その他の基本矩形をパラグラフに分類する．章題のみが最終行に記載されることはないので，本文に分類された基本矩形が一つしかない場合は，その基本矩形をパラグラフに分類する．

次に，本文領域ではパラグラフの区切りでインデントが発生することに着目するため，図7-7に示すように領域内の基本矩形に対しインデントレベル (idl) を定義する．具体的には，本文領域に分類された基本矩形を左から順 (左上端x座標の小さい順) にソートし，最左端の基本矩形の左上端x座標との差が一定値 (cw:文字幅の半分程度) を超えるまで, 基本矩形のインデントレベルを0とする．同様に, インデントレベル0の基本矩形を除く最左端の基本矩形からcwを超えるまでの基本矩形のインデントレベルを1とし，以下これらを繰り返して全ての基本矩形に対してインデントレベルを求める．

求めたインデントレベルを用いて，複数行にわたる章題やパラグラフを統合し, それぞれの領域を決定する. 具体的には, 本文領域の基本矩形を上から順 (左上端y座標の小さい順) に以下の条件をチェックし，何れかの条件を満たすところで領域を分割する．その他のところでは，基本矩形領域を統合し，章題あるいはパラグラフの領域を作成する．

- 基本矩形の分類が，章題からパラグラフへ，またはパラグラフから章題へかわるところ
- 連続する2つの基本矩形の分類がパラグラフの場合で，インデントレベルが1増加するところ

最後に，先頭行がパラグラフでインデントレベルが 0 である場合，最初のパラグラフは前の段組またはページからの続きのパラグラフであると判断し，当該パラグラフ領域にマークを施しておく．例えば，図 7-8 に示された本文領域に対して，以下のように章題とパラグラフが作成される．まず，2 行目と 3 行目の間隔がしきい値 h 以上であるために，2 行目までが章題領域，3 行目以降がパ

ラグラフ領域となる．さらに，5 行目から 6 行目にかけてインデントレベルが 1 増加しているため，ここで 2 つのパラグラフ領域に分割され，最終的に 1 つの章題領域と 2 つのパラグラフ領域が得られる．こうして求めた章題領域とパラグラフ領域を y 軸方向に読み順に並べ，章題とパラグラフの対応を記述しておく．

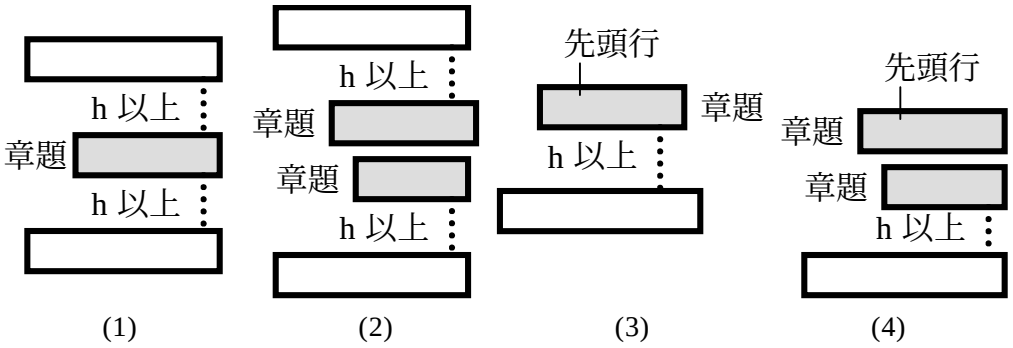


図 7-6　章題に分類される基本矩形

Fig. 7-6　Basic rectangles labeled in section title.

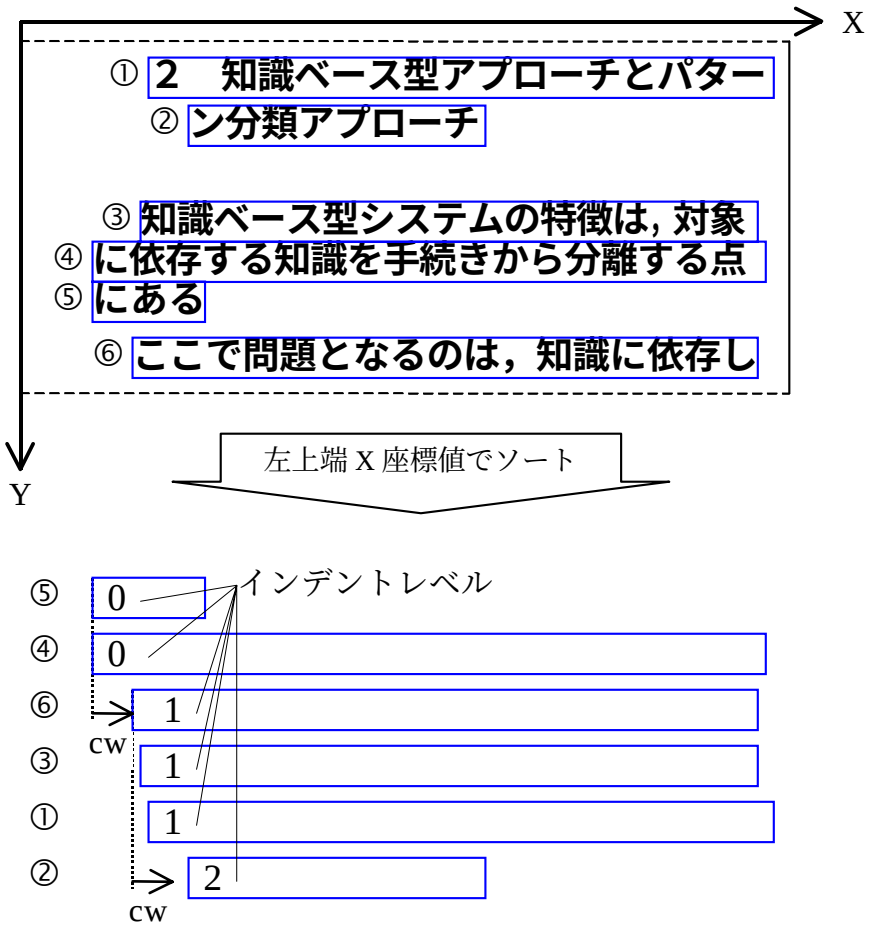


図 7-7　インデントレベルの求め方

Fig. 7-7　Example of the indent levels.

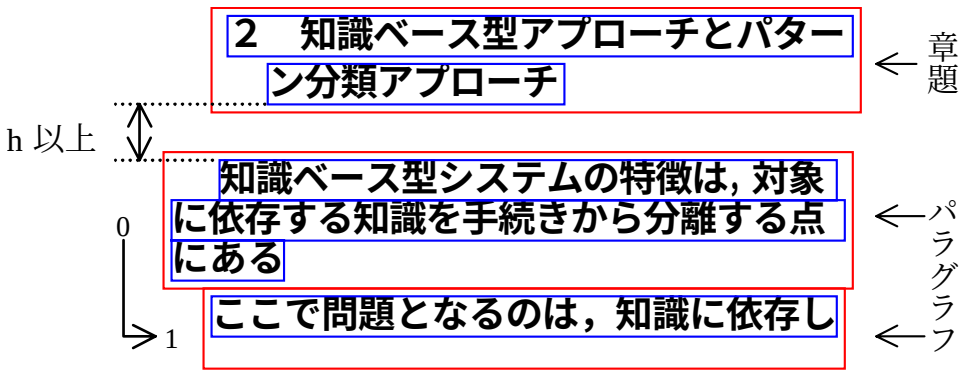


図 7-8　本文領域からの章題とパラグラフの抽出例

Fig. 7-8　Examples of section title and paragraphs in a body region.

