博　士　論　文

全体最適化戦略に基づく
複雑な帳票画像の自動認識に関する研究

A Study on Automatic Recognition of Complex Form Document Images based on Global Optimization Strategy

東京農工大学大学院

工学府　電子情報工学専攻

田中　宏

2012 年 9 月

目　次

1. 序論

本章では，本研究の背景と目的，本論文の構成について述べる．

1.1. 本研究の背景

文書画像認識（OCR）技術は，紙文書に書かれた情報を計算機で活用するために，スキャナやデジタルカメラなどで読み込んだ文書画像を認識してデータに変換する技術である．OCR 技術が実用化され始めた当初は，郵便番号読取機のように，固定の書式で読み込んだ文字画像をデータ化する処理が主な目的であった．後に認識できる紙文書の書式が多様化し，帳票データの入力支援としての用途が拡がった．近年では文書画像を PDF 等の電子文書として保存した文書ファイルに検索用のキーワードを自動付与するという用途も拡大している．主に業務目的で用いられる帳票画像認識では認識結果の正確さが要求されるが，近年は認識が困難な複雑な帳票が増えており，実用的な認識精度の維持が困難になっている．一方，帳票に限らず様々な文書が認識対象となるキーワード付与の用途では，従来の技術では扱うことができない複雑かつ多彩な文書が認識対象に加わるため，OCR 技術の更なる高度化が求められる．以上の要因により，複雑で多彩な文書画像を高精度に認識するための技術開発が必要とされている．

多くの業務では，計算機によるデータ処理は業務を効率的に推進するために欠かせないものとなっている．しかし，紙文書は軽く読みやすく特別な装置が無くても読み書きが可能であるなど，人間にとって扱いやすい媒体であるため，多くの業務で紙帳票は依然として重要な位置を占めている．そのため，紙帳票に書かれた情報（表中に書かれた商品名，金額，個数など）を計算機で扱うためにデータ化して読み込むデータ入力作業が必須である．

従来は，紙帳票に書かれた情報は大きな工数をかけて手作業で計算機に入力されていた．その工数を削減するための一手段として，スキャナ等で読み込んだ文書画像を認識して自動的にデータに変換する OCR 技術が利用されている．しかしながら OCR の精度は 100%が保証されているわけではなく，認識誤りを訂正する作業が必要なため，認識精度が不十分な場合にはデータ入力工数がかえって増加したり，入力データの品質が低下してしまうおそれがある．そのため，OCR 技術には実用に耐えるだけの認識精度の確保が求められる．実用化のために必要とされる OCR の認識精度は目的によって異なり，例えば納品書や請求書のような業務帳票では，金額，個数などの数値の誤りは致命的であるため高い認識精度が要求されるが，商品名や会社名のように，後で検索したり，人が読んだりするためのデータの認識では若干の誤りは許容される．

OCR の認識精度を実用レベルに保つためには，かつては郵便振替用紙のような OCR 専用帳票が使われることが多かった．しかし，近年では人にとって読みやすいデザインが優先され，OCR にとって認識が困難な，多彩で複雑な帳票が多用されるようになってきている．また，かつては認識対象となる帳票の書式をあらかじめシステムに登録しておくことによって，既知の書式フォーマットに基づいて帳票画像の認識を行う技術が用いられていたが，近年では書式フォーマットを用いずに帳票画像を認識する技術が求められている．これらの背景により，帳票画像認識技術はより高い精度が要求されるようになっている．

更に，OCR 技術はスキャン画像へのキーワード付与という用途への適用がより重要になっている．紙文書をスキャナ等で画像として読み込み電子的に文書を保存することによって，紙文書の保管コ

ストを削減し，大量な文書を効率的に活用するためのシステムが，主に企業向けの電子ファイリングシステムとして広く提案され，普及している．最近では ECM（Enterprize Content Management）システムという名称が一般化し，スキャン画像だけでなく，Word や PDF 等の電子文書も含めて一貫して文書を管理するシステムが使われている．ECM システムでは，大量に保存された文書ファイルの中から目的とする文書を見つけるためにキーワード検索が用いられることが多い．スキャン画像には文字コードが無いためキーワードを新たに追加する必要があるが，そのキーワード付与を自動化するために OCR 技術が用いられる．ECM システムが扱う文書の多くはビジネス文書であり，従来は単純な書式の文書が多かったが，最近では帳票文書と同様に多彩で複雑な文書も保存されるようになっている．

また，紙文書をスキャンして利用する動きは個人にも広まりつつある．その背景として高性能で安価なドキュメントスキャナの普及，電子書籍端末ブームなどが挙げられる．ドキュメントスキャナは大量の紙文書を高速かつ確実にスキャンできるスキャナである．かつて業務用に数十万円もした高速スキャナと同等性能のスキャナ製品が，数万円を割る安価で発売されるようになり，例えば㈱PFU の ScanSnap シリーズのように累計百万台を超える製品も出てきている．また Amazon 社の Kindle や Apple 社の iPad のように，タブレット型端末で電子書籍を快適に読める製品が次々に発売されており，紙書籍を裁断して一冊まるごとスキャンし，PC や電子書籍端末で読むというスタイルがブームとなっている．このような，紙書籍をまるごと電子ファイルに変換する作業を「自炊する」と呼ぶ（電子書籍をユーザ自身が作るという意味から）ような，新たな流行語も生まれている．この，いわゆる「自炊」作業で作られる電子文書は，文書をスキャンする際に同時に OCR で認識され，全文のテキストが付与されることにより，キーワード検索で目的とする文書ファイルやページを見つけることができる．

ECM システムや個人用途での電子文書の活用（「自炊」を含む）では，様々な書式の文書画像が保存されるため，全文認識を行う OCR 技術は非常に高い精度が要求される．これは帳票画像認識の場合と同様であるが，個人向けの用途でスキャンされる文書は，例えばカラフルな雑誌やマンガ本，新聞，パンフレットなどのように，帳票文書に比べてもより多彩で多様である．また文書中の図表とテキストを分離したり，表の中の文字列を高精度で認識するなどの高精度な認識機能は，帳票認識と同様に必要とされる．ただし，キーワード検索のために文字を認識するという用途に限って言えば，一部の文字に認識誤りが生じても，良く似たキーワードを検索する「あいまい検索」技術を用いれば目的とするキーワードを見つけることはできるので，帳票認識ほど完全な認識精度が要求されるわけではない．

以上で述べたように，OCR 技術の主な用途は，帳票文書からのデータ入力の支援・効率化と，電子文書へのキーワード付与の自動化であるが，いずれの用途においても認識対象の文書はますます多彩で複雑化しており，実用的な認識精度の実現はより困難になっている．

1.2. 本研究の目的

本研究は，OCR 技術の主な適用対象である帳票からのデータ入力支援において，認識対象文書の多彩・複雑化によって実用的な認識精度の維持が難しくなっているという問題に対処するために，帳票画像を高精度に認識する技術を開発するものである．ここで帳票と呼んでいるのは，文書画像

の大半が表領域で構成され，表に記述された文字列データが文書内容の主要な要素であるような文書である．一般に「帳票」とは，業務上の取引を記録する伝票類を意味するが，本研究では雑誌等のページに表が記された場合でも，表領域を含む文書という意味で帳票の一種と考えることとする．

帳票画像認識の高精度化のために，第一に「多彩な」文書画像を高精度に認識する技術を開発する．多彩というのは，様々な色が用いられていると同時に，画像のボケや裏写りなどのノイズも含む劣化画像も意味する．すなわち，従来のように背景と前景（文字・罫線など）が単色でノイズの少ない画像だけではなく，様々な色やノイズが混在した品質の低い文書画像でも精度良い認識する技術を開発する必要がある．そのために，劣化画像から高品質な文字画像を抽出する文字抽出用二値化技術と，カラー文書画像から罫線を抽出する技術を開発する．

第二に「多様な」帳票画像を高精度に認識する技術を開発する．ここで多様というのは，帳票の書式についての多様さである．先に述べたように，複雑な書式の帳票が認識対象画像の中に増えているので，従来のように単純な表だけでなく，複雑な構造を持つ表でも高い精度で認識できる，新たな表認識アルゴリズムを開発する．

以上の開発技術を用いることにより，スキャナやデジタルカメラなどで読み込んだ帳票画像を実用的な精度で認識して，データ入力作業の効率化を支援することが本研究の目的である．

1.3. 本論文の構成

本論文は七つの章から構成される．第 1 章（本章）では，本研究の背景と問題意識について述べ，論文の構成についての概要を記す．第 2 章では，本研究が対象とする帳票画像認識技術の典型的な処理手順を記し，関連する先行研究と本研究の位置付けのについて述べる．続いて第 3 章から第 5 章において，帳票画像認識を構成する要素技術それぞれの改善についての詳細を述べる．第 3 章では表認識高精度化の一環として，劣化画像から罫線を高精度に抽出する技術について述べる．第 4 章は第 3 章で抽出した罫線情報を用いて，複雑な構造の表でもロバストに解析して，表を構成するセル領域を抽出するセル抽出技術について述べる．第 5 章では劣化画像から文字の二値画像を高精度に抽出する技術について述べる．第 6 章では，帳票画像認識を高精度化するための要素技術を実用化するために必要とされる，周辺技術について述べる．最後に第 7 章において，これまでに述べた内容をまとめた結論を記す．

各章の内容については，下記であらためて詳細に記述する．

第 2 章　背景技術と課題

第 2 章では，帳票画像認識技術について概観し，本研究の位置づけを述べる．まず，帳票画像認識の典型的な処理手順について記述し，その主要な処理モジュールが「レイアウト解析」,「表認識」,「文字認識」であることを述べる．続いて，それぞれの処理モジュールに対応した主な先行研究について概観する．それを受けて，帳票画像認識の精度が主に表認識と文字認識によって左右されることを示し，本研究が表認識と文字認識の精度を向上させるための要素技術として，「罫線抽出技術」,「セル抽出技術」,「文字抽出用二値化技術」についての検討を行うものであることを示す．

第 3 章　罫線抽出技術の研究

第 3 章では，表画像から罫線を高精度に抽出する技術について述べる．多彩なデザインの表画像から罫線を抽出するためには，背景ノイズの影響や，低解像度画像における画像のボケ，近接文字と罫線の誤認などによる罫線の付加・脱落誤りを解決する必要がある．本研究においては，ボケやノイズの影響を軽減するために，ラン線分抽出とエッジ線分抽出を併用した罫線抽出方式を開発した．また，局所的二値化の閾値補正技術を開発し，薄い罫線の途切れを選択先に抑制する技術を開発した．更に，文字画像から誤抽出した罫線を削除する技術も開発した．これらの技術を用いて，多彩で低品質の表画像から罫線が高い精度で抽出できることを示す．

第 4 章　セル抽出技術の研究

第 4 章では，抽出した罫線の情報に基づいて，表を構成する項目セル領域を抽出する技術について述べる．多彩なデザインの表画像からセル領域を抽出する場合，罫線抽出の結果には誤りが含まれる可能性が高いため，セル抽出技術には罫線誤りの影響を受けにくい頑強性が求められる．更に，複雑な構造の表であっても解析できるアルゴリズムも必要である．本研究では，入力された罫線情報に基づいて罫線が交差する交点を尤度付きで生成し，交点の組み合わせ探索を用いた最適化アルゴリズムによってセル領域を抽出する新しいセル抽出技術を開発した．開発した技術により，多彩で複雑な表画像から高い精度でセル領域が抽出できることを示す．

第 5 章　文字抽出用二値化の研究

第 5 章では，文字認識の精度向上を目指して，文字の二値画像を高精度に生成する二値化技術について述べる．多値の文字画像を認識する場合でも，多くのシステムでは，文字を白黒の二値画像に変換してから文字認識を行っている．そのため，文字認識に適した二値画像を生成する，文字抽出用二値化の改善が，文字認識の精度向上のために有効である．本研究では，低解像度画像における文字画像のボケや，近接罫線の影響による文字の二値画像の劣化を改善するために，局所的二値化の閾値補正技術を用いた文字抽出用二値化技術について述べる．更に，大域的二値化の代表的な手法である大津二値化を入力画像の解像度に応じて併用することによって，解像度に寄らずに高精度な二値化が実現できることを示す．

第 6 章　結論

第 6 章では，本研究の開発成果を実用化する上で必要となる周辺技術について述べる．帳票画像認識を高精度化するための要素技術は，一般文書画像認識の精度向上にも貢献するため，帳票画像認識と一般文書画像認識それぞれの主要な用途について，実用化のために考慮すべき点について考察した．帳票画像認識は，主に帳票に記述された業務データを入力するために用いられるため，誤りの無い正確なデータが入力できなければ業務に影響が出てしまう．そのために，帳票画像認識の結果を人手で再確認したり，認識結果を二重チェックするなどの対策が必要とされるという点について述べる．一方，一般文書画像認識は，電子文書を再構成するためにレイアウト解析技術の高精度化が必要であるという点や，認識結果をテキスト検索に用いる

ために，認識候補ラティスを利用できる仕組みが望ましいという点を示す．更に，本研究で開発した技術の実用化状況についても報告する．

第 7 章　結論

最後に第 7 章において，本論文の内容を概観し，本研究のまとめを記す．

2. 背景技術と課題

概要

本章では，本研究の背景技術について述べ，本研究の位置付けと解決しようとする課題について記述する．

2.1. はじめに

近年，多くの業務では計算機を用いたデータ処理が不可欠となっている．しかし納品伝票や請求書，各種契約の申込書などのように，依然として紙帳票が使われ続けている業務も少なくない．紙帳票が業務に使われている場合でも，元データは計算機上で作成・管理されている場合がほとんどであり，紙に記入されたデータを計算機に登録する作業が必要とされる．そのようなデータ入力作業の効率化のために，帳票画像認識技術が用いられる．

帳票とは文書の一形態であり，文書コンテンツの主要部分が表によって構成され，表に記された文字列や数値などのデータを記録することを主目的とした文書である．帳票も文書の一種なので，帳票画像認識技術は一般文書を対象とした文書画像認識技術と同様，「レイアウト解析」，「表画像認識」，「テキスト認識」という要素技術から構成される．**図 2-1** が文書画像認識の典型的な構成であり，処理の流れを模式図で示したのが**図 2-2** である[1].