7.3.2.3 参考文献グループ

参考文献に対応する部分領域に分類された基本矩形から，個々の参考文献領域を抽出する．本文領域と同様，図7-3のような2段組の場合，2つの参考文献グループ領域が抽出される．以下，それぞれの領域に対して個別に処理する．

まず，各参考文献の 2 行目以降がインデントされていることに着目する．具体的には, 本文領域と同様の方法で各基本矩形のインデントレベル (idl) を求める．次に, 参考文献領域に分類された基本矩形を上から順 (左上端 y 座標の小さい順) にソートし，基本矩形のインデントレベルが 1 から 0 へかわるところで領域を分割する．その他の場所では基本矩形領域を統合し，個々の参考文献領域を生成する．ここで，生成された個々の参考文献領域の先頭行のインデントレベルが 0 でない場合，この参考文献領域は前の段組またはページからの続きの領域であると判断し，マークを施しておく．

7.3.2.4 著者紹介グループ

部分領域作成過程で著者紹介グループに分類された基本矩形には，図7-3や図7-9(a) に示すように，著者名，著者写真，著者紹介といった細分化された領域名が定義されている．また，複数の著者紹介グループが存在する場合，図7-2に示した意味構造を得るため，個々の著者名，著者写真，著者紹介領域を抽出し，それぞれを個別に扱う必要がある．

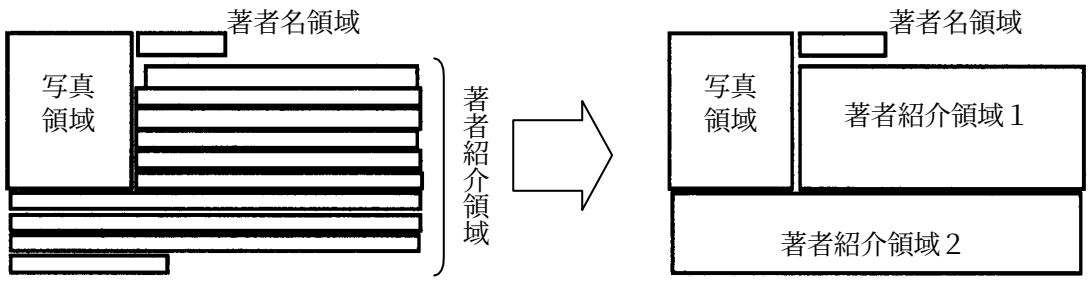
具体的には，まず個々の著者写真領域に着目し，写真領域の右側に隣接する著者名領域との対応づけを行う．次に，同じく写真領域の右側に隣接する著者紹介領域に分類される基本矩形を統合し, 写真領域と対応づける．これにより，図 7-9(b) の著者紹介領域 1 が抽出される．また，写真領域の下にも著者紹介に分類された基本矩形がある場合は，これらを著者紹介領域の続きであると見なし, 図 7-9(b) の著者紹介領域 2 のように統合し, 同じく写真領域と対応づける．このように，著者紹介領域を写真領域の右と下で分割して作成するのは，後に文字認識する領域を単純な矩形領域とするためである．

7.3.3 タグ付き文書の生成

これまでの処理で抽出された意味構成要素のうち，テキストとして扱う領域に対して文字切り出しおよび文字認識処理を施す．得られたテキストデータに対し，意味構成要素名と，付加情報として領域座標をSGML文書のタグに相当する論理タグとして付加する．この際，本文および参考文献グループの処理でマークされた段をまたがるパラグラフや参考文献領域のテキストをつなぎ合わせる．また，図，表，写真などの非テキスト領域に対しては，当該領域のページ画像を画像符号化し，個別ファイルとして蓄積するとともに，ファイル名で

参照できるようにする．

以上の処理により，SGML文書の素材となるテキストデータをページ単位に生成する．



(a)　　(b)

図 7-9　著者紹介グループの例

Fig. 7-9　Example of author's profile groups.

7.3.4 論文単位の SGML 文書の生成

ページごとに得られた前述のタグ付きデータをページ順に結合し，論文の意味構成要素全体を含むSGMLインスタンスを生成する．特に，本文や参考文献グループでマークされたページをまたがるパラグラフや参考文献領域のテキストデータをつなぎ合わせる．こうして生成されたSGMLインスタンスに，文字コード体系の宣言などを行うSGML宣言部[7-3)]やDTDを追加してSGML文書が完成する．

7.4　学術論文誌認識システムの評価実験

7.4.1 システムの構成

7.3で述べた手法を用い，学術論文誌認識システムを構築した．構築したシステムは，図7-10に示すように入力サーバと認識サーバから構成され，両サーバともUNIXワークステーションを基本としている．システムの概観を図7-11に示す．以下，各サーバの処理内容について述べる．

入力サーバでは，論文誌を高速スキャナを用いて両面同時にスキャンし，2値画像を得る．論文誌はページ順にスキャンされ，先に述べた前処理を施したのちページ画像をMMR圧縮し，画像データベースに蓄積する．

次に認識サーバでは，オペレータの介在により以下の処理を行う．

① まず，SGML文書を生成したい論文紙面を選択する．図7-12に示すようなペ

ージのサムネイル画像がスキャナで読み取られた順にディスプレイに表示され，オペレータがページ画像を選択する．この際，ページ画像がページ順に表示されていない場合は，オペレータが順番を修正する．また，基本矩形のパターン分類で用いる参照ベクトルが論文フロント・中間・最終ページ別に作成されるため，入力画像の論文フロントページと最終ページをオペレータが指定し，その間を中間ページとして処理する．

② 次に，選択されたページ画像が入力サーバの画像データベースから認識サーバに読み込まれ，7.3節で述べた処理を行う．途中，意味構成要素の抽出結果をオペレータに提示し確認・修正を行う．この際，図，表，数式など意味構成要素への対応づけに失敗した領域は，オペレータがマウスを用いて修正し，メニュー選択により要素名を指定する．

③ テキスト領域に対しては，直ちに文字認識処理が実行されるとともに，文字認識結果の確認を要する場合はオペレータが確認し，文字認識誤りを訂正する．

④ 最後に，所望論文の全ページを結合した SGML 文書をデータベースに出力する．

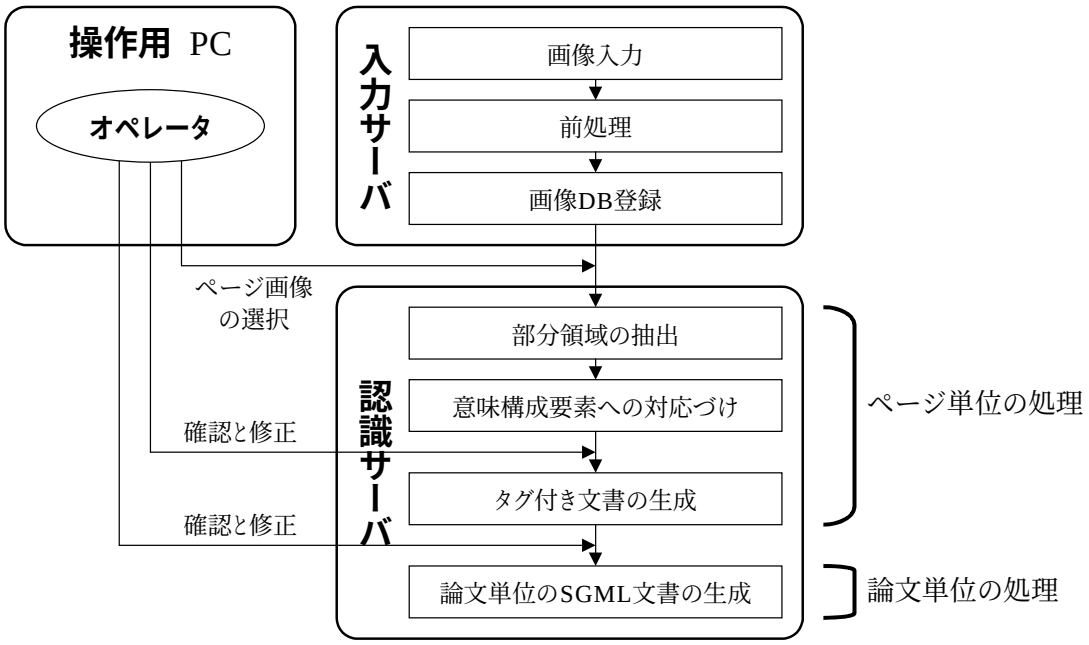


図 7-10　システム構成と処理の流れ

Fig. 7-10　System architecture and process flow.

図 7-11　学術論文誌認識システムの概観

Fig. 7-11　Developed system.

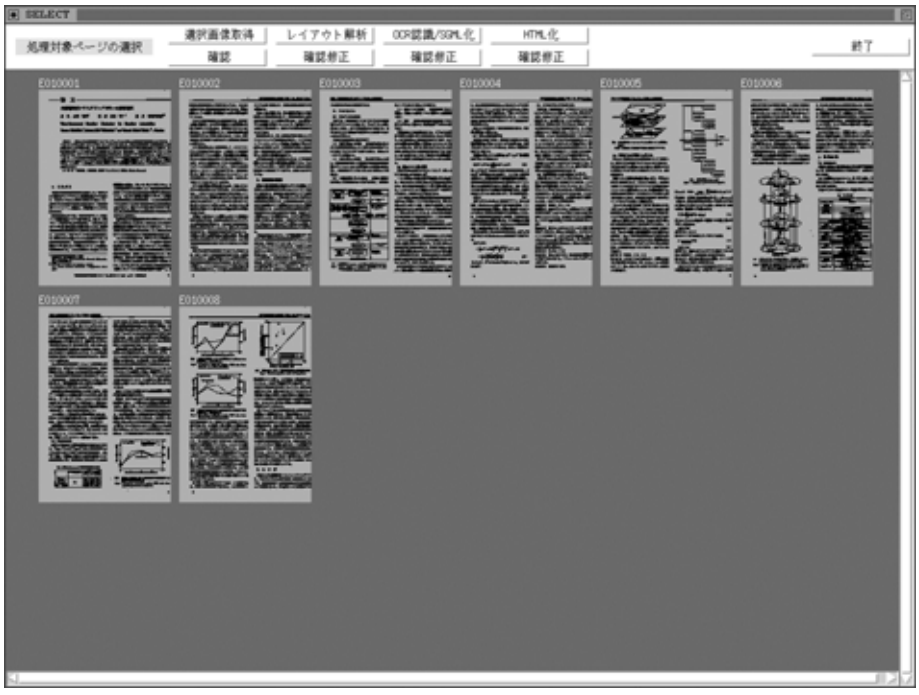


図 7-12　システムの操作画面の例

Fig. 7-12　Example of operation windows.

7.4.2 実験と考察

7.4.2.1 実験の内容

構築したシステムの処理能力を検証するため，実際の学術論文誌を用いて図7-2に示したDTDに基づくSGML文書を生成する実験を行った．

実験には，表7-3に示す電子情報通信学会論文誌に収録された338ページの画像を用いた．入力画像は解像度400[dots/inch]でスキャンし，前処理後の画像の大きさを水平方向X=2800[dot]，垂直方向Y=3800[dot]とした．また，前処理で印字領域の左上端を揃えるための基準点をSX=200，SY=150，インデントレベルを求める際のしきい値をcw=22[dot]，本文領域を章題とパラグラフに分類する際のしきい値をh=50[dot]とした．また，基本矩形の分類過程で必要となる参照ベクトルは，ページ種別ごとに全データを学習データとして用いて作成した．

評価項目として，意味構成要素抽出率，対応づけした要素の対応率，紙面通過率，文字認識率，および処理時間を取り上げた．ここで，意味構成要素抽出率は，表 7-2 に示した処理で得られる全領域のうち，個々の領域が正しく抽出された割合を，対応率は，正しく抽出された領域に対し全て正しく対応づけされた割合をそれぞれ表している．これらの指標によって，7.3 節で提案した部分領域の抽出と意味構成要素への対応づけの結果を評価する．また，紙面通過率は，図 7-10 に示す意味構成要素への対応づけ後，確認修正作業の不要なページの割合を示すものであり，文字認識率やシステムの処理時間と合わせて実際の作業量の目安となるものである．

表 7-3　実験に用いたデータ

Table 7-3　Number of pages used in experiment.

ページ種別	フロント	中 間	最 終	合 計
枚数	110	110	118	338

7.4.2.2 実験の結果

意味構成要素抽出率，および対応率を表 7-4 および表 7-5 にそれぞれ示す．表中の“ 個々の所属機関名 ”の領域数は，“ 現在の所属機関名 ”が別記される場合の領域数を含めてカウントした．表 7-6 には紙面通過率を示す．全実験データに対して 80.2%の紙面通過率を得た．文字認識率は，和文著者名，章題，パラグラフ，参考文献，最終ページの著者名および著者紹介に含まれる 11,021 文字に対して 274 文字の誤読が発生し，正読率は 97.5%であった．また，処理時間は入力サーバで 1 ページあたり平均約 19 秒，認識サーバではオペレータの作

業時間を除いて 1 ページあたり平均 24 秒であった.

表 7-4　意味構成要素抽出率

Table 7-4　Extraction rates of logical components.

ページ種別	構成要素名	総数	領域正解数	要素抽出率
フロント	個々の和文著者名	302	298	98.7%
	個々の所属機関名	174	174	100%
	章題	117	116	99.2%
	パラグラフ	502	464	92.4%
中間	章題	65	60	92.3%
	パラグラフ	1166	1081	92.7%
最終	章題	33	20	90.9%
	パラグラフ	272	264	97.1%
	個々の著者名	270	262	97.0%
	個々の著者紹介	398	370	93.0%
	個々の写真	270	267	98.9%
	個々の参考文献	1021	1010	98.9%

表 7-5　対応率

Table 7-5　Linking rates.

部分領域	構成要素名	領域数 対応率
和文著者グループ	個々の和文著者名	298
	→ 個々の所属機関名	98.7 %
本文	章題	215
	→ パラグラフ	100　%
著者紹介グループ	個々の著者名	262
	→ 個々の著者紹介, 写真	100　%

表 7-6　紙面通過率

Table 7-6　Page passing rates.

ページ種別	紙面総数	修正不要な紙面数	紙面通過率
フロント	110	100	90.9%
中 間	110	82	74.5%
最 終	118	89	75.4%
合 計	338	271	80.2%

7.4.2.3 考察

7.2で述べた学術論文誌の構造解析における3つの課題について，実験結果を基に考察を加える．

(a) レイアウトが変動するページ画像からの部分領域の抽出

論文フロントページのみならず，図，表，写真などが混在する中間ページや最終ページに対して，提案システムでは90%を超える部分領域抽出率を達成できることが分かった．また，図や表などを画像としてデータベースに格納し，本文中に画像へのリンクを埋め込むことにより，本文のテキストデータ以外の図表も参照することができる．

(b) 部分領域の意味構成要素への対応づけ

個々の著者名や著者所属機関名の対応づけなど，部分領域の細分化と対応記号の抽出により，今回実験に用いたDTDに対応した意味構成要素の抽出と要素間の対応づけを実現できることが分かった．

(c) 論文単位の構造化

連続する複数のページ画像から論文単位にSGML文書を生成できることが分かった．特に，段組やページをまたがる文章も，レイアウトに基づく特徴を用いて，次段あるいは次ページの文章と連結できることが分かった．

実用システムに求められる処理効率について考察を加える．表7-6に示したように，全紙面を対象とした紙面通過率は80%を超えており，オペレータの負担も少ないシステムを構築できることが分かった．実験結果の中で，オペレータによる修正が必要な領域として，図や表のキャプションをパラグラフと誤るケースが多く，図表領域の抽出の自動化が今後の課題である．

また，従来のようにオペレータがキーボード入力によりSGML文書を作成する場合と，提案システムを用いた場合の処理時間を比較する．提案システムの

処理時間は入力サーバと認識サーバをあわせて1ページ当たり約43秒であった．一方，オペレータがキー入力する場合，入力後の文字校正も含めてA4版で1ページ当たり約 15 分[7-17]かかるとされる．提案システムにおけるオペレータ作業時間を考慮しても，実験結果で得られたように領域生成率や文字認識率が高い場合，提案システムの方が優位であることは明らかである．

7.5 まとめ

学術論文誌のページ画像から意味構成要素のフルテキストを抽出し，学術論文のデータベース化を支援するシステムを構築した．実際の学術論文誌からのSGML文書生成では，意味構成要素の抽出結果が全て正しいページ数の割合を示す紙面通過率が80.2%となることが分かった．このことにより，SGML文書の生成時間の大幅な短縮が図れることを述べた．

最後に，今後の課題について言及する．まず，他の多くの学術論文誌を対象としたロバスト性の検証が必要である．今回実験に用いた学術論文誌での実験結果は上述のとおり良好であったが，レイアウトの異なる他学会の論文誌での検証が必要である．筆者らは，先のパターン分類手法に基づく文書画像の構造解析[16]で複数誌を対象とした部分領域の抽出を行っており，本論文で述べたSGML文書生成も可能であると考える．