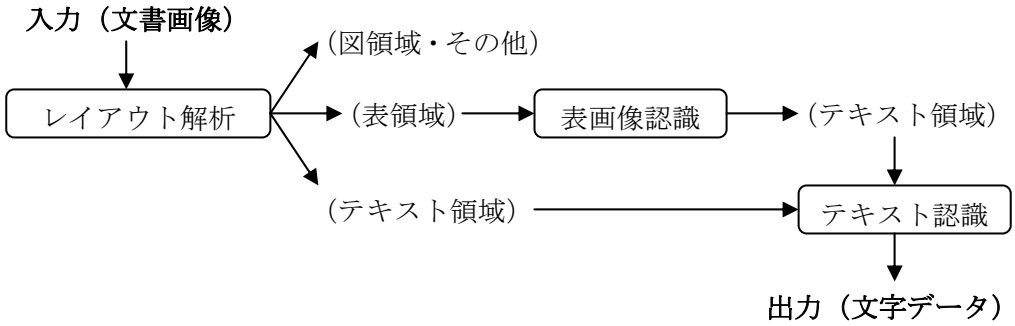


図 2-1　文書画像認識の典型的な構成

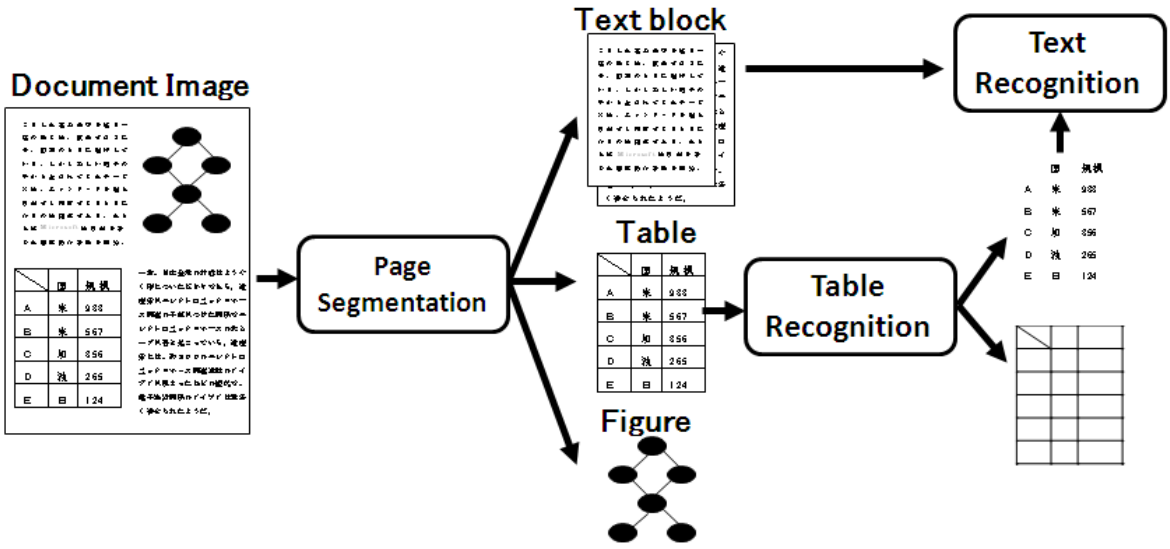


図 2-2　文書画像認識の典型的な処理の流れ

帳票画像は，文書の主要部分を表が占めており，帳票が表す内容の大半が表内の文字列として記述されているような文書画像である．例えば**図 2-3** に一般文書画像の例，**図 2-4** に帳票画像の例を示す．一般文書画像では，文書中に複数の表やグラフ，テキスト領域が混在しているため，文書画像を高精度で認識するためには，各領域を分類するレイアウト解析の精度が重要である．一方，帳票画像は表領域が占める範囲が多く，またレイアウトが単純であるため，レイアウト解析の負荷は比較的小さい．したがって，帳票画像認識の精度を向上するためには，レイアウト解析よりも，表認識とテキスト認識が重要であることが分かる．

図 2-3　一般文書画像の例

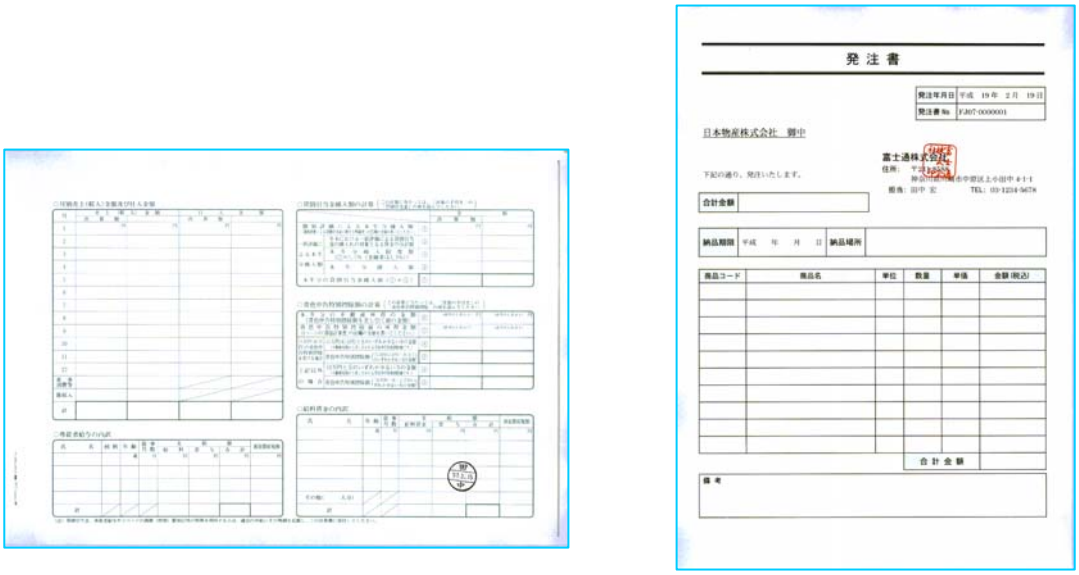


図 2-4　帳票画像の例

本章では，帳票画像認識技術を構成する各要素技術について，先行研究の紹介と本研究との関連について述べる．まず第 2.2 節では，レイアウト解析の標準的な手法と，先行研究について記す．

前述のように，帳票画像認識ではレイアウト解析の負荷は比較的小さく，簡便な手法でも十分に実用になる．しかし，帳票画像でも表外に文字列や図は存在するので，複雑なレイアウトを解析する技術も将来的には必要となる可能性がある．そのため，非常に複雑な文書レイアウトを解析するための先行研究についても合わせて述べる．第 2.3 節では，表画像認識の背景技術について述べる．文書中で表を表現する方法には，罫線で表項目を区切る罫線表と，罫線を用いずに表項目の位置関係だけで表を表現する無罫線表がある．一般に国内では罫線表が多く用いられ，海外では無罫線表や，一部の罫線が省略された表が用いられることが多い．第 2.3 節では，先ずこれらの区別について述べ，本研究が罫線表のみを対象としていることを述べる．また，表画像認識技術を構成する，罫線抽出とセル抽出のそれぞれについて，先行研究の紹介を行う．第 2.4 節では，テキスト認識技術の概要と，特に文書画像の二値化技術について述べる．テキスト認識は，レイアウト解析で抽出したテキスト領域の画像を認識して，文字コード列に変換する処理である．そのためには文字認識技術が必要であるが，文字認識技術には，二値画像を認識する技術と，多値画像を認識する技術とに分けられる．本研究では二値画像を認識する文字認識技術を用いるが，多値画像を用いた文字認識技術についても概観する．また，二値画像を文字認識で用いるため，あらかじめテキスト領域の画像を二値化する技術が文字認識の精度に大きく影響する．本研究では，文字抽出用の二値化技術を改善することによって文字認識の精度を向上するというアプローチを取っており，そのための背景技術として，二値化技術の先行研究を記述する．文書画像を二値化する方法は，文書全体で単一の二値化閾値を用いる大域的二値化と，画素ごとに閾値を設定する局所的二値化とに分類できる．本研究が対象とする文字抽出用二値化技術は，これらの２種類の方式を適宜組み合わせたものである．ここでは，２種類の二値化技術の基本知識と，様々な工夫を施した先行研究の紹介をする．更に，文字抽出用二値化という観点で本研究の範囲を設定し，関連技術の説明と本研究の位置付けについて述べる．最後に第 2.5 節で本章のまとめを述べる．

2.2. レイアウト解析

帳票画像はレイアウトが比較的単純なことが多いので，通常は一般文書のレイアウト解析を簡略化して用いることが多い．そこで，本節では一般文書のレイアウト解析の先行研究について述べ，それを簡略化することで帳票画像のレイアウト解析が実現できることを示す．

文書のレイアウトを解析する試みは古くから行われている．初期の試みにおいては，文書の書式にある程度の制約を設けて解析する方法が主流である．代表的なのは，いわゆるマンハッタンレイアウト（Manhattan Layout）と呼ばれる，文書を構成するテキストや図表などの領域が縦横の直線で分割でき，各領域が矩形で囲まれたレイアウトに文書を限定する方法である．文献[2]のサマリーによれば，マンハッタンレイアウトの文書を解析する主な手法には４つの種類がある．第一の手法はランレングスを用いた領域分割である[3]．**図 2-5** は文献[3]から抜粋した処理画像の例である．左端の画像から，ｘ方向とｙ方向にそれぞれ連続した黒画素（ラン）を求め，一定長を超えるランが存在する領域の近傍を塗り潰して論理和を求めることにより，テキストとそれ以外の領域を分離している．第二の手法は X-Y カットによる領域分割である[4]．**図 2-6** に示すように，ページを縦・横に順に直線でカットすることによって領域を分離するものである．第三の手法は領域間の空白領域（White Space）を用いた分割である[5]．文字や図表などの前景画像に比べて，背景の方が単一

性が高くグループ化しやすいという観点に基づき，領域間を区切る空白領域を検出して領域の分離を行うというものである（**図 2-7**）．第四の手法は X-Y カットと空白領域による分割を組み合わせたものとされている[6]．

文献[2]では以上の４種類に手法が分類されているが，第三の手法である空白領域の利用は，領域間をどのようにカットするかという問題に対する解であり，第二の手法とは矛盾しない．第四の手法は第二，第三の手法の組み合わせなので，独立した手法に分類するのは適当ではない．したがって，マンハッタンレイアウトの文書を解析する手法は第一と第二の手法の２種類に分類でき，解析の精度を向上させるために，第三，第四の手法のようなバリエーションが存在するのだと解釈することができる．過去の多くの研究では X-Y カットに基づく手法が多用されている．

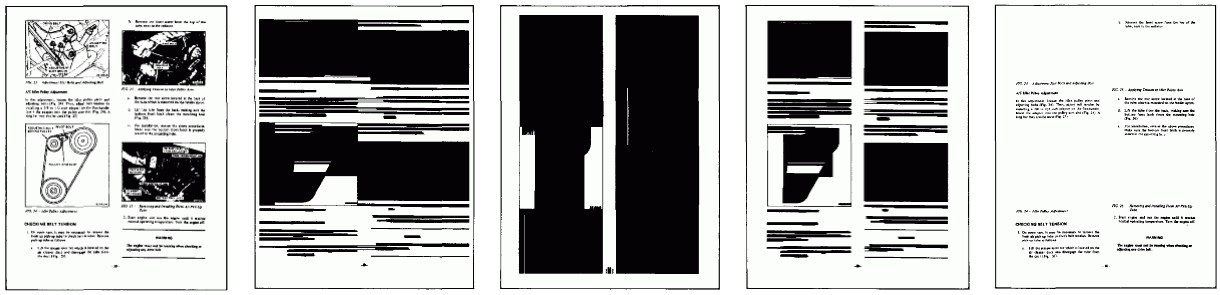


図 2-5　ランレングスに基づく領域分割（文献[3]の Figure 2 より）

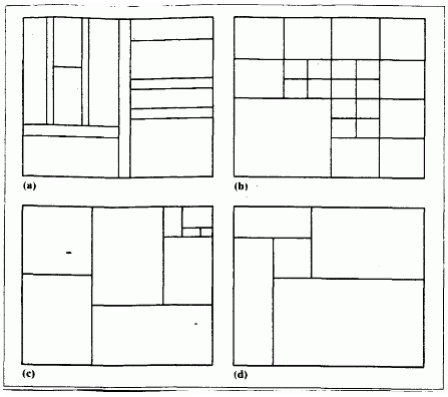


図 2-6　Ｘ－Ｙカットに基づく領域分割（文献[4]の Figure 2 より）

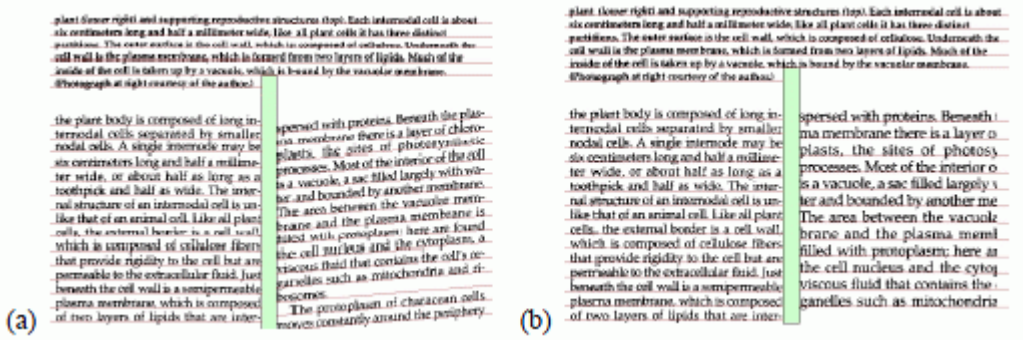


図 2-7　空白領域による領域分割（文献[6]の Figure 3 より）

一方で，非マンハッタンレイアウトの文書を対象とする場合は別の手法が必要である．主なアプローチの一つは，画素の連結要素（Connected Components）を用いて領域を分類する手法である．文献[7][8]は，文字や図表を構成する連結要素を相互に関連付けてテキストや図表を分類する．**図 2-8** は文献[8]において，連結要素を用いてテキスト行を抽出する例である．文献[7][8]は文字や図表などの前景要素の連結要素を用いた例であるが，文献[9]は，背景の空白領域の連結要素を連結して，図表などの領域を分割する．これらの手法は，マンハッタンレイアウトの解析で仮定していた「領域は直線で区切ることができる」「領域は矩形である」という条件を排除して，連結要素とを用いて非直線の領域切断を可能にしたものだと言える．

このような非直線による領域切断というアプローチをより柔軟かつ一般化したものがボロノイ図を用いた領域分割である[2][10][11]．**図 2-9** に文献[11]から抜粋したボロノイ図の生成過程と，ボロノイ図を用いた領域分割の例を示す．これを**図 2-8** と比較すると，ボロノイ図は文字や図表の間の背景領域に着目して文字・図表の区切り線を生成するものであり，連結成分を用いた手法と類似していることが分かる．文字や図表が複雑な位置に配置されている文書では，このように自由度が高い手法が必要とされる．