また，本稿で作成したDTDで除外した意味構成要素の抽出も必要となる．本稿で取り上げたシステムにおいては，書誌情報として扱われる要素や本文等検索時に利用価値の高い要素に限定したDTDを想定した．このDTDには含まれていない脚注や付録，また，今回の実験でオペレータ介在作業の多かった図表のキャプションや数式の抽出を自動化することも必要である．これらについては，提案システムで実装されている本文領域の抽出処理の後，文字認識結果を活用して，高精度の意味構成要素の抽出を実現することが考えられる．

第 8 章　結論

8.1　本研究で得られた成果

本研究で得られた成果を要約し，残された課題について述べる．

本研究の目的は，これまで紙を媒体として扱われてきた文書をコンピュータで処理できる電子文書に変換することにある．文書本来の情報伝達・保存機能に立ち返ってみると，文書は記録すべく意味構造と，それを人間にとって視覚的に分かり易く表現するレイアウト構造から成り立っていることが分かる．紙を媒体とする文書のコンピュータでの自動読み取りに際して，これらの性質に着目し，それぞれの構造を抽出することが必要である．

本研究論文では，文字・図形が混在する文書画像のレイアウト構造，ならびに意味構造の認識に関する新しい手法について論じた．また，その応用システムの構成法について述べるとともに，実際に実用化システムを構築し，評価を行った結果について述べた．以下，各章で得られた成果について述べる．

第 1 章「序論」では，本研究の目的と，本研究論文の構成について述べた．

第 2 章「文書画像の構造認識」では，文書画像の構造認識に関する基本的枠組みとして，文書のレイアウト構造および意味構造について述べた．また，本研究で対象とする文書の範囲とその性質，ならびに構造認識に用いる文書画像の特徴に関する基本的事柄について述べた．

第 3 章「文字・図形の分離」では，文字・図形が混在した文書の図表領域を対象として，文字と図形を分離抽出する新しい手法について論じた．オフィスで扱われる文書・図面は多種多様な書式を有しており，また文字と図形が混在していることが多い．これらの文書・図面を認識入力するためには，文字・図形が混在した領域の文字と図形を分離抽出する必要がある．本章では，着目する点 (画素) の近傍の線密度を近接線密度として定義し，近接線密度の値によって文字と図形を分離抽出する手法について述べた．文字・図形が混在した領域の近接線密度を調べると，文字領域上に近接線密度の高いピークが現れ，このピークの高さは文字サイズに関連している．このことにより，近接線密度は文字・図形の分離抽出に有効な特徴であることを示した．また，提案手法を用いた実験を行い，抽出された文字・図形に対して文字認識，図形認識を行った結果，良好な認識結果が得られたことを示した．

第 4 章「文書画像のレイアウト構造認識」では，文書をイメージ情報として入力し，文字・図形を分離抽出してそれぞれを文字認識，図形認識することにより，コンピュータ処理に適したコード情報に変換する文書画像の認識手法に

ついて論じた．提案手法では，文字・図形の認識に先立って文書のレイアウト構造を抽出し，文字・図形の分離抽出や一文字切り出しを行うことにより，多種多様な文書を認識可能としている．特に，オフィス文書の多様性に対処するため，文書の本質的特徴である周辺分布特徴，黒連結特徴，近接線密度特徴を併用する手法を提案した．次に，オフィス内の印刷文書を用いた認識実験により，テキスト領域，表領域，図領域が正しく抽出されること，およびテキスト領域は文字コード列に，表領域は文字コード列とフィールド情報に，図領域は文字コード列と線分コマンドに変換できることを示した．

第 5 章「レイアウト構造認識に基づく文書認識システムの構成法」では，イメージ形式の文書をコンピュータ処理に適したコード情報に変換する文書認識システムについて論じた．まず，文書認識処理の実現に必要なレイアウト解析，文字認識，図形認識，イメージ圧縮の各処理技術について，既存アルゴリズムの適用性を吟味し，文書認識の実現性について述べた．次に，これらの各処理を組み合わせて文書認識処理を実行する文書認識モデルを提示し，モデルに基づき文書認識システムの性能評価を試みた．評価実験の結果，文書ファイル量の削減，文書入力の高速化・省力化が図れることを示した．さらに，認識結果からシステム構成に向けて幾つかの指針を導出し，LAN を用いたシステムの実現例を示した．

第 6 章「文書画像の意味構造認識」では，文書画像から電子的な文書へのメディア変換に際し，文書の意味構造の認識手法について論じた．これまでに，意味構成要素の位置関係などをルール化し，これを対象文書に関する知識として用いるモデルベース型の意味構造認識手法が提案されている．しかしながら，こうした従来の手法では，意味構造モデルとして構成要素領域の位置関係を記述する必要があり，構成要素の相対位置が変動する文書画像に対しては適用できないなどの問題があった．そこで，構成要素の位置，行間隔，文字の大きさ，文字数などの特徴を画像特徴としてとらえ，これらを総合的に判断して文書画像の意味構造を認識するパターン分類に基づく手法を提案した．提案手法は，文字列領域に相当する基本矩形領域に対して特徴ベクトルを算出し，あらかじめ作成した参照ベクトルと比較して最も特徴の類似した構成要素に対応づけるものである．利点としては，構成要素の位置関係が定まらない領域や矩形表現できない領域でも構造認識が可能となること，構成要素の確からしさを表現できること，基本矩形単位に並列処理が可能となることが挙げられる．学術論文誌の論文フロントページを対象とした実験により，98．5%の基本矩形分類率が得られることを示し，提案手法の有効性について述べた．

第 7 章「学術論文誌認識システムの実用化」では，学術論文誌に含まれる連

続した複数のページからなる個々の論文の意味構造を認識し，データベースに登録するためのシステムを構築した結果について論じた．従来から，印刷文書の意味構造認識に関しては，レイアウト構造と意味構造の対応関係に着目した手法が提案されている．しかし，レイアウトが比較的固定的な論文フロントページの認識では良好な結果が得られているが，継続するページについてはレイアウトが変動するため，論文単位の意味構造認識結果を得るのは困難であった．そこで，第 6 章で論じたパターン分類手法を応用した論文単位の意味構造認識システムを提案した．意味構造認識結果の記述形式として SGML (Standard Generalized Markup Language) を用い，現在発行されている学術論文誌を用いて実際に SGML 文書を生成するシステムを構築し，その有効性を評価した結果について述べた．

8.2 今後の研究課題

本研究論文の締めくくりにあたり，今後解決すべき課題について述べる．

一頃，日本人のコンピュータ利用の障害としてキーボード文化に馴染まないと言われたことがあるが，今日のワードプロセッサ，あるいはパーソナルコンピュータの普及は必ずしもそうではなかったことを示している．ところが，文書の作成，保存が電子化されても，依然として紙を媒体とした情報の伝達や保存は減らない．むしろ紙文書の増大を招くこととなった．一方で，電子文書が増大するにつれて，紙と比べて電子文書の有用性も認知され，大量に存在する遡及文書を電子化したいと言う要求も高まってきた．

第 1 章で述べたように，究極は紙文書と電子文書がコンピュータを介してシームレスに扱うことができるようになることである．

現在，電子図書館の構築が進められている．ここでの電子図書館は，図書館サービスの一つである書誌情報 (二次情報) 検索を，コンピュータネットワークを介して遠隔地から利用できるようにするのみならず, 図書の全ての内容 (一次情報) を電子化し, コンピュータネットワークを介して提供できるようにすることである．これは，いわゆる図書館の機能を超えたデジタルライブラリの構築とも言えるものである．なぜならば，デジタル化された情報は，今後長い歴史の中で保存され，また自在に編集加工も可能であるからである．こうしたデジタルライブラリの主要なコンテンツとして，本研究で扱った学術論文は非常に有用であり，おおいに期待が押せられるところである．この意味で，本研究で成し得た文書画像の構造解析技術は極めて有効なものではあるが，一方多種多様な出版物等の文書を今後全て扱うことを考えると，本技術も未だ制約の多い技術であると言える．これらのことから，本研究論文で論じた技術は，まだ完成されたものではなく，今後さらに発展させるべき技術である．

本研究の基本的な前提の一つに，レイアウト構造と意味構造の関連性に着目したことがあげられる．一般的には，文書に記すべき情報の意味構造を読者に分かり易く表現するために，表示・印刷等においてレイアウト構造がそれぞれ割り当てられることが多い．このことにより，画像として捉えやすいレイアウト構造から意味構造を引き出すことが可能であることを示した．

一方，近年のインターネットに代表される情報流通環境では，むしろオンラインドキュメントと称される，ページ出力を意図しない文書が増えつつある．こうした文書は，作成段階から電子化されてはいるものの，レイアウト構造を特定せず作成されたものが多い．これらの文書から，本論文でも述べた SGML 等の意味構造を持った文書に変換するには，テキスト情報に対する言語解析が不可欠である．これらの技術に関しては，近年の自然言語処理技術の進歩が目覚ましく，今後これらの技術に負うところが大きいと考えられる．

謝　辞

本論文をまとめるにあたり，数々のご指導とご教示を賜った東京工業大学大学院 総合理工学研究科電子システム専攻 河原田 弘教授に心から感謝致します．河原田 弘教授には，筆者の東京工業大学大学院 総合理工学研究科 博士課程前期課程の研究でご指導頂いて以来，本研究に対して激励を頂いた．また，本論文の細部にわたり適切なご批判とご指導を賜った東京工業大学 大学院情報理工学研究科 小川 英光教授，同 精密工学研究所 上羽 貞行教授，佐藤 誠教授，同 工学部附属像情報工学研究施設 長橋 宏教授，長尾 智晴助教授に深く感謝致します．

本論文は，筆者が日本電信電話公社 横須賀電気通信研究所 (昭和 55 年 4 月～昭和 60 年 3 月まで)，日本電信電話株式会社 横須賀電気通信研究所，同複合通信研究所 (昭和 62 年 9 月まで)，および，NTT データ通信株式会社 (現 株式会社 NTT データ) 開発本部，同技術開発本部 情報科学研究所において行った研究成果をまとめたものである．

この間，本研究の開始当初からご指導頂いた株式会社 NTT データ 取締役 技術開発本部長 荒川弘熙博士に深く感謝致します．荒川弘熙博士には本研究論文をまとめることを強く動機づけて頂いた．また，本研究の遂行にあたり上司としてご指導頂いた三菱電機 川田圭一氏 (当時 NTT 複合通信研究所 分散処理装置研究室長)，株式会社 NTT データ 技術開発本部 情報科学研究所歴代所長の安部孝二氏 (現 日本アドバンストカードシステム株式会社)，管村 昇博士 (現 人材開発部長)，現所長の中村太一博士に深く感謝致します．

さらに，アライド・リソーシズ・コミュニケーションズ代表取締役 久保田一成氏には，NTT 横須賀電気通信研究所当時，文字・図形の画像特徴に関する数々のご討論とご助言を頂いた．株式会社 NTT データ 技術開発本部担当部長 木田博巳氏には，文書情報の記述方式に関するご討論，ならびに文書認識システムの性能評価に関するご助言とご支援頂いた．同 新世代情報サービス事業本部担当部長 遠城秀和氏には，文書認識システムの設計・試作に際してご支援を頂いた．高知工科大学教授 岡田 守博士には，NTT データ通信株式会社当時，学術論文誌からの SGML 文書生成システムの実用化研究に際し，数々のご助言を頂いた．ここに記して各位に深く感謝致します．

また，株式会社 NTT データ 技術開発本部 山岡正輝氏，ならびに岩根和巳氏，同 公共システム事業本部 佐藤道弘氏には，それぞれ本研究を進めるにあたり，数々の実験システムの構築において力添え頂いた．あわせて各位に深く感謝致します．

最後に，本研究論文を執筆する機会を与えて頂くと共に，暖かいご支援を頂いた株式会社 NTT データ 藤田史郎代表取締役会長, 神林留雄代表取締役社長, 田中義昭前常勤監査役 (当時 取締役開発本部長)，NTT ソフトウェア株式会社 代表取締役社長 鶴保征城博士 (前 株式会社 NTT データ 常務取締役技術開発本部長) に心からお礼申し上げます.

参考文献

【第 1 章】

1-1) 徳升,岩城, “フィールド情報に基づく帳票識別の一検討,” 信学技報, PRU88-114 (1989).

1-2) 長谷, 米田, 酒井, 吉田: “図書目録カードの認識・理解の試み”, 信学技報, PRU86-42 (1986).

1-3) 黄瀬, 山田, 田中, 馬場口, 手塚: “名刺画像認識における項目仮説生成”, 情処研報, CV52-12 (1988).

1-4) 辻本, 麻田: “文書画像理解による記事の自動抽出”, 情処研報, CV52-13 (1988).

1-5) 東野, 藤澤, 中野, 江尻: “書式定義言語を用いた文書画像の理解”, 画電学誌, Vol.17, No.5, pp.267-277 (1988).

【第 2 章】

2-1) CCITT 勧告 T.410 シリーズ/ISO8613: “Open Document Architecture (ODA) and Interchange Format” (1988).

2-2) “ISO8879, Standard Generalized Markup Language” (1986).

2-3) 山本, 服部, 中尾: “幾何学モード変換による FAX の線密度変換”, 信学技報, IE75-71 (1975).

2-4) 午坊, 相原: “ファクシミリ線密度変換の一方法”, 画像電子学会全大予稿, 10 (1975).

2-5) 水野, 臼渕, 飯沼, 石黒: “細め処理を用いた画像の縮小”, 昭 54 信学全大, 5-80 (1979).

2-6) 宮井, 首藤: “イメージの拡大縮小方式”, 情処学会第 20 回全大, 2E-6 (1979).

2-7) 吹抜, 吉本: “ランレングス領域における画素密度変換”, 信学技報, IE79-60 (1979).

2-8) 久保田: “文字・図形の線密度変換の一検討”, 昭 57 信学全大, 1173 (1982).

2-9) 例えば 長尾監訳 A. Rosenfeld: “Digital Picture Processing”,近代科学社 (1978).

2-10) 大津: “判別および最小 2 乗基準に基づく自動しきい値選定法”, 信学論(D), Vol.J63-D, No.4, pp.349-356 (1980).

2-11) 樋野, 東田, 大田, 坂井: “尾根点・谷点に基づく線図形の二値化方式”, 情処学会第 24 回全大, 4E-1 (1982).

2-12) K. Y. Wong, R. G. Casey, F. M. Wohl: “Document Analysis System”, IBM Journal of Research and Development, Vol.26, No.6, pp.647 (1982).
2-13) 村尾, 坂井: “文書画像における構造情報の抽出”, 情処学会第 21 回全大, 7H-1 (1980).
2-14) 大谷: “ボカシ法による文字列認識の一検討”, 情処学会第 22 回全大, 4I-4 (1981).
2-15) 浅田, 沖野, 荒木, 辻, 小谷: “文字列抽出法の検討”, 昭 56 信学会情報システム部門全大, 209 (1981).
2-16) 庭田, 中村, 南: “欧文テキスト画像を対象とした文字領域の抽出”, 情処学会第 26 回全大, 2B-3 (1983).
2-17) 秋山, 増田: “書式指定情報によらない紙面構成要素抽出法”, 信学論(D), Vol.J66-D, No.1, pp.111-118 (1983).
2-18) 長谷: “2 次元フーリエ変換を用いた文字列抽出法の検討”, 情処学会第 22 回全大, 4I-2 (1981).
2-19) 伊藤, 坂谷, 高井, 鷹尾, 飯坂: “フリーフォーマット文書の並列フィールドセグメンテーション手法”, 情処学会第 20 回全大, 2E-1 (1979).
2-20) 寺嶋, 首藤, 川井, 渡辺: “図面に書かれた文字の切り出し方式”, 情処学会第 23 回全大, 6C-4 (1981).
2-21) 井上, 吉田: “文字・図形分離方式の検討”, 昭 56 信学総合全大, 1344 (1981).
2-22) 金子, 三ツ矢, 奥平: “図面における線要素と文字・記号の分離”, 昭 57 信学会通信部門全大, S6-5 (1982).
2-23) 久保田, 荒川: “線順次アルゴリズムを用いた論理回路図面入力法”, 信学技報, PRL82-17 (1982).