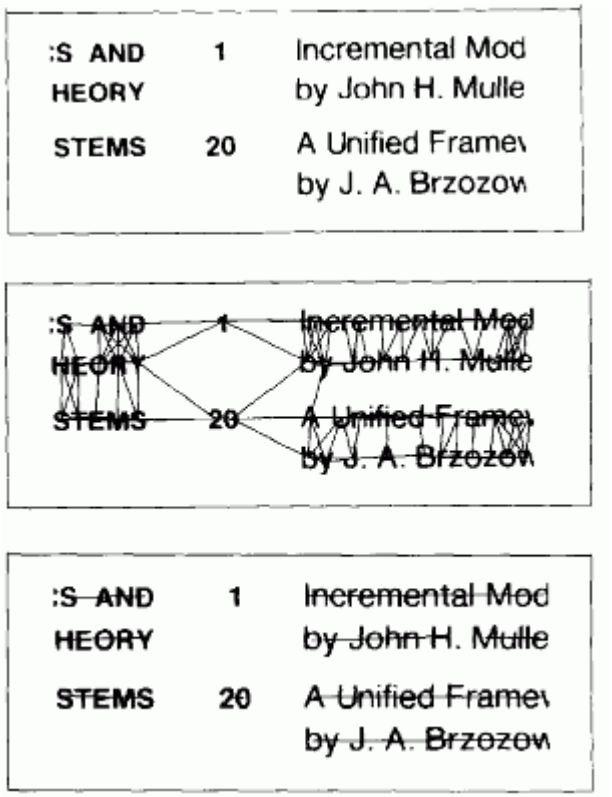


図 2-8　連結要素によるテキスト行の抽出（文献[8]の Figure 1,7 より）

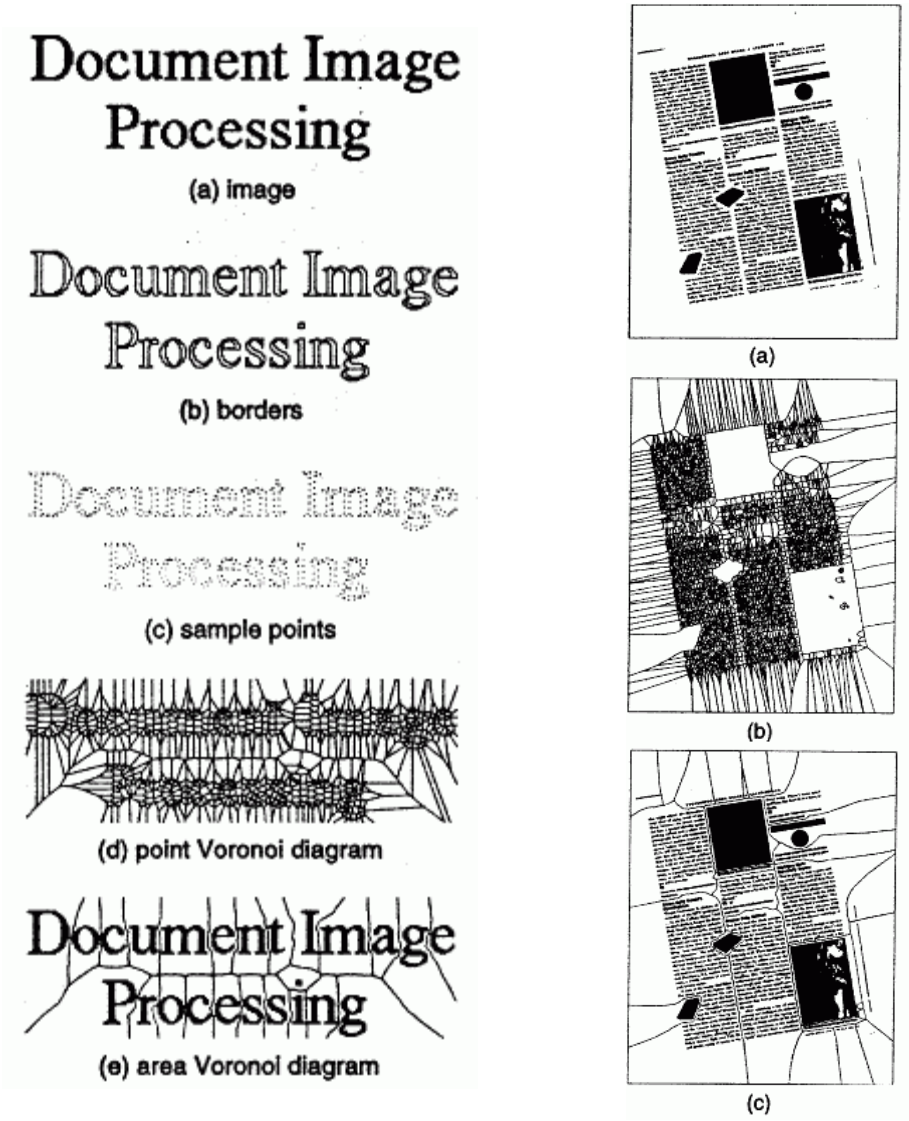


図 2-9　ボロノイ図の生成と文書領域分割（文献[11]の図１と図４より）

さて，以上に示したように文書のレイアウト解析には様々な手法が提案されているが，帳票画像認識で用いる場合には，さほど複雑なレイアウトを解析する手法は必要とされていない（ただし近年では帳票でも複雑なレイアウトが使われるようになっており，上記のような自由度の高い手法が将来的に必要とされる可能性は高い）．

図 2-1 に示したように，文書画像認識の典型的な構成では，文書画像からテキスト領域を抽出して，文字データを取得することが第一の目的である．つまり，文字情報を高精度で認識することが目的であり，図や写真などの文字以外の領域は「文字ではない」ということが分かれば良い．そこで，テキスト領域を抽出する，という点に特に注目した方式も提案されている．

文献[12]では，文字を表す連結要素を抽出し，近傍の連結要素をグループ化（文献中では群化と呼んでいる）することによってテキスト領域を抽出する．連結要素は，その外接矩形のサイズによって文字要素か否か判定されるが，文字とみなされた連結要素の中にも文字以外のものが含まれる可能性があるため，グループ化が終了した後に各テキスト領域を文字認識し，その結果が悪いものは文字以外に分類するなど，精度向上の工夫が施されている．文献[12]の方法を基本として，複雑

なテキスト配置に対しても高精度なテキスト抽出を実現するために，文字間の空白矩形を繰り返し抽出してテキスト領域抽出の精度を高めるという試みも提案されている[13][14].

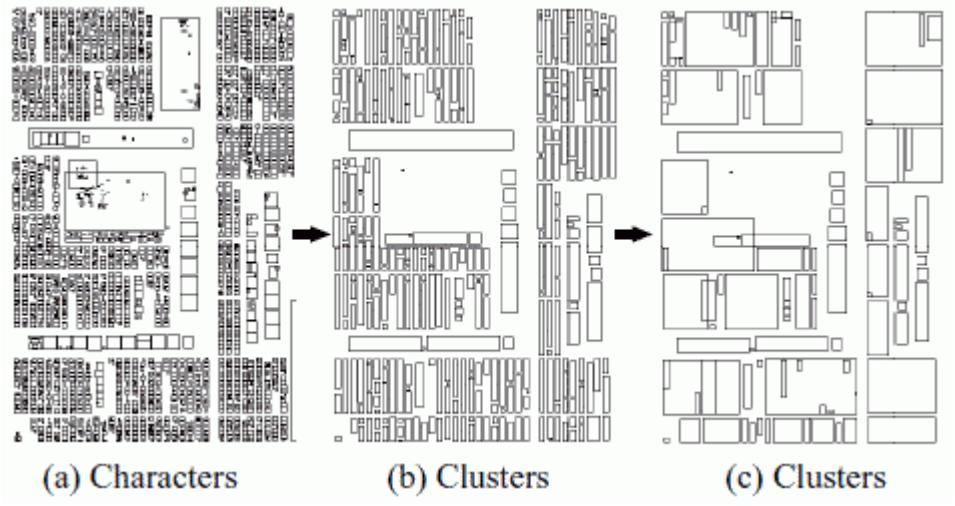


図 2-10　文字成分の群化（参考文献[12]の図４より）

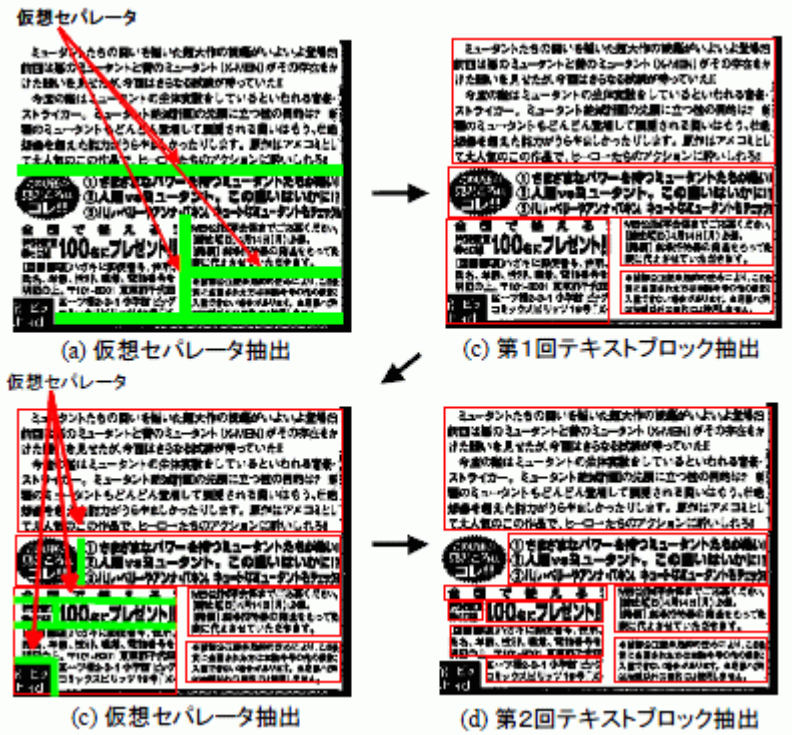


図 2-11　再帰的に空白矩形で領域を分離（参考文献[13]の図１より）

本研究では，レイアウト解析の改善は研究対象とはしておらず，既存の手法に基づいて適切な方式を利用する．特に，帳票画像認識の目的が文字データの抽出であるため，文献[12]の方式を用いてテキスト領域と図表領域を抽出し，更に図表領域を表認識することによって表の項目セルを抽出（表認識に失敗すれば図領域と判定）し，セル内のテキスト領域と，表外のテキスト領域とを文字認識することによって，文字データを抽出するという方法を採用している．

2.3. 表画像認識

表画像認識は，レイアウト解析によって抽出された表領域の部分画像を対象として，表を構成するセル項目領域を抽出する技術である．本研究で対象とする表は，表項目が罫線で区切られた罫線表のみであるが，世の中では，罫線で区切られていない無罫線表や，一部の罫線しか描かれていない表などもよく使われる．本節では，まず罫線表と無罫線表について述べ，続いて本研究の主要テーマである罫線抽出技術とセル抽出技術について記述する．

2.3.1. 罫線表と無罫線表

帳票画像認識は業務向け用途への適用が先行したので，先行研究には納品書や請求書などの伝票類を認識対象とするものが多い．海外の研究では，インボイス（invoice＝送り状）の認識が主要な研究対象となっている．例えば文献[15]に掲載されている例（**図 2-12**）のように，表の行と列を区切る罫線は存在せず，同じ高さ（文字列の y 座標が同じ）や桁（文字列の先頭の x 座標が同じ）にある項目を対応付けることで，表構造を表したものがある[16][17].

このような無罫線表は，先に述べたレイアウト解析の手法（例えば**図 2-10**）では，個々の表項目が単独の文字列と認識され，まとまった表領域とはみなされない．したがって，無罫線表を表と認識するためには，文字認識を行った後で，各テキストの位置関係を解析する必要がある．

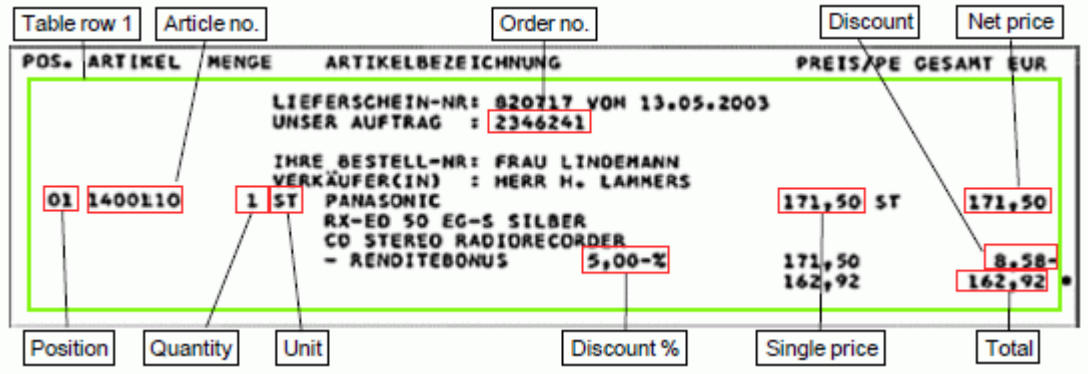
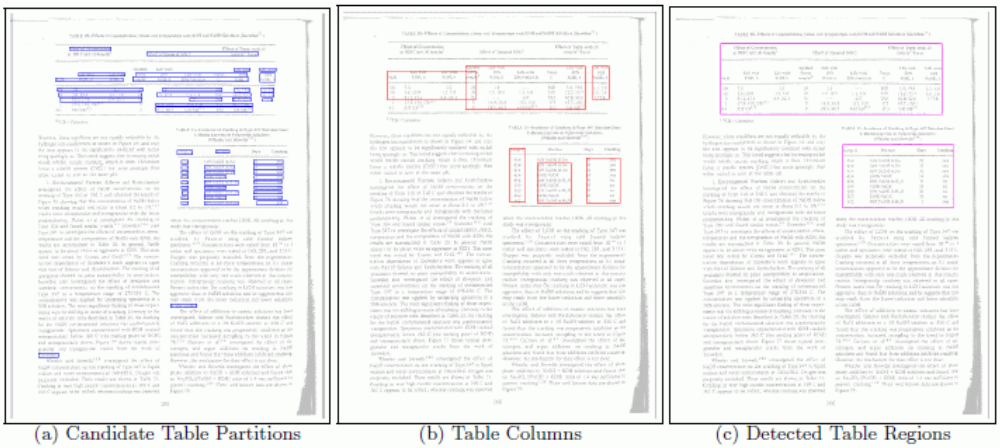