【第 3 章】

3-1) 秋山, 増田: “書式指定情報によらない紙面構成要素抽出法”, 信学論(D), Vol.J66-D, No.1, pp.111-118 (1983).
3-2) 村尾, 坂井: “文書画像における構造情報の抽出”, 情処学会第 21 回全大, 7H-1 (1980).
3-3) Ito, Sakatani and Takai: “Parallel Segmentation Algorithm of Free Format Document”, ICCS, pp.519-523 (1979).
3-4) 西村, 野口, 豊田: “新関記事の本文を構成する文字の切出し”, 情処学会第 24 回全大, 3E-7 (1982).
3-5) 目黒, 梅田: “マルチフォント印刷漢字の認識”, 信学論(D), Vol.J65-D,

No.8, pp.1026-1033 (1982).
3-6) 久保田, 荒川: “線順次アルゴリズムを用いた図面入カ方式の一検討”, 昭57 信学通信部門全大, S6-1 (1982).

【第 4 章】

4-1) 木田, 岩城, 久保田: “LAN における文書理解サービスとメディア処理ステーションの構成”, 情報処理学会「LAN/マルチメディアの応用と分散処理」シンポジウム, pp.133-140 (1984).
4-2) 木田, 岩城, 荒川: “文書自動認識システムの構成法”, 画像電子学会誌 ,15, 2, pp.107-115 (1986).
4-3) Inagaki, Kato, Hiroshima and Sakai: “MACSYM: A Hierarchical Parallel Image Processing System for Event-Driven Pattern Understanding of Documents”, Pattern Recognition, Vol.17, No.1, p.85 (1984).
4-4) 恒川, 下辻: “図面読取装置 TOSGRAPH”, 信学論(D), Vol.J68-D, No.4, pp.466-472 (1985).
4-5) 長田, 井上, 吉田: “論理回路図面の自動入力処理”, 信学論(D), Vol.J68-D, No.4, pp.837-844 (1985).
4-6) 中野, 藤澤, 国崎, 岡田, 花野井: “文字認識と協調した表形式文書の理解”, 信学論(D), Vol.J69-D, No.3, pp.400-409 (1986).
4-7) 秋山, 増田: “書式指定情報によらない紙面構成要素抽出法”, 信学論(D), Vol.J66-D, No.1, pp.111-118 (1983).
4-8) 村尾, 坂井: “文書画像における構造情報の抽出”, 情処学会第 21 回全大, 7H-1 (1980).
4-9) Ito, Sakatani and Takai: “Parallel Segmentation Algorithm of Free Format Document”, ICCS, pp.519-523 (1979).
4-10) Toyoda, Noguchi and Nishimura: “Study of Extracting Japanese Newspaper Article”, 6th-ICPR, pp.1113-1115 (1982).
4-11) 馬場口, 塚本, 相原: “手書き日本文字列からの文字切り出しの基礎的考察”, 信学論(D), Vol.J68-D, No.12, pp.2123-2131 (1985).
4-12) 岩城, 久保田, 荒川: “近接線密度法による文字・図形分離抽出”, 信学論(D), Vol.J68-D, No.4, pp.821-828 (1985).
4-13) 久保田: “文字・図形の線密度変換の一検討”, 昭 57 信学全大, p.1173 (1983).
4-14) Akamatsu and Kawatani: “Hierarchical Classification of Hand-printed Kanji Using Density Feature and Configuration

Feature”, ICTP’83, pp.175-180 (1983).
4-15) 久保田, 荒川: “線順次アルゴリズムを用いた図面入力方式の一検討”, 昭57 信学会通信部門全大, S6-1 (1982).
4-16) 遠城, 岩城, 木田: “文書認識のためのプロダグションシステムの構成”, 信学技報, PRL84-67 (1985).
4-17) 駱, 渡邉, 杉江: “ルールベースの適用による日本語新聞紙紙面の構造認識”, 信学論(D-II), Vol.J75-D-II, No.9, pp.1514-1525 (1992).

【第 5 章】

5-1) OA 実態調査報告書, 日本 OA 協会 (1983).
5-2) 大町, 臼渕: “文書のイメージ編集”, 画像電子学会誌, Vol.13, No.2, pp.129-139 (1984).
5-3) 木村, 伊藤, 富永: “FAX 入力による文書画像編集システム”, 昭 59 信学全大, No.1412, pp.5-193 (1984).
5-4) 木田, 岩城, 久保田: “LAN における文書理解サービスとメディア処理ステーションの構成”, 情報処理学会「LAN/マルチメディアの応用と分散処理」シンポジウム, pp.133-140 (1984).
5-5) 秋山, 増田: “書式指定情報によらない紙面要素抽出法”, 信学論(D), Vol.J66-D, No.1, pp.111-118 (1983).
5-6) 西村, 野口, 豊田: “新聞記事の本文を構成する文字の切出し”, 情処学会第24 回全大, 3E-7 (1982).
5-7) Akamatsu and Kawatani: “Hierarchical Classification of Hand-printed Kanji Using Density Feature and Configuration Feature”, ICTP’83, pp.175-180 (1983).
5-8) 目黒, 梅田: “マルチフォント印刷漢字の認識”, 信学論(D), Vol.J65-D, No.8, pp.1026-1033 (1982).
5-9) 岩城, 久保田, 荒川: “近接線密度法による文字・図形分離抽出”, 信学論(D), Vol.J68-D, No.4, pp.821-828 (1985).
5-10) 岩城, 荒川: “図表領域中の文字の認識”, 信学技報, PRL84-50, pp.31-36 (1984).
5-11) 長田, 井上, 吉田: “論理回路図の自動入力処理”, 信学論(D), Vol.J68-D, No.4, pp.837-844 (1985).
5-12) 村瀬, 若原, 梅田: “候補ラティス法による手書きフローチャートのオンライン認識”, 信学論(D), Vol.J66-D, No.6, pp.675-682 (1983).
5-13) 久保田, 荒川: “線順次アルゴリズムを用いた論理回路図面の入力法”, 信学技報, PRL82-17, pp.17-24 (1982).

5-14) Facsimile Coding Schemes and Coding Control Functions for Group 4 Facsimile Apparatus, CCITT Recommendation T.6.
5-15) W. Horak, G. Kronert: “Techniques for Preparing and Interchanging Mixed Text-Image Documents at Multifunctional Workstations”, Siemens Forsch. -u. Entwickl-Ber, Bd.12, Nr.1, pp.61-69 (1983).
5-16) 白鳥: “かな漢字変換入力方式における入力作業特性の検討”, 昭 57 信学全大, No.2460, p.6-244 (1982).
5-17) 遠城, 岩城, 木田: “文書認識のためのプロダクションシステムの構成”, 信学技報, PRL84-67, pp.13-18 (1985).
5-18) 駱, 渡邉, 杉江: “ルールベースの適用による日本語新聞紙紙面の構造認識”, 信学論(D-II), Vol.J75-D-II, No.9, pp.1514-1525 (1992).

【第 6 章】

6-1) 駱, 渡辺, 吉田, 稲垣, 斎藤: “知識ベースに基づいた図書目録カードの理解”, 情処学論, Vol.31, No.12, pp.1755-1767 (1990).
6-2) K. Kise, K. Momota, M. Yamaoka, J. Sugiyama, N. Babaguchi and Y. Tezuka: “Model based Understanding of Document Images”, Proc. of MVA'90, pp.471-474 (1990).
6-3) A. Yamashita, T. Amano, H. Takahashi and K. Toyokawa: “A Model Based Layout Understanding Method for the Document Recognition System”, Proc. of 1st ICDAR, pp130-138 (1991).
6-4) K. Kise, M. Yamaoka, N. Babaguchi and Y. Tezuka: “Model Based System for Analyzing Document Images”, Proc. of 11th ICPR, pp.647-650 (1992).
6-5) 山岡, 岩根, 岩城: “パターン分類手法に基づくレイアウト解析”, 信学技報, PRU93-125 (1994).
6-6) K. Iwane, M. Yamaoka and O. Iwaki: “A Functional Classification Approach to Layout Analysis of Document Images”, Proc. of 2nd ICDAR, pp.778-781 (1993).
6-7) T. Akiyama and N. Hagita: “Automated Entry System for Printed Documents”, Pattern Recognition, 23, 11, pp.1141-1154 (1990).
6-8) 岩城, 木田, 荒川: “文書認識アルゴリズムの一検討”, 画電学誌, Vol.15, No.4, pp.254-264 (1986).
6-9) ラッヘンブルック著, 鈴木義一郎, 三宅章彦訳: “判別分析”, 現代数学社.
6-10) K. Fukunaga: “Introduction to Statistical Pattern Recognition”, Academic Press (1972).