図 2-10　海外のインボイスの例（参考文献[15]の Figure 2 より）



(a) Candidate Table Partitions　(b) Table Columns　(c) Detected Table Regions

図 2-13　無罫線表の領域抽出の例（参考文献[16]の Figure 3 より）

図 2-13 は無罫線表の領域を抽出した例である[18]．先ず表領域の候補となるテキストを抽出し，表のカラムを検出した後に表領域を同定している．このように，無罫線表の場合はテキスト領域を抽出した後に，その後処理として表領域の抽出を行うという方法が一般的である．この時点でテキスト領域は先に抽出できているので，**図 2-1** に示すように表領域を抽出してから表認識を行い，更に表内のテキスト領域を抽出するという手順を取る必要は無い．

本研究においては，表項目が罫線で区切られた罫線表のみを表認識の認識対象とする．**図 2-1** に示すように，表認識はあくまで表内のテキスト領域を高精度で抽出するための手段と位置付けているので，先にテキスト領域が抽出できている無罫線表は，本研究における表認識とは問題領域が異なるためである．

罫線表を対象とした表認識は，表を構成する項目セル領域（その内部にテキストが存在する）を抽出するために行われる．セル領域は罫線で囲まれているので，セル領域の抽出は罫線で囲まれた閉領域を抽出する問題に帰着する．そのため，典型的な手法の一つは，まず罫線を抽出し，続いて抽出した罫線を用いて閉領域を同定するという二段階の処理を行う手法である．一方で，罫線抽出は行わずに，罫線が交差した交点を直接用いる方法や，一般化 Hough 変換のように画像中から矩形を抽出する方法などのように，罫線を明には用いない方法もある．これらの詳細は 2.3.3 節で述べる．

次節では，まず罫線抽出についての先行研究について記す．続いて，罫線抽出を用いる方法，用いない方法の両方を含めて，セル抽出の先行研究について述べる．

本研究では，表認識を罫線抽出とセル抽出の二段階で行う方法を採用している．よって，それぞれの先行研究と本研究との関連については，それぞれ第 3 章と第 4 章でも詳しく記述している．

2.3.2. 罫線抽出

罫線抽出は，表画像の中から罫線と思われる直線を抽出する技術である．従来，比較的単純な表画像が認識対象であった場合には特別な処理は必要では無く，単に二値画像の黒画素が縦横それぞれの方向に一定長を超えて連続する並び（ラン）を抽出するだけで事足りていた．例えば文献[19]には，水平方向と垂直方向のランによる画像を太らせて，それらの論理和を求めることによって罫線のみの画像が得られる，という記述があるのみである（**図 2-14**）．

表画像が傾いている場合には，水平，垂直方向のランでは罫線が正しく抽出できないため，傾いた直線でも抽出できる方法を用いる必要がある．典型的な手法としては，Hough 変換を用いて直線を抽出する方法があるが[20], 文献[21]においては, Hough 変換は処理時間がかかるという理由で，罫線のエッジを短い線分の集合で表して，線分のチェーンコードを追跡することによって，表画像の傾き補正と罫線抽出を実現している．

しかし表の構造が複雑になり，罫線の座標をより正確に求める必要が生じた場合や，表中の文字が多く罫線と文字が接触しているような場合, 表画像の品質が悪い（解像度が低い, ノイズが多い, など）場合などには，精度の高い罫線抽出技術が必要となる．罫線と文字が接触した場合の問題については，如何にして文字画像から罫線を分離して高精度な文字画像を生成するかという問題が文献[22]において議論されているが，文字画像が罫線に誤認する問題については，先行研究では十分に議論されていないので，本論文の第 3 章において詳細に議論する．ノイズの多い低品質画像から

罫線を抽出する方法については文献[23]などの例がある（**図 2-15**）.

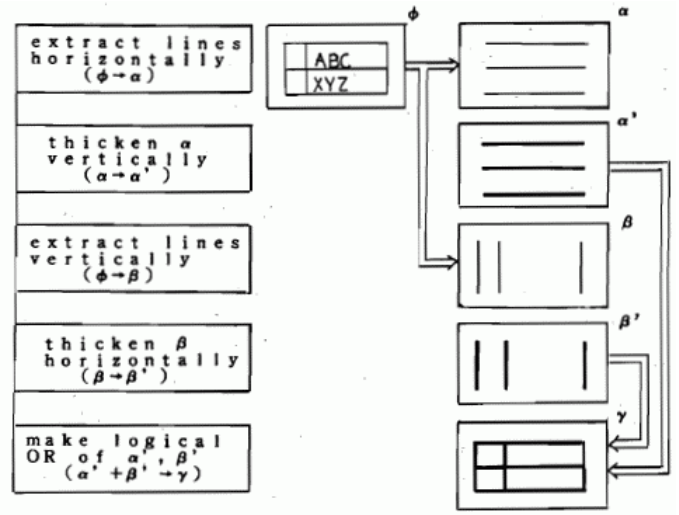
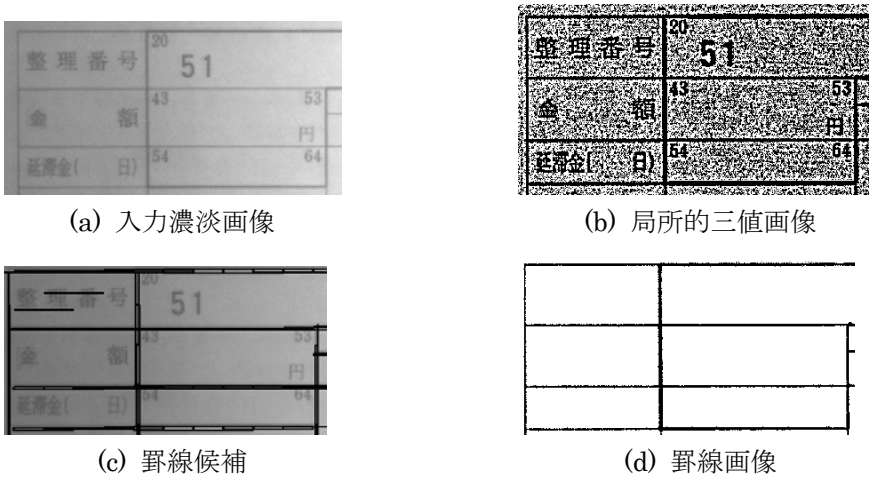


図 2-14　罫線画像を生成する手順の例（参考文献[19]の図５より）



(a) 入力濃淡画像　　(b) 局所的三値画像

(c) 罫線候補　　(d) 罫線画像

図 2-15　低品質画像からの罫線抽出（参考文献[23]の図１～５より）

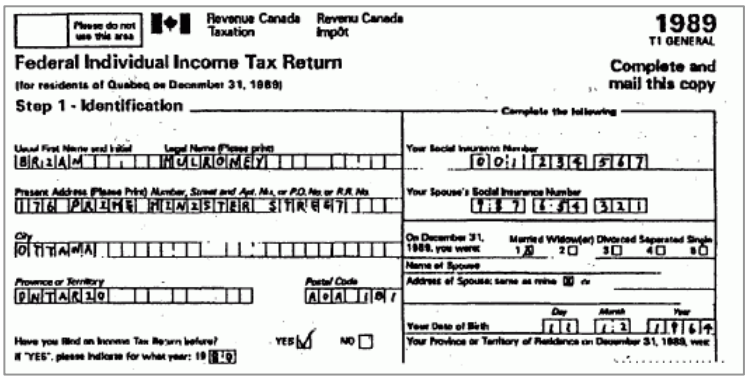
2.3.3. セル抽出

セル抽出は，表を構成する項目領域を抽出する技術である．先に述べたように，ここでは罫線表を対象とした先行研究について述べる．

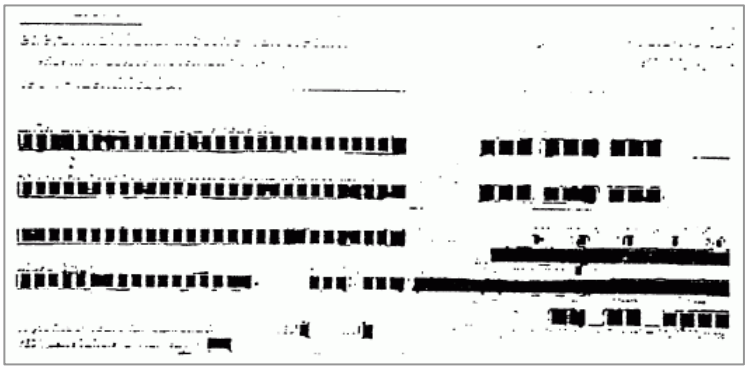
セル領域は文字列データを記入するための領域なので，単純な構造の表では矩形であることが仮定できる．そのため，多くの先行研究においては，セル抽出は矩形の閉領域を抽出する技術として実現されている．

画像から矩形領域を抽出する方法には，罫線情報を用いずに画像情報から直接矩形領域を抽出する方法と，罫線に基づく情報を用いる方法がある．前者には，Geometric Hashing によって矩形の中心座標を抽出する方法[24]や，エッジ画像から一般化 Hough 変換を用いて矩形領域を抽出する方法[25]が提案されている．後者には，表を縦横の平行な罫線で順に分割する方法[26]～[28]（2.2 節で述べた X-Y カットによる領域分割と同様）や，罫線が交差する交点を利用する方法がある[29][30].

画像から直接抽出する方法は，抽出した矩形領域が表項目セルなのか，文字や図形なのかの区別が難しいという問題がある（例えば「口（くち）」や「目」「曲」など）．また，セルの中に小さなセルが配置された入れ子構造の場合には，矩形領域が重なって抽出されるため，矩形抽出結果から表構造を再現するのは困難である．そのため，文献[25]では文字を記入するための小さな枠のみが抽出対象となっている（**図 2-16**）．



(a) 帳票画像



(b) 矩形領域抽出結果

図 2-16　Hough 変換による矩形領域抽出（参考文献[25]の Figure 3,4 より）

これらの例で分かるように，画像処理的に矩形領域を抽出する技術は，それ単独で項目セルを抽出する技術として用いるのは難しいように思われる．もちろん，Hough 変換によって矩形領域の候補を求めた後で領域の検証を行うなど，セル抽出の要素技術として用いるという方法は考えられるが，先行研究においては，表認識のために Hough 変換などを用いた例はあまり多くない．

セル抽出で頻繁に用いられるのは，罫線を抽出して，罫線に囲まれた閉領域を求めるという方法と，交点を抽出して４つの交点の組み合わせを求めるという方法である．前者の罫線を用いる方法の例を**図 2-17** に示す[28]．これはマンハッタンレイアウトの文書を X-Y カットで分割する方法と原理は同じである．表が矩形のセルで構成されているという前提の元で，表領域を直線で再帰的に切断することによって矩形のセル領域に分割する．実際には，セルが全て矩形であっても X-Y カットで分割できない表構造もあり得る（**図 2-18**）が，先行研究では，そのような構造は対象外とされている．

交点を用いた代表的な先行研究[30]では，**図 2-19** のように表の構造を格子状の交点で表している．交点は罫線情報から生成されたもので，文献の中には４つの交点（corners）の組み合わせがセル領

域に対応するという記述がある．

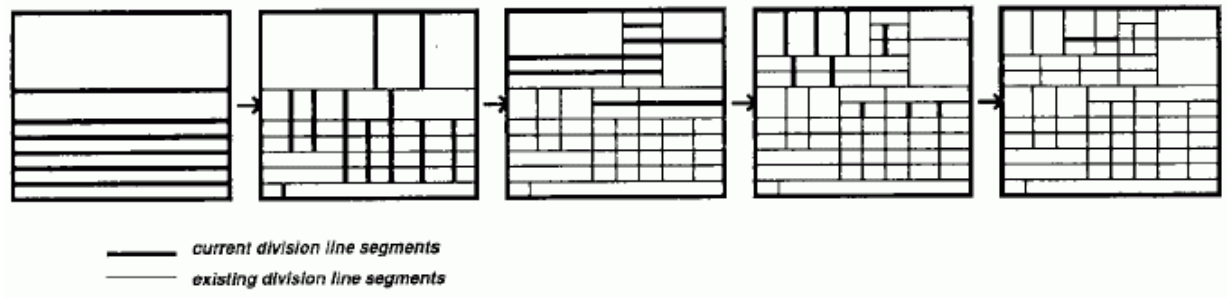
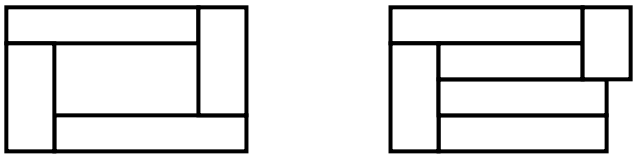


図 2-17　表を罫線で X-Y カットする方法（参考文献[28]の図３より）



(a) 卍型の表構造　　(b) 卍型の変形

図 2-18　X-Y カットでは分割できないセル領域

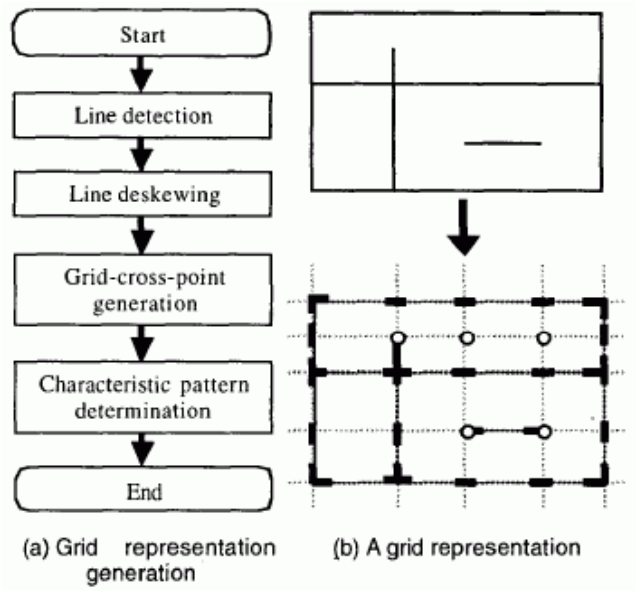
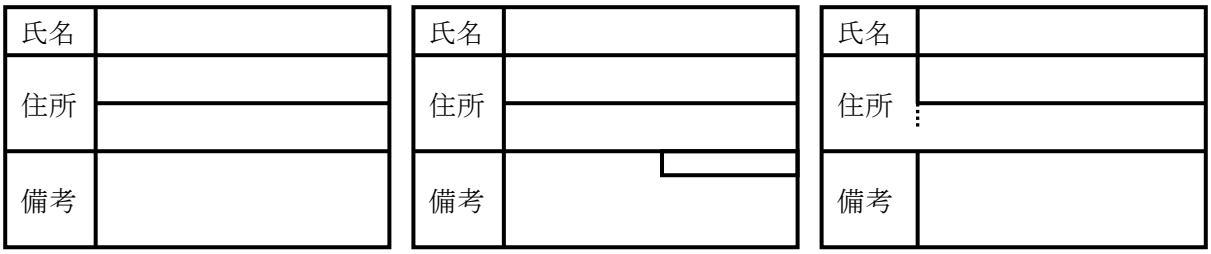