6-11) 石塚: “SGML 形式による学会誌全文データベースの構築と印刷”, 情報知識学会誌, Vol.2, No.1, pp.23-48 (1991).
6-12) 黄瀬, 山岡, 馬場口, 手塚: “文書画像構造解析のための知識ベースの一構成法”, 情処学論, Vol.34, No.1, pp.75-87 (1993).

【第 7 章】

7-1) “JOIS 活用の手引き”, 日本科学技術情報センター (1991).
7-2) 根岸: “フルテキスト・データベースの応用動向”, 情処学論, Vol.33, No.4, pp.413-420 (1992).
7-3) “ISO8879, Standard Generalized Markup Language” (1986).
7-4) 田中: “文書記述言語 SGML とその動向”, 情処学論, Vol.32, No.10, pp.1118-1125 (1991).
7-5) Bruce Schatz and Hsinchum Chen: “Building Large-Scale Digital Libraries”, Computer, Vol.29, No.5, pp.22-26 (1996).
7-6) 梶, 藤澤: “電子図書館システムの技術動向”, 信学誌, Vol.79, No.9, pp.910-919 (1996).
7-7) J. Adachi, and H. Hashizume: “NACSIS electronic library system : Its design and implementation”, Proceedings of The International Symposium on Digital Libraries 1995,pp.36-41 (1995).
7-8) Guy A. Story, Lawrence O'Gorman, David Fox, Louise Levy Schaper and H.V. Jagadish: “The RightPages Image-Based Electronic Library for Alerting and Browsing”, Computer, Vol.25, No.9, pp.17-26 (1992).
7-9) 石塚: “SGML 形式における学会誌全文データベースの構築と印刷”, 情報知識学会誌, Vol.2, No.1, pp.23-48 (1991).
7-10) 森田, 鈴木, 宮川, 浜中: “SGML 方式による情報管理誌全文データベース化の可能性と HTML による電子版情報管理誌の試作”, 情処研報情報学基礎, Vol.95, No.45, pp.7-14 (1995).
7-11) 駱, 渡邉, 杉江: “ルールベースの適用による日本語新聞紙紙面の構造認識”, 信学論(D-II), Vol.J75-D-II, No.9, pp.1514-1525 (1992).
7-12) 山田: “文書画像の ODA 論理構造化文書への変換方式”, 信学論(D-II), Vol.J76-D-II, No.11, pp. 2274-2284 (1993).
7-13) S. Tsujimoto and H. Asada: “Understanding Multi-articled Documents”, Proceedings of the 10th ICPR, pp.551-556 (1990).
7-14) A. Yamashita, T. Amano, H. Takahashi and K. Toyokawa: “A Model Based Layout Understanding Method for the Document Recognition System”, Proceedings of the first ICDAR, pp.130-138 (1991).

7-15) K. Kise, M. Yamaoka, N. Babaguchi, and Y. Tezuka: “Model Based System for Analyzing Document Images”, Proceedings of the 11th ICPR, pp.647-650 (1992).
7-16) 山岡, 岩根, 岩城: “パターン分類手法に基づく文書画像の構造解析”, 信学論(D-II), Vol.J79-D-II, No.5, pp.756-764 (1996).
7-17) 笠原: “SGML 適用による電子化の実践”, 情処研報デジタルドキュメント, Vol.96, No.76, pp.1-8 (1996).

付　録

学術論文誌認識システムで用いたSGML-DTD

```
<!DOCTYPE article [
<!-- GENERAL ENTITIES -->
<!ENTITY        gt      ">"     -- Greater than -->
<!ENTITY        lt      "<"     -- Less than -->
<!ENTITY        amp     "&"     -- Ampersand -->

<!-- PARAMETER ENTITIES -->
<!ENTITY %data"(#PCDATA|sup)*"          -- 解析対象文字列と上付き文字 -->

<!ELEMENT article       - - (fm,bdy,bm) +(pgno)     -- 論文 -->
<!ATTLIST article
        vol     NMTOKEN     #IMPLIED        -- Vol --
        num     NUTOKEN     #IMPLIED        -- Number -- >
<!-- 前付け -->
<!ELEMENT fm            - O (tg,aug,abs)            -- 前付け -->
<!ELEMENT tg            - O (jt,et)                 -- タイトルグループ -->
<!ELEMENT jt            - O (%data;)                -- 和文タイトル -->
<!ELEMENT et            - O (%data;)                -- 英文タイトル -->
<!ELEMENT aug           - O (jaug,eaug)             -- 著者グループ -->
<!ELEMENT jaug          - O (jauff+)                -- 和文著者グループ -->
<!ELEMENT jauff         - O (jau,aff,paff?)         -- 個々の著者 -->
<!ELEMENT jau           - O (%data;)                -- 和文著者名 -->
<!ELEMENT aff           - O (%data;)                -- 所属機関名 -->
<!ELEMENT paff          - O (%data;)                -- 現在の所属機関名 -->
<!ELEMENT eaug          - O (%data;)                -- 英文著者名 -->
<!ELEMENT abs           - O (%data;)                -- アブストラクト -->
<!-- 本文 -->
<!ELEMENT bdy           - O (sec+) +(floats)        -- 本文 -->
<!ELEMENT sec           - O (st,p+)                 -- 章 -->
<!ELEMENT st            - O (%data;)                -- 章題 -->
<!ELEMENT p             - O (%data;)                -- パラグラフ -->
<!ELEMENT floats        - O EMPTY                   -- 図表 -->
<!ATTLIST floats
```

(1)

```
        file        CDATA           #REQUIRED     -- ファイル名 -->
<!-- 後付け -->
<!ELEMENT bm          - O (bibg,jaupg)          -- 後付け -->
<!ELEMENT bibg        - O (bib+)                -- 参考文献グループ -->
<!ELEMENT bib         - O (%data;)              -- 参考文献 -->
<!ELEMENT jaupg       - O (jaup+)               -- 著者紹介グループ -->
<!ELEMENT jaup        - O (name,prf,photo)      -- 個々の著者の紹介 -->
<!ELEMENT name        - O (%data;)              -- 著者名 -->
<!ELEMENT prf         - O (%data;)              -- 著者紹介 -->
<!ELEMENT photo       - O EMPTY                 -- 写真 -->
<!ATTLIST photo
        file        CDATA                   #REQUIRED     -- ファイル名 -->
<!-- ページ番号 -->
<!ELEMENT pgno        - - (%data;)              -- ページ番号 -->
<!ATTLIST pgno
        file        CDATA           #REQUIRED     -- ページ画像のファイル名 -->
<!-- フォントの指定 -->
<!ELEMENT sup         - - (%data;)              -- 上付き文字 -->
]>
```

発表文献

【投稿論文】

1. 河原田,横沢,岩城, “階層的パターン構造に基づく文字連想システム,” 信学論, Vol.J65-D, No.3, pp.362-369 (1982).
2. 岩城,久保田,荒川, “近接線密度法による文字・図形分離抽出,” 信学論, Vol.J68-D, No.4, pp.821-828 (1985).
3. 木田,岩城,荒川, “文書自動認識システムの構成法,” 画電学誌, Vol.15, No.2, pp.107-115 (1986).
4. 岩城,木田,荒川, “文書認識アルゴリズムの一検討,” 画電学誌, Vol.15, No.4, pp.254-264 (1986).
5. 山岡,岩根,岩城, “パターン分類手法に基づく文書画像の構造解析,” 信学論, Vol.J79-D-II, No.5, pp.756-764 (1996).
6. 山岡,佐藤$_{道弘}$,岩根,佐藤$_{弘行}$,岩城,岡田, “文書レイアウトに着目した学術論文誌からのSGML文書生成システム,” 画電学誌, Vol.27, No.5, pp.591-601 (1998).