図 2-19　交点を用いた表構造の表現（参考文献[30]の Figure 1 より）

このように，ほとんどの先行研究ではセル領域は４つの角を持つ矩形の領域であることが前提となっている．しかしながら，例えば図 2-20(a)のように矩形セルのみで構成された表がある場合，図 2-20(b)のように小さな矩形セルが一つ追加されると，そこに非矩形セル（Ｌ字型セル）が生じる．また，図 2-20(c)の例では，罫線の一部が途切れた場合にＬ字型セルが生じている．このように，本来は矩形のみで構成されていた表構造も，若干の修正や罫線抽出の誤りによって，非矩形のセルが容易に生ずることが分かる．セルの形状を矩形のみに限定したセル抽出技術では，このような場合に，表を構成するセルの一部が抽出できなくなる可能性がある．



(a) 矩形セルのみの表　(b) 小矩形によるＬ字型セル　(c) 罫線途切れによるＬ字型セル

図 2-20　Ｌ字型セルが生成する例

本研究では，先行研究に見られるようにセル領域の形状を矩形に限定するということは行わず，Ｌ字型などの非矩形セルも抽出対象とする．また卍型の表構造も扱うことができる方式を目指す．これにより，**図 2-20(c)**に見られるように，罫線抽出の誤りがセル抽出に大きく影響することが無いような技術を開発する．開発技術の詳細は第 4 章にて述べる．

2.4.　テキスト認識技術

2.4.1.　テキスト行認識技術

レイアウト解析によってテキスト領域が抽出されると，次にテキストを行に分離して，各行の認識を行うことによって文字列認識結果を得る．レイアウト解析が文字の並びを意識してテキスト領域を抽出した場合（**図 2-8**）には，抽出したテキスト領域の行はあらかじめ分かっているが，そうでない場合には，複数行を含むテキスト領域から行を抽出する処理が必要となる．典型的な処理は**図 2-21** に示すように黒画素の数を横方向にカウントしてヒストグラムを作り，谷の部分を行間だとみなす方法である．これは行が横方向の場合の例であるが，縦方向でも同様の処理を行って，行間の谷の深さや幅を比較したり，縦横で文字認識を行って，平均認識得点が良い方を行方向とみなすなどにより，行方向も判定できる．このような方法は，文献[31]をはじめとして多くの先行研究が採用している．

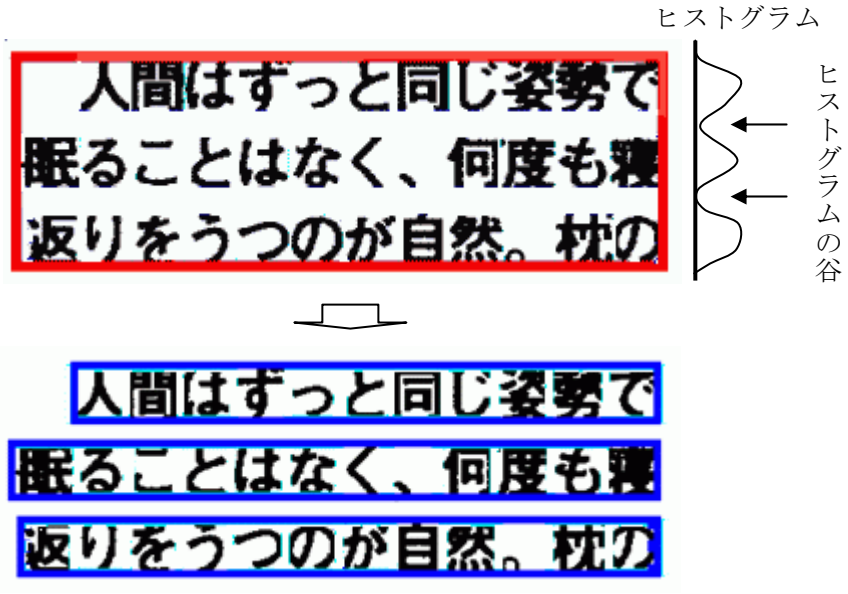


図 2-21　画素のヒストグラムに基づく行抽出

テキスト領域の行が抽出されると，行を一文字単位に分離することができれば，一文字ごとに認識を行う文字認識技術を適用することによって，認識結果テキスト列を得ることができる．しかし文字の境界は事前には確定していないので，複数の文字境界候補を生成して，最適な文字境界のセットを選択するという方法（オーバーセグメンテーション）がしばしば採られる．この方法も非常に一般的なもので，数多くの先行研究が存在するので，ここでは代表的なものを一つ示すにとどめておく[32].

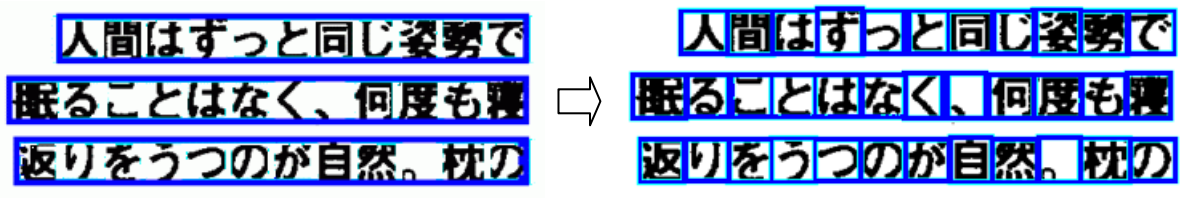


図 2-22　テキスト行の文字単位への分離

2.4.2.　文字認識技術

文字画像を認識する技術は OCR 技術の根幹であり，長い歴史の中で様々な特徴量や距離計算の方法，識別アルゴリズム等が提案されている．例えば，旧郵政研究所が主体となって，平成 6 年と平成 8 年の 2 回に渡って手書き数字を認識するコンテストを行い，有望な特徴抽出法と距離計算法の比較や，それらのいくつかを組み合わせた効果に関する研究を行った例がある[33][34]．その結果によれば，セル特徴[35]，加重方向ヒストグラム特徴[36]，外郭方向寄与度特徴[37]などが優れた特徴として挙げられている．一方，距離計算法についても多くの方式が提案されており，二次識別関数の誤差問題を改善した疑似ベイズ識別関数[38]や，類似文字識別のための補正項をマハラノビス距離に導入した混合マハラノビス関数[39]などが知られている．

ここで述べた先行研究では，認識対象の文字画像としては二値画像を用いている．それぞれの手法は二値画像でなければ適用できないというものではないが，文字は本質的に背景平面の上に乗ったオブジェクトの形状を示しており，理想的には二値で表し得るものであることや，明度やカラー値のような多値画像は扱いが困難である（例えば値の変動が激しい）こと，認識処理の計算量の負荷が大きいことなどから，多くの先行研究は二値画像を用いている．

しかしながら，入力画像の品質が低い場合，二値画像に変換する処理（二値化）における文字形状の乱れが文字認識の精度に大きな影響を及ぼすという問題が深刻である．そのため，多値画像を二値化せず多値のままで認識する方式も提案されている．例えば文字の濃度勾配を用いた方式[40]や，二重固有空間（Dual Eigenspace）を用いた方式[41]などである．更に，文字品質が良い場合は多値画像よりも二値画像を用いた方が精度が高いという知見を利用して，二値画像ベースの識別手法と多値画像ベースの識別手法を組み合わせることによって高精度化を図る研究も発表されている[42].

多値画像を用いた文字認識手法は，識別技術の改良によって文字画像の品質劣化に対応しようとする試みである．それに対して，従来と同様の二値画像ベースの認識技術を用いつつ，二値化技術を改善することによって文字認識の精度を向上させようという方向性も考えられ，本研究はそのような立場を取っている．そこで，本研究の関連技術を記述する上で重要なので，文書画像の二値化

に関する関連技術について次節であらためて記述する．

2.4.3. 文字抽出用二値化

画像の二値化とは，多値画像の各画素が文字や図表などの前景に属するか，背景に属するかを判定して，二値画像を出力する処理である．これにより文字の形状（輪郭線）の位置が確定するため，文字形状の比較を行うことによって文字認識を行うことができる．画素の全景と背景とに分類することから，二値化技術は２クラス識別問題に帰着する．

初期の基本的な二値化技術では，画像全体に共通の閾値を設定して，各画素の値が閾値よりも大きいか否かによって２クラス判別を行っていた．このように画像全体に対して一つの閾値を用いる二値化技術を大域的二値化と呼ぶ．**図 2-23** に示すように，大域的二値化は画素値の明度を横軸に，出現頻度を縦軸にプロットした頻度分布において，２クラスの識別境界である閾値を求める問題に相当するため，“binarization”（二値化）ではなく“thresholding”（閾値決定）と呼ばれることがある．文書画像認識では，大津が提案した閾値決定手法が最もよく用いられており，一般に「大津二値化」（Otsu binarization）と呼ばれる[43]．これは判別分析の基準を用いて，２クラスの分離度が最も大きくなるクラス間境界を二値化閾値とする手法である．更に大津二値化が各クラスのサンプル数の偏りの影響を受けやすいという問題点を改良するための修正を行った方式を Kittler らが提案している[44]．

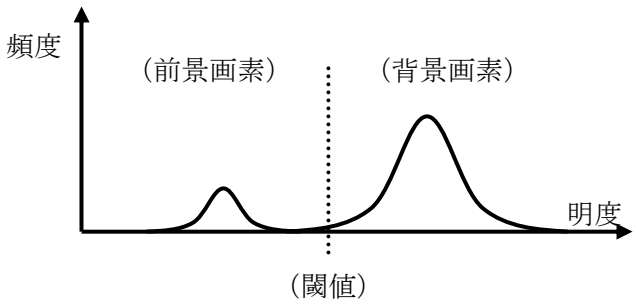


図 2-23　大局的二値化の閾値設定

大局的二値化は文書画像全体で一つの閾値を用いるため，文書の位置によって照明の明るさが異なっていたり，背景の色や明るさが場所によって異なるような場合には適切な二値化を行うことができない．そのため，画素ごとに異なった閾値を用いて二値化を行う局所的二値化が用いられることが多い．局所的二値化には様々なバリエーションがあり，二値化の目的によって最適な手法は異なる．例えば同一の多値の文字画像を複数の二値化手法を用いて取得した二値画像をそれぞれ文字認識して，文字認識の精度を基準として最適な二値化手法を選択するという研究が発表されており[45]，その中では Niblack が提案した二値化手法（Niblack 二値化）[46]にノイズ削減技術を組み合わせたものが最も良いとの結果が得られている．Niblack 二値化は極めてシンプルな手法なので，様々な用途の二値化技術を構築するための基本技術としてしばしば用いられる．

文字認識のために多値画像から二値画像を生成する場合は，二値化の対象が文字を含む画像であることを利用して二値画像の品質を高める工夫が可能である．一般に文字の画像は劣化が無い状態では文字輪郭はステップエッジであるため，文字の周辺はエッジ強度が強く，文字以外の部分はエッジ強度が弱いという傾向がある．そのような傾向を利用して，Sauvola らは文書画像をエッジが

強い領域（過渡的領域：transient region）とエッジが弱い（平坦領域：uniform region）とに分類し，それぞれの領域に適した二値化手法を使い分けるという方法を提案した[47]．Sauvola らはエッジが強い領域のための二値化を”text binarization method”と呼んでおり，文字近傍とそれ以外（背景や情景など）とを分けることによって高品質な文字画像を取得しようとしている（**図 2-24**）．

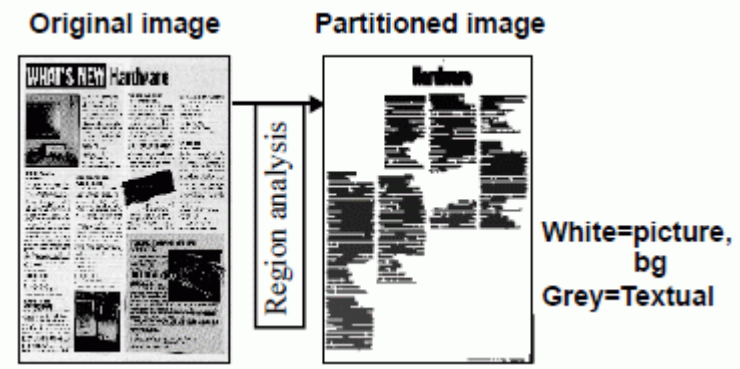


図 2-24　テキストと図・背景の領域分離例（参考文献[30]の Figure 3 より）

更に文字形状を直接的に意識した二値化手法も数多く提案されている．鎌田らは文献[48]にて，先ず大域的二値化を用いて文字近傍領域を抽出し，文字の近傍領域についてのみ局所的二値化（Niblack 二値化[46]）を用いて高品質な二値画像を生成するという手法を発表している．また同著者は文献[49]において，文字の近傍領域を Sobel 勾配値を二値化した画像から求め，文字近傍の色の分布も考慮することによって，多色の文書画像から高品質な文字の二値画像を生成するという，文献[48]の発展手法も発表している．

近年，これらの手法に類似した手法も提案されている．文献[50]では Canny エッジ抽出法を用いて文字の輪郭線を求め，その輪郭を Connected Component Analysis （CCA）で文字ごとに分離する．文字の CCA は，輪郭の内側と外側での色成分を分析して文字色と背景色を判定して，個々の文字 CCA ごとに適切な二値画像を生成する（**図 2-25**）．この手法では，文献[49]と同様にエッジ強度を用いて文字の近傍領域を求め，それぞれの文字近傍に対して局所的二値化を適用しており，技術の方向性は類似している．

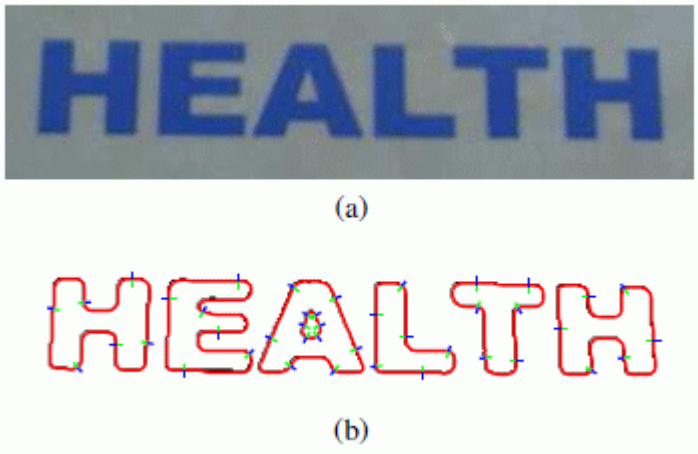


図 2-25　文字輪郭を利用した二値化（参考文献[48]の Figure 2 より）

本節で述べた複数の文字抽出用二値化に共通するのは，文字画像が輪郭に沿って強いエッジ成分を持っていることや，文字近傍と比較して背景領域は画素値の変化が少なく平坦であることなど，文字を含む文書画像の特徴を手法の設計に用いているという点である．

本研究では，文献[49]の手法をベースとして，二値化手法を更に改良することによって文字認識の精度を改善している．その改良方法も，文書画像の特徴を有効に活用したものである．その詳細は第五章で述べる．

2.5. 本研究の位置づけ

本章では，文書画像認識を構成する主な要素技術である，「レイアウト解析」，「表画像認識」，「テキスト認識」についての主な先行研究について述べた．それぞれの技術の関係は**図 2-1** に示した通りであり，更に本論文で述べる各章の位置付けを加えた図を**図 2-26** に示す．

本研究のテーマは帳票画像認識の高精度化である．既に述べたように，文書の中でも帳票はレイアウトが単純であるため，精度向上のためにレイアウト解析精度はあまり重要ではない．そこで本研究では，帳票画像認識を高精度化するため，表画像認識とテキスト認識の高精度化を目指す．その中でも特に，表画像認識の構成要素である罫線抽出技術（第 3 章）とセル抽出技術（第 4 章），そしてテキスト認識の精度向上のために，多値画像から二値の文字画像を生成する二値化技術（第 5 章）について詳しく述べる．

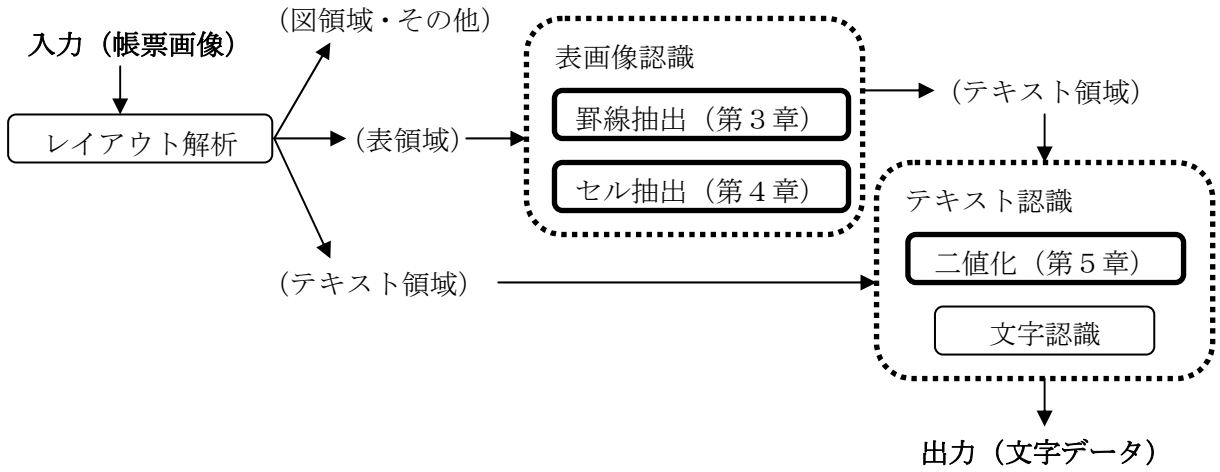


図 2-26　帳票画像認識の構成と本研究の位置付け

3. 罫線抽出技術の研究

概要

帳票画像から様々な種類の罫線を高精度に抽出する方式について述べる．近年の帳票 OCR は，デジタルカメラの普及などに伴う帳票画像の多様化に対応するため，様々な帳票書式や画質に対しても高い帳票認識精度が求められている．多彩なデザインの帳票を認識対象とするため，帳票画像から罫線を抽出する技術は，実線罫線だけでなく色や模様の領域境界による罫線（境界罫線）など様々な種類の罫線を扱う必要がある．中でも一様な模様領域（テクスチャ領域）の境界を罫線として抽出する技術は帳票画像の多様化における新たな課題である．テクスチャ境界も含む様々な罫線を抽出するため,我々はラン線分抽出とエッジ線分抽出を併用した罫線抽出方式を開発した.また，罫線抽出誤りを脱落誤りと付加誤りとに分類し，それぞれの主な原因である罫線途切れと文字列からの罫線誤抽出を改善する技術を開発した．本稿では，まず 2 種の線分抽出によって罫線の候補を求め，各罫線候補の種類や属性を局所的な画像情報により判定する罫線抽出技術について述べる．続いて，罫線脱落の原因となる罫線途切れを解消するための二値化閾値の補正技術について記す．更に，文字からの罫線誤抽出を防ぐため，前処理で文字画像を消去する技術と，後処理で罫線ノイズを除去する技術，罫線の形状を詳細に判定して誤抽出した罫線を削除する技術について述べる．最後にサンプル帳票画像を用いた評価を行い，本方式の効果と課題について考察する．

3.1. はじめに

近年，業務データの多くは作成時から電子化されているが，例えば商品と共に送付される納品書や窓口で顧客が記入する各種申込書などのように，データのやり取りに紙文書が用いられる場合も少なくない．そこで，紙文書と電子データの比較やデータ入力の効率化のため，表を含む紙文書を電子化する帳票 OCR 技術は益々必要性を増している．

これまでの主な帳票 OCR 技術は，主に実線や点線で構成された比較的単純な表のみが認識対象であった．そのため表認識の一要素である罫線抽出技術は，二値画像から黒画素のランを抽出するなどの単純な処理が中心であり[1]～[3]，罫線の種類としてはせいぜい点線[4]が考慮されるのみであった．

しかし近年の多彩なデザインのカラー帳票では，表の端での罫線の省略や （**図 3-1(a)**），異なった色で塗り潰された領域の境界を罫線とする境界罫線（**図 3-1(b)**），一様な模様(テクスチャ)で表わした領域の境界によるテクスチャ境界罫線（**図 3-1(c)**）なども多用される．特にテクスチャ境界は従来のようなラン抽出では対応できず，新たな罫線抽出技術が必要である．

帳票で表領域の分割に用いられるテクスチャのパターンは細かな模様が一様に配置されたものが多い.テクスチャには**図 3-2(a)(b)**のような不規則なものもあるが,帳票で用いられるのは**図 3-2(c)(d)**のような一様な模様やディザ,ノイズが重畳したようなパターンなど規則性があるものに限られる．テクスチャ境界を抽出する方法には,テクスチャ特徴を表す特徴量によって領域を分割する方法や,テクスチャ境界線を直接抽出するエッジ抽出法がある．前者の特徴量には濃度ヒストグラムや濃度共起行列特徴（GCLM）[5]などが知られており，後者には Canny エッジ抽出（Canny 法）[6]などがある．今回我々は，エッジ境界が比較的安定して得られると言われる Canny 法を参考にテクスチャ境界に対応することとした．Canny 法では模様等による画像の変動はランダムノイズとしてモ

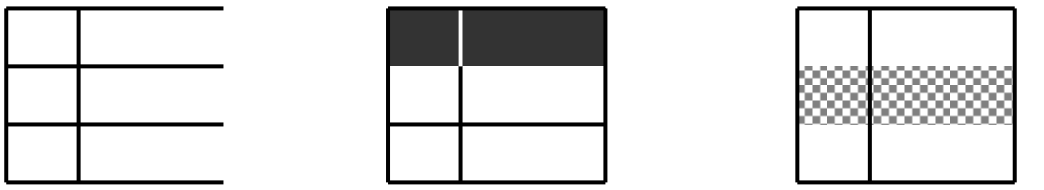
デル化され，Gaussian フィルタで平滑化される．これは帳票中のテクスチャ領域の特性（一様性）に合致する．

我々が開発した罫線抽出方式は，実線や面塗り境界（**図 3-1(b)**）を抽出するためのラン線分抽出と，テクスチャ境界を抽出するためのエッジ線分抽出（Canny 法ベース）を併用する．抽出された線分は，2 種類のマスクにより不適切な位置の線分を削除し，それぞれの線分を統合して罫線候補を生成する．その後で改めて罫線の種類や座標を詳細に判定する（**図 3-3**）．

多彩な帳票から罫線を抽出する場合，文字から罫線が誤抽出される問題に対処する必要がある．例えば**図 3-4(a)(b)**は「引」の縦棒が罫線に類似していることを示す例であり，画像が劣化している場合など両者の区別は困難である．またテクスチャ境界を抽出対象とすると，**図 3-4(c)**のように密度の高い文字列の外接線からエッジ線分が誤抽出されることもある．

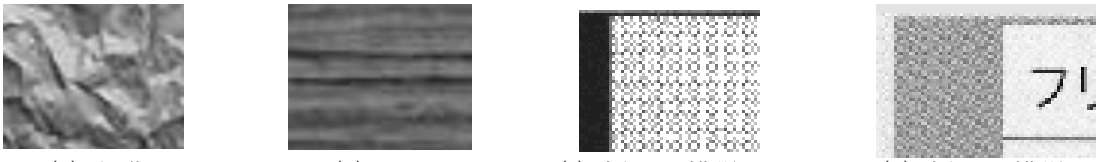
我々は，このような文字からの誤抽出を防ぐため二つの対策を行った．第一は罫線抽出の前に入力画像から文字画像を消去する文字消去処理であり，第二は罫線抽出後に不適切な罫線を削除する誤罫線除去処理である（**図 3-5**）．

以上，新たな罫線候補生成技術と誤罫線除去技術により，多彩なカラー帳票画像からテクスチャ境界を含む様々な罫線を高精度で抽出することが可能になった．本稿ではまず開発技術について記し，続いて実帳票画像を用いた評価について述べ，最後に考察を行う．



(a) 仮想罫線　(b) 境界罫線　(c) テクスチャ境界罫線

図 3-1 様々な罫線の例



(a) 紙袋　(b) 木目　(c) 帳票の模様 1　(d) 帳票の模様 2

図 3-2 テクスチャの例

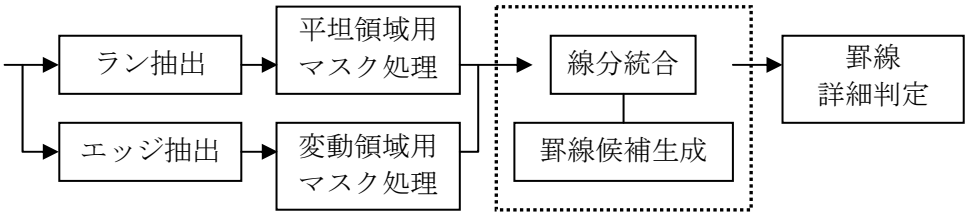
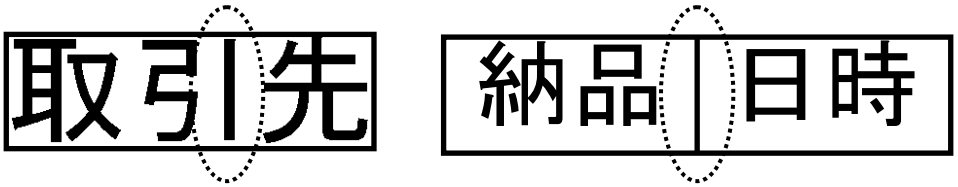


図 3-3 罫線候補生成の流れ



(a)文字中の直線成分　　(b)短い罫線

(c)画素密度の高い文字列

図 3-4 文字と罫線の区別が困難な例

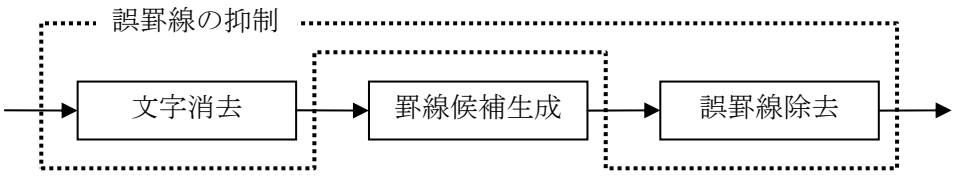


図 3-5 開発した罫線抽出方式の構成

3.2. 罫線候補生成

3.2.1. 線分抽出とマスク処理

ラン線分抽出は，入力画像を二値化して縦方向または横方向のランを求め，一定の長さ以上のものを線分として抽出する．二値化処理には背景判別付き局所的二値化を行う[7]．これは濃淡画像の局所的二値化[8]を行い，注目画素の近傍領域において白－黒画素の濃度差が一定の閾値を越えた場合にのみ二値化の結果を用いる（それ以外は白画素）処理である．これにより面塗り領域の境界も実線罫線と同様に抽出できる．

エッジ線分抽出は Canny 法と同じ手順であり（**図 3-6**），カラー画像を対象とするため明度勾配の代わりに色勾配を用いる．色勾配生成は Sobel フィルタで隣接画素間の明度差分の計算を RGB 空間上の距離に置き換えたものである．ただし簡単のため RGB 要素間の相関は考慮していない（スカラ勾配[9]）．通常 Canny 法では勾配方向に沿ってエッジ点を探索するが，今回は垂直／水平の罫線が対象なので，縦横いずれか片方向の勾配値の極大点を求め，方向別のエッジ抽出を行う．

次にマスク処理により不適切な位置の線分を削除する．マスク処理は罫線が存在しないと判断された領域にマスクをかけ，不要な罫線候補が生成されないようにする処理である．マスクは平坦領域用マスクと変動領域用マスクの２種類が用いられる．

平坦領域用マスクは元画像から局所的な分散値（Σ値）を求め，Σ値が大きな部分にマスクを設定して線分抽出を抑制する．画像の局所変動が少ない領域を残すので実線や面塗り領域を含み，テクスチャ領域は含まない．一方，変動領域用マスクは元画像のΣ値を計算した後，抽出したい罫線の方向とは直交した方向にΣ値の差分を求める．差分値が小さな領域はΣ値が一様な領域だと思われるので，テクスチャ境界は存在しないと判断できる．そこで，このΣ値の差分が小さな領域に変動領域用マスクを設定する（**図 3-7**）．