【講演発表】

[国際会議]

1. K.Kubota, O.Iwaki, H.Arakawa, “Image Segmentation Techniques for Document Processing,” Proc. 1983 Int. Conf. on Text Processing with a Large Character Set (ICTP'83), pp.73-78 (1983).
2. K.Kubota, O.Iwaki, H.Arakawa, “Document Understanding System,” Proc. IEEE Seventh Int. Conf. on Pattern Recognition (7th ICPR), pp.612-614 (1984).
3. H.Kida, O.Iwaki, K.Kawada, “Document Recognition System for Office Automation,” Proc. IEEE Eighth Int. Conf. on Pattern Recognition (8th ICPR), pp.446-448 (1986).
4. O.Iwaki, H.Kida, H.Arakawa, “A Segmentation Method based on Office Document Hierarchical Structure,” Proc. 1987 IEEE Int. Conf. on Systems, Man, and Cybernetics, pp.759-763 (1987).
5. A.Tokumasu, O.Iwaki, “A Study of Document Format Identification based on Table Structure,” Proc. IEEE Int. Conf. on Image Processing (ICIP'89), pp.593-597 (1989).
6. T.Matsunaga, A.Tokumasu, O.Iwaki, “A Study of Document Format Identification based on Table Structure,” Proc. 1989 IEEE Int. Conf. on Systems, Man, and Cybernetics, pp.845-846 (1989).
7. C.Ono, O.Iwaki, M.Okada, “Drawing Pattern Classification using Contour Convexity,” Proc. IEEE Int. Conf. on Image Processing (ICIP'91) (1991).
8. K.Iwane, M.Yamaoka, O.Iwaki, “A Functional Classification Approach to Layout Analysis of Document Image,” Proc. Int. Conf. on Document Analysis and Recognition (ICDAR'93), pp.778-781 (1993).
9. M.Yamaoka, M.Sato, K.Iwane, O.Iwaki, “A Document Understanding System for Converting Printed Documents to SGML Instances,” Proc. Int. Symposium on Digital Libraries 1995, pp.287-288 (1995).
10. M.Yamaoka, O.Iwaki, “Document Layout Analysis Using Pattern Classification Method,” Proc. 3rd Int. Computer Science Conf. (ICSC'95), Lecture Notes in Computer Science 1024, Image Analysis

Applications and Computer Graphics, Springer, pp.524-525 (1995).

11. A.Takasu, N.Katayama, M.Yamaoka, O.Iwaki, K.Oyama, J.Adachi, “Approximate Matching for OCR-processed Bibliographic Data,” Proc. IEEE 13th Int. Conf. on Pattern Recognition (13th ICPR'96), Vol.III, pp.175-179 (1996).

［国内学会全国大会、研究会等］

1. 徳永,久保田,大谷,岩城, “16 ドット/mm 高解像度サーマルヘッドを用いた新しい階調記録法,” 信学技報, EMC81-12 (1981).
2. 久保田,岩城,徳永, “3-L 階調記録法における最適濃度パターンの設計,” 昭 56 信学情報・システム全大, 184 (1981).
3. 徳永,久保田,大谷,岩城, “3-L 法による階調画像記録,” 第 12 回画像工学コンファレンス論文集, 4-3 (1981).
4. 岩城,久保田,石井, “近接線密度法による文字・図形切り分け処理の検討,” 信学技報, PRL81-81 (1982).
5. 岩城,久保田,荒川, “近接線密度法による文字・図形切り分け処理の検討,” 昭 57 信学通信全大, S6-4 (1982).
6. 岩城,久保田,荒川, “不特定書式文書・図面中の文字認識,” 昭 58 信学情報・システム全大, 87 (1983).
7. 岩城,久保田,遠城,荒川, “文字・図形分離処理におけるプロダクション・システム導入の一検討,” 信学技報, PRL83-63 (1984).
8. 久保田,岩城,荒川, “文書理解システム構成法の一検討,” 昭 59 信学総全大, 1544 (1984).
9. 遠城,岩城,久保田, “文書・図面理解システムにおける処理制御の一検討,” 昭 59 信学総全大, 1545 (1984).
10. 岩城,久保田,荒川, “文字・図形分離処理におけるプロダクション・システム導入の一検討,” 昭 59 信学総全大, 1587 (1984).
11. 木田,岩城,久保田, “ＬＡＮにおける文書理解サービスとメディア処理ステーションの構成,” 情処学「LAN/マルチメディアの応用と分散処理」シンポジウム論文集 (1984).
12. 岩城,荒川, “図表領域中の文字の認識,” 信学技報, PRL84-50 (1984).
13. 遠城,岩城,木田, “文書認識のためのプロダクション・システムの構成,” 信学技報, PRL84-67 (1985).
14. 岩城,木田,遠城, “文書認識ソフトウェアの構成,” 昭 60 信学総全大, 1544 (1985).
15. 遠城,岩城,久保田, “文書認識のためのプロダクション・システムの構成,” 昭 60 信学総全大, 1556 (1985).
16. 岩城,木田, “文書認識システムにおける処理アルゴリズムの検討,” 昭 60 信学情報・システム全大, S4-5 (1985).
17. 岩城,木田,荒川, “機能分散形文書認識システム,” 信学技報, PRU86-32 (1986).

18. 岩城,木田, “文書認識処理の高速化を指向した専用ハードウェアの検討,” 情処学第 33 回全大, 1P-3 (1986).
19. 岩城,木田, “文書認識システムにおける文書構造表現法の検討,” 信学 70 周年記念総全大, 1507 (1987).
20. 村上,岩城, “周辺分布の高周波成分に着目した帳票識別法の一検討,” 情処学第 37 回全大, 6W-9 (1988).
21. 徳升,岩城, “フィールド情報に基づく帳票識別の一検討,” 信学技報, PRU88-114 (1989).
22. 徳升,岩城, “表の領域情報に基づく帳票識別法の検討,” 情処学第 38 回全大, C-8 (1989).
23. 宮島,岩城, “商標の類似度に関する一検討,” 情処学第 42 回全大, 1D-8 (1991).
24. 山岡,岩城, “テキスト・図・表混在文書に対する構造解析手法の一検討,” 1992 信学春季全大, D-569 (1992).
25. 山岡,岩根,岩城, “パターン分類手法に基づくレイアウト解析,” 信学技報, PRU93-125 (1994).
26. 山岡,岩城, “文書画像の SGML 文書への変換に関する一検討,” 信学技報, PRU94-36 (1994).
27. 道坂,吉野,岩城, “レイアウト解析による文書画像分類法の検討,” 情処学第 53 回全大, 3T-7 (1996).
28. 若松,岩城, “Dusk View を用いた時空間断片情報の視覚化,” 地理情報システム学会講演論文集, Vol.6/1997, pp.21-25 (1997).
29. 若松,橋場,岩城, “Dusk View: 透明度を用いた属性表現,” 日本ソフトウェア科学会 WISS’97, インタラクティブシステムとソフトウェア V, pp.107-112 (1997).
30. 橋場,若松,岩城, “3 次元リンク構造を用いた多変量解析結果の視覚化,” 日本ソフトウェア科学会 WISS’97, インタラクティブシステムとソフトウェア V, p.214 (1997).

【出願特許】

1. 久保田,岩城,荒川,石井, “文字領域と図形領域とを区別する方法,” 特願昭56-189005 (1981).
2. 久保田,荒川,岩城, “文字領域と図形領域を区別する図形処理方式,” 特願昭57-110966 (1982).
3. 岩城,徳升,木田,荒川, “帳票認識装置,” 特願平 1-49800 (1989).
4. 佐藤,木田,岩城,荒川, “文字切り出し装置,” 特願平 1-59593 (1989).
5. 宮島,岩城,荒川, “商標類似検索装置,” 特願平 3-30475 (1991).
6. 徳升,木田,岩城, “画像ニ値化装置,” 特願平 3-86578 (1991).
7. 山岡,岩城, “文書の構造解析方法,” 特願平 4-37604 (1992).
8. 山岡,岩根,岩城, “文書レイアウト解析装置,” 特願平 5-260570 (1993).
9. 山岡,岩根,岩城, “文書レイアウト解析装置及び文書フォーマット識別装置,” 特願平 6-50866 (1994).
10. 山岡,岩根,佐藤,岩城, “文字コード生成方法及び文書データベース登録システム,” 特願平 6-218868 (1994).
11. 山岡,岩根,佐藤,岩城, “文字認識装置及び文書内容表示システム,” 特願平7-264234 (1995).
12. 山岡,岩城, “レコード照合方式及び文書検索装置,” 特願平 8-318057 (1996).