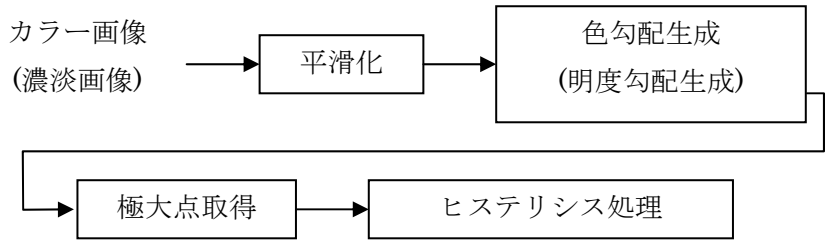


図 3-6 エッジ抽出の流れ

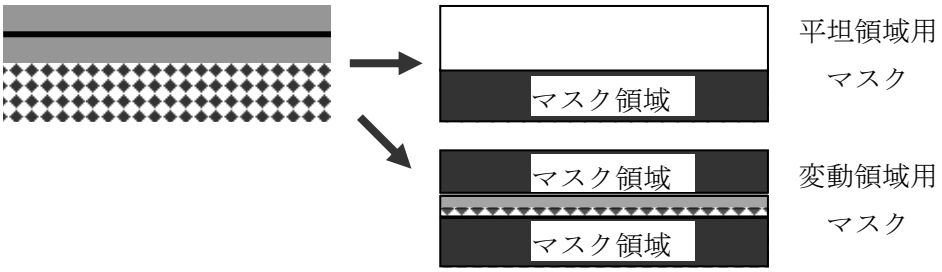


図 3-7 マスク領域

3.2.2. ラン線分とエッジ線分の統合

ラン抽出とエッジ抽出で得られた線分画素は一つの画像に統合し，各画素にラン線分画素，エッジ線分画素，またはその両方のフラグを振る．この統合画像をラベリングして罫線候補を求め，それぞれの罫線候補の種類を判定する．

罫線の種類判定の原理を**図 3-8** に示す．実線から得られた罫線候補はラン線分に 2 本のエッジ線分が隣接しており，境界罫線ではそれぞれ 1 本ずつ，テクスチャ境界罫線ではエッジ線分のみである．これ以外は仮に実線罫線とする．

ここで一定の閾値よりも短い罫線候補は削除するが，その閾値は罫線候補の種類に応じて決定する．例えばテクスチャ境界罫線は実線罫線に比べて長い傾向があるので，短いテクスチャ罫線候補は誤りである可能性が高い．そこで実線の場合よりもテクスチャ境界の方が大きな長さ閾値を用いる．

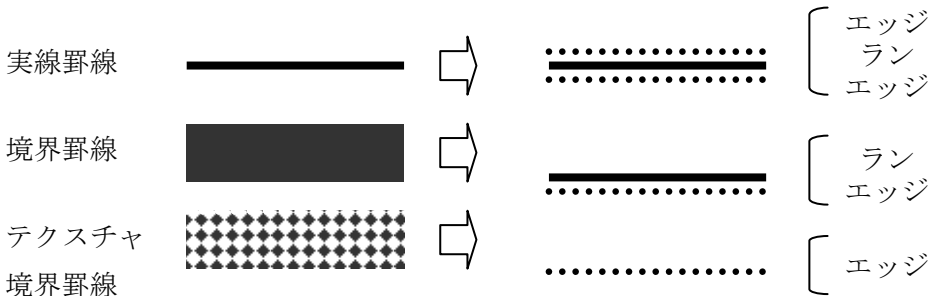


図 3-8 線分抽出パターン

3.2.3. 罫線詳細判定

罫線候補ごとに罫線の座標や種類（属性）の詳細な判定を行う．罫線候補には異なった種類の罫線が混在している場合があるので,縦横の罫線候補が直交する位置で罫線候補を部分罫線に分割し，部分罫線ごとに属性判定する（**図 3-9**）.

罫線属性の判定は，部分罫線候補の近傍領域の二値画像を用いる．二値画像を罫線の長さ方向に射影して画素の頻度と変化率（画素値が変化した回数の割合）の分布を求め，この画素頻度と変化率の分布により罫線の種類を判定する.

例えば実線罫線であれば，変化率が一定値以下で画素頻度が大きい領域と画素頻度が小さな領域とが交互に現れ，罫線の近傍に画素頻度が大きく変わる点を二つ持つ．テクスチャ境界の場合は変化率が大きな領域と小さな領域が隣接する．実線罫線の判定例を**図 3-10** に示す.

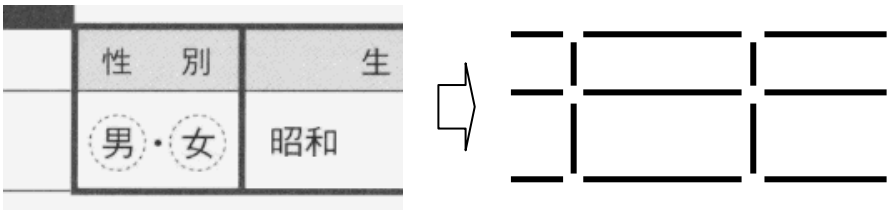
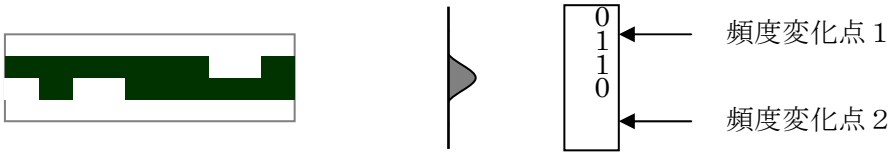


図 3-9 直交罫線での部分罫線への分割



(a)部分罫線画像　　(b)画素頻度分布　　(c)頻度変化点

図 3-10 部分二値画像による実線判定

3.3. 途切れによる罫線脱落の改善

3.3.1. ラン罫線抽出の概要

開発した罫線抽出方式（**図 3-3**）では，実線と面塗り境界をラン罫線抽出で求め，テクスチャ境界をエッジ罫線抽出で求める．ここではラン罫線抽出における途切れの問題について検討する.

ラン罫線抽出は，画素が連続した領域または境界線を抽出するものである．ラン罫線のうち，実線罫線の条件は下記（①～⑤）のように記述できる.

①罫線画素は同一（類似）色の画素で構成される

②罫線画素が縦横方向に直線的に連続した領域（＝ラン）である

③ランの長さは閾値（例えば 20pixel）以上

④領域の幅（罫線幅）は閾値（例えば 5pixel）未満

⑤同じ罫線内では領域の幅はほぼ一定

一方，面塗り境界罫線の条件は下記（A～C）の通りである.

A) 同一（類似）色の画素が連続して存在する領域の直線的な境界線（＝エッジ）である
B) エッジの長さは閾値（例えば 20pixel）以上
C) 領域の幅は閾値（例えば 5pixel）以上

我々の方式では，これら 2 種類の罫線を局所的二値化（Niblack 二値化[8]）による二値画像から閾値より長いランを抽出することによって同時に抽出している．

Niblack 二値化は対象画素の周辺 $w\times w$ の局所領域（w は窓サイズ）における明度平均 m と標準偏差 δ より，二値化閾値 T を以下の**式 3-1** で求めるものである．

$$T = m + k\delta \qquad \text{(式 3-1)}$$

図 3-11 は実線と領域境界を w=5 の局所領域で二値化した例であり，各画像の下のグラフは局所領域内の画素の頻度分布を示す（縦軸が頻度，横軸が明度）．

実線の二値化の例は局所領域内で罫線画素と背景が閾値 T で分離され，線幅 2 の二値画像が得られている様子を示す．一方，領域境界の二値化の例は局所領域が背景を含まない場合を示しており，中心画素は背景（白画素）と見なされる．その結果，領域境界は罫線幅(w-1)/2 の実線と同じように二値化される．なお実線の線幅が窓サイズ 5（=w）以上となると罫線内に白画素が生ずる（太い罫線が面塗り領域と見なされる）．つまり Niblack 二値化の結果から黒画素のランを抽出するという処理は，条件④と条件 C)の閾値が w に固定された出力が得られることになる．

以上が Niblack 二値化の基本的な動作説明だが，実用上は更にノイズ除去機能が実装されることが多い[13]．これは**式 3-1** では背景に細かなノイズが生じやすいという問題を回避するためである．Niblack 二値化では標準偏差 σ の項により黒画素が出にくい方向に閾値がシフトするため，**図 3-12** のように境界付近では黒画素が出力され，平坦な領域では何も出力されない．しかし**図 3-13** のように平坦な領域に微小な凹凸があるとノイズが生ずることがある．これは平坦な領域では標準偏差 σ が小さく，ノイズ抑制の効果が得にくいためである．

背景ノイズを削減するため，Sauvola ら[14]は標準偏差項に明度値を加えた補正式を提案しているが，我々はもっと単純に，局所領域内での前景・背景画素の群内平均の差分が一定値を下回った場合に平坦領域と見なすことによってノイズを抑制している（＝背景判別）．なお文献[7]によれば Eikvil らの手法[15]にも同様の判定が用いられていることが記されている．

なお，我々の方式では実線罫線の幅の制限（条件④）に対応するため，罫線の方向（縦横）ごとに幅を制限するフィルタを適用している．

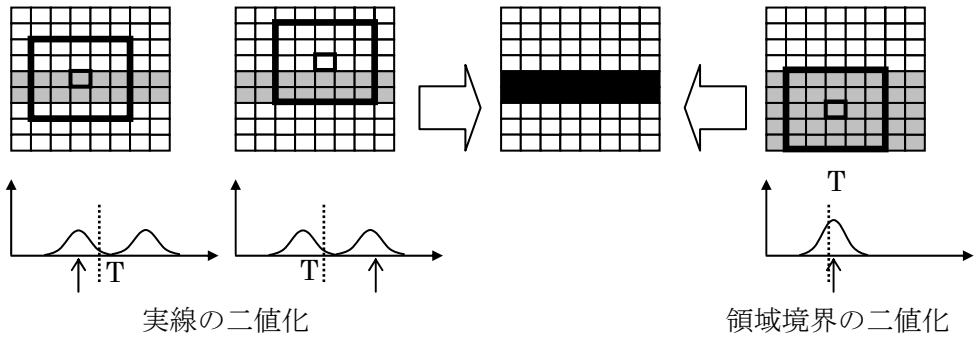
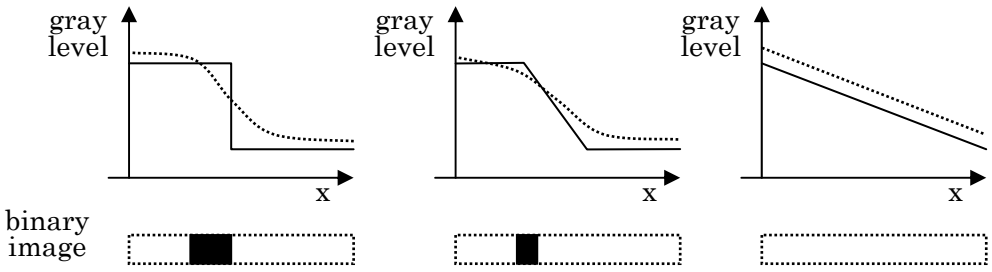


図 3-11　実線と境界罫線の局所的二値化



(a) ステップ境界　　(b) 勾配境界　　(c) 平坦勾配

図 3-12　境界の種類と閾値の比較

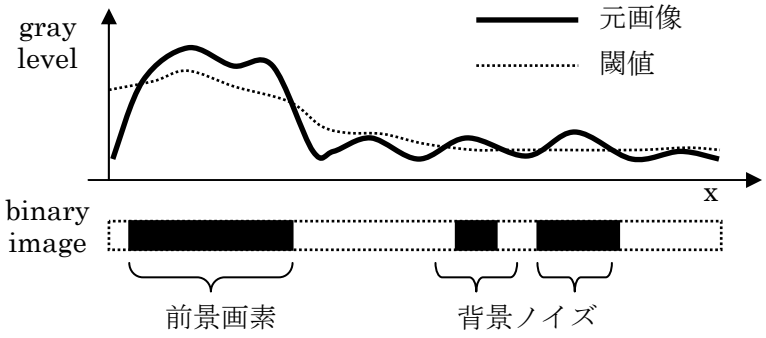


図 3-13　元画像と閾値

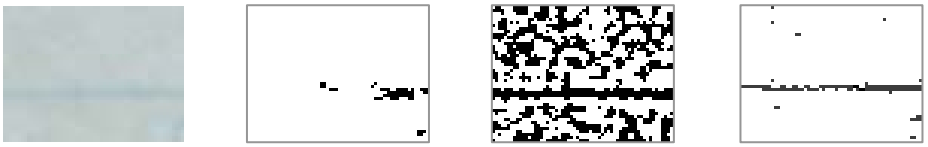
3.3.2. 背景判別の補正による途切れ解消

ラン罫線が途切れる原因の第一は，罫線画素と背景の明度差が小さい（コントラストが低い）ことである．

薄い罫線（**図 3-14(a)**）が途切れた例を**図 3-14(b)**に示す．途切れ位置の局所領域を見ると，例えば**図 3-15** のように罫線画素は周囲の背景よりも小さな値であり，Niblack 閾値で罫線画素（前景）と背景画素の区別は可能である．しかし前景・背景の群内平均の差分が小さいため平坦領域と判断され，白画素が出力されている．

例えば**図 3-15** の値では，前景平均=213.2，背景平均=203.3，差分 d=9.9 である．通常，背景判別閾値（*dmin*）は 10～15 程度とするので $d < dmin$ となり平坦と判定される．*dmin* を小さくすれば途切れは解消されるが，罫線以外からのノイズが増加するので，単に閾値を小さくすれば良いというわけではない（**図 3-14(c)**）．

そこで，罫線方向ごとに**図 3-16** のようなマスクを設定し，領域 A の白画素数と領域 B の黒画素数に基づいて背景判別閾値 *dmin* の補正を行う．具体的には，領域 B の黒画素数が一定値（*bcnt*）以上，領域 A のいずれかで白画素数が一定値（*wcnt*）以上，という条件を満たした場合に，背景判別閾値 *dmin* の値を変更する（例えば *bcnt*=*wcnt*=4）．これにより，**図 3-17(a)～(c)**のようなパターンでは *dmin* は補正され，**図 3-17(d)**のようにランダムな場合には *dmin* は変更されない．これにより**図 3-14(c)**のノイズは**図 3-14(d)**のように抑制することができる．



(a)元画像　(b) dmin=10　(c) dmin=2　(d) 補正

図 3-14　薄い罫線の途切れ

213	212	213	213	211
202	199	204	204	201
204	202	206	206	205
213	211	212	213	212
215	215	215	215	215

図 3-15 途切れ周辺の画素

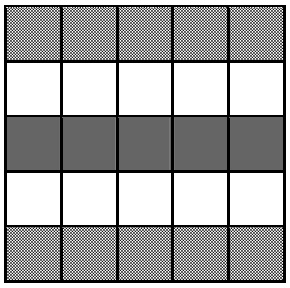
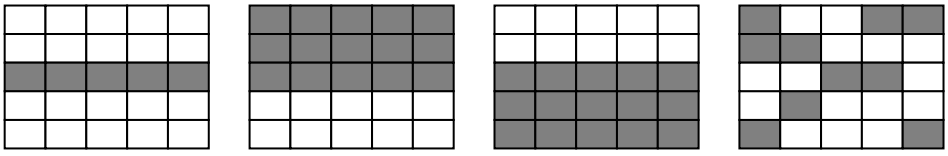


図 3-16 罫線判別マスク



(a) 実線　(b) 境界-1　(c) 境界-2　(d) ノイズ

図 3-17　背景判別補正の適用対象

3.3.3. 二値化閾値の補正

ラン罫線途切れのもう一つの原因は，罫線の近くに別の濃い画素が存在することによる二値化閾値の変動である．**図 3-18(a)**では二値化判定を行う局所領域の中に罫線画素と背景画素が存在し，頻度分布の中間点付近に閾値が得られている．一方，**図 3-18(b)**では局所領域内に近接画素を含むため頻度分布が三つの中心を持ち，近接画素とそれ以外を分ける閾値が得られてしまっている．

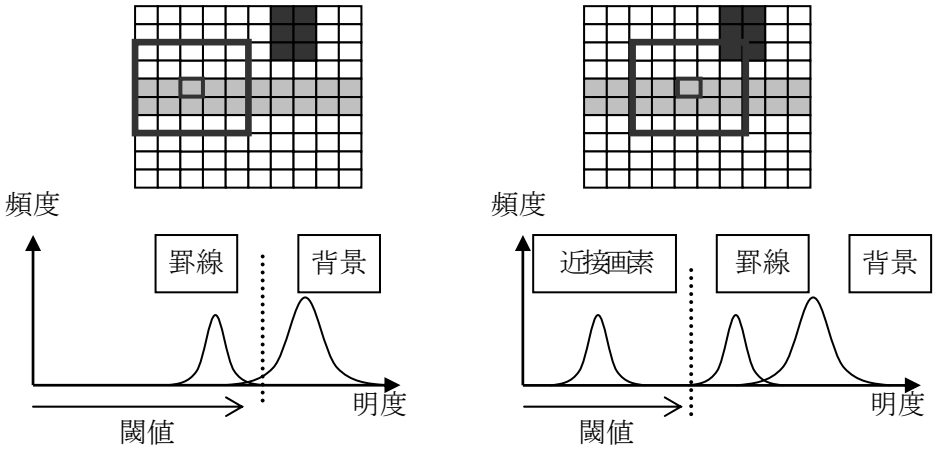
近接画素の影響を避けるためには，近接画素を含まない局所領域から得られた閾値を用いれば良い．例えば**図 3-19** のように局所領域を右にシフトすると近接画素を含まない局所領域が得られる

ので，局所領域を上下左右にシフトし，それぞれの領域で求めた閾値の最大値を新たな閾値とすれば途切れは解消できる．なお，以上の処理はあらかじめ全ての座標での閾値を求めておき，座標ごとに周辺の閾値の最大値を求める操作と等価である．

閾値補正により罫線途切れは解消できるが（**図 3-20(b)⇒(c)**），一方で**図 3-20(d)～(f)**のように中間画素が黒画素となって付加誤りの原因となるという問題がある．これは，閾値補正によって閾値が変わる画素は罫線途切れの画素とは限らず，あくまで罫線途切れが生じている「可能性がある」画素に過ぎないためである．実際，局所領域だけを見ても，罫線の近くに濃い近接画素があるのか，濃い画素がボケて中間値が生じているのかは判定できない．

罫線補正によって白画素を黒画素に変更するか否か，すなわち閾値補正を適用するか否かは，局所領域の外部を見なければ判断できない．そこで 3.3.1 節で述べた罫線の条件を参照し，条件①②および条件 A)に基づいて罫線画素の判定を行う．具体的には閾値補正によって白画素が黒画素になる部分は保留画素として，**図 3-21(a)**のような三値化画像を中間的に生成する．続いて罫線モデルの定義に基づいて直線的な黒画素の並びを抽出し，これを決定画素とする．最後に一定長以上の決定画素の延長上にある保留画素を黒画素とする．

縦罫線抽出用の二値画像（太さ制限フィルタ適用済）での閾値補正の効果を示す．**図 3-22(b)**での途切れは**図 3-22(c)**で解消されているが，文字が必要以上に膨らんでいる．保留画素を利用した補正を行うことで不要な膨張をある程度は抑えることができる（**図 3-22(d)**）．



(a)近接画素を含まない　　(b)近接画素を含む

図 3-18　近くに濃い画素が存在する場合

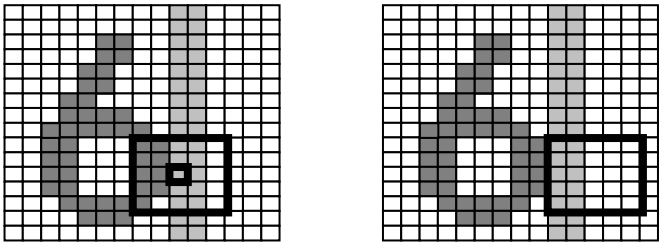
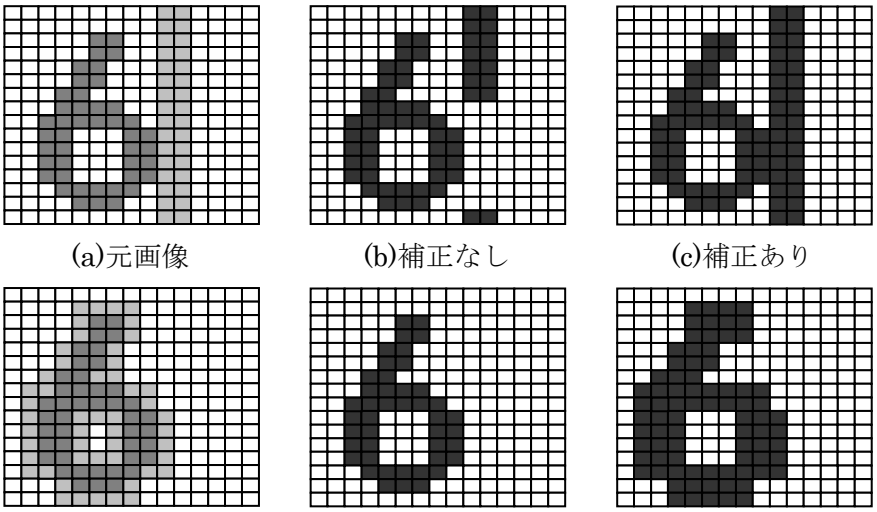


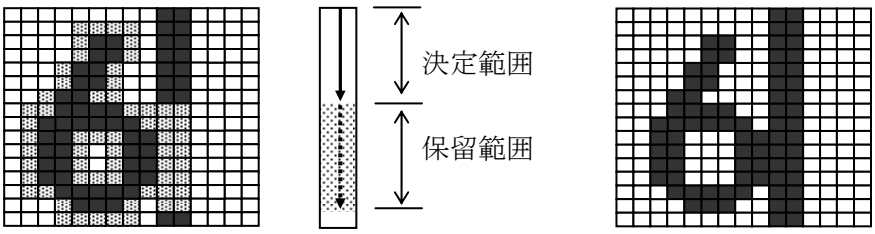
図 3-19　近接画素を避けた周辺領域



(a)元画像　(b)補正なし　(c)補正あり

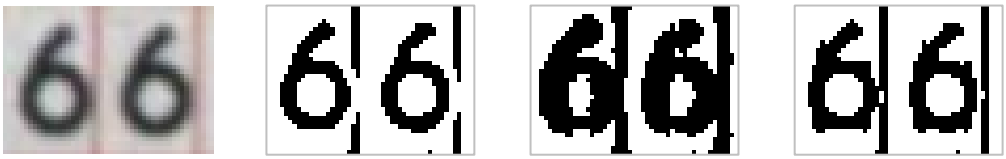
(d)元画像　(e)補正なし　(f)補正あり

図 3-20　閾値補正の効果と問題点



(a)三値化画像　(b)保留画素の補間　(c)補正後

図 3-21　保留画素の補間



(a)元画像　(b)途切れ　(c)閾値補正　(d)保留画素

図 3-22　閾値補正の効果

3.4.　文字からの付加誤りの改善

3.4.1.　文字消去処理

文字から罫線候補が誤抽出される誤りを改善する方法として，前処理にて文字画像をあらかじめ消去する方法と，後処理で誤り罫線を除去する方法とを実装した．

文字消去処理は，入力画像から文字を構成する画素を抜き出し，周辺画素から求めた背景色で塗り潰す．文字を完全に見えなくするのは困難だが，罫線が誤抽出されない程度に色の差を少なくするのは可能である．文字の消去は文字領域判定と文字画像消去の 2 段階で行われる（**図 3-23**）．

文字領域判定は，二値化画像をラベリングして，各ラベルのサイズや形状に基づいて文字を構成

する画素か否かを判定する．そこで罫線と接触した文字を分離するため，まず縦横の長いランを二値画像から消去する（**図 3-24(a)**）．続いて画素を 8 方向の連結画素でラベリングし，ラベルごとに罫線，文字，それ以外の判定を行う．

ここで誤って短い罫線を消してしまわないため，ラベル形状が細長いものや，ラベル中に長いランが存在するものは罫線と判定するようにする．したがって文字中に罫線に類似した直線を持つものは消去されないが，ここでの文字消去処理はあくまで誤抽出の削減が目的であり，確実に文字だと判定できたラベルのみを消去すれば良い．

文字画像消去処理では，文字と判定された領域について文字画像を構成する画素の周辺色を取得し，その色で文字の画素を塗り潰す．続いて色の不連続の影響を軽減するため塗り潰した画素を近傍領域でぼかす（**図 3-24(b)**）．

周辺色は文字周辺の画素（周辺画素）の値の平均値を求めるが，背景以外の画素による影響を避けるため，周辺画素の平均値から一定の閾値より値が離れた画素を無視してもう一度平均を取得する[10]．今回は閾値を固定とし，RGB 値のいずれかが一定値よりも離れている画素を無視するようにした．

文字画素に周辺色をセットしたら Gaussian フィルタで平滑化する．これにより文字領域の外接矩形内の画素をぼかし，消去した文字画像と背景との間に境界線ができるのを防ぐ．

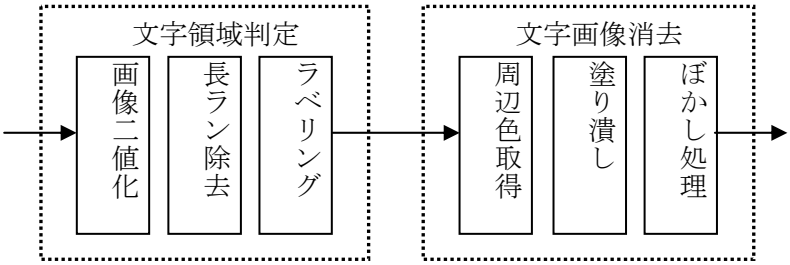
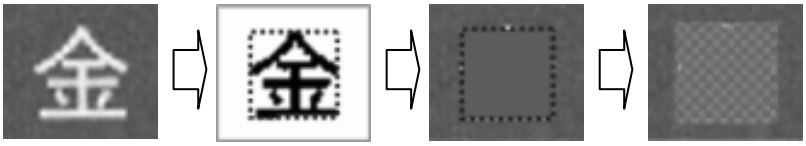


図 3-23 文字消去処理の流れ

(a) 罫線接触文字の分離



(b) 文字画像の消去

図 3-24 文字消去処理の例

3.4.2. 文字からの誤り罫線の除去

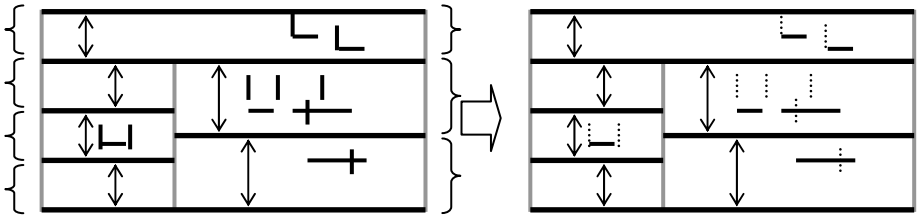
文字から誤って抽出される罫線の多くは文字消去処理で削除されるが，完全に削除できるわけではない．そこで後処理にて不要な罫線候補を除去する．

誤り罫線除去の原理を**図 3-25** に示す．まず抽出された罫線候補の中で，一定の長さを越えた横罫線を確定罫線とみなす．次に確定罫線の隣り合ったペアを求め（**図 3-25(a)(b)**），罫線ペアに挟まれた矩形領域を設定する．この領域ごとに罫線長の最小値（長さ閾値）を設定し，閾値より短い縦罫線を削除する．

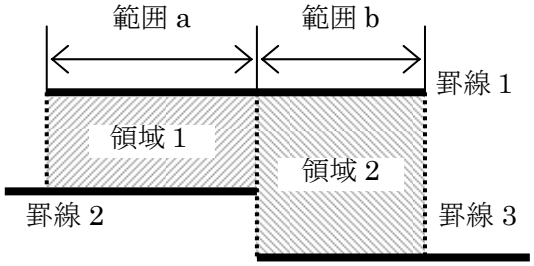
長さ閾値は領域の高さよりも若干小さな値を閾値として設定する（ただし既定の最大値を越えない値）．領域内の文字サイズに基づいて長さ閾値を設定するのであれば，例えば領域内の罫線候補の長さの分布に基づいて文字の平均サイズを求め，推定文字サイズに基づいて閾値を設定するなどの方法も考えられるが，今回は簡易な方法を用いている．

横罫線のペアに挟まれた領域は**図 3-25(c)**に示すように設定する．まず罫線 1 に最も近い罫線 2 との間で領域 1 を求め，領域 1 に対応する範囲 a を既使用範囲とする．ここでまだ罫線ペアに対応していない範囲 b が存在すれば，更に対応する罫線ペアを探し，罫線 3 との間で領域 2 を構成する．以上を未使用範囲が無くなるまで続けて誤罫線除去領域を求める．

一般に表内の文字列は横書きが多く，複数の文字を跨いだ誤り罫線はほとんどが横罫線である．そのため文字サイズを意識した長さ閾値による誤罫線除去は縦罫線候補がより効果的である．最終的に表を構成しない罫線は削除できるため，縦罫線の誤りが減ればそれに伴い横罫線の誤りも減少する．予備実験によれば横罫線の誤りを除去しても効果に差が出なかったので，現在は縦罫線のみを削除対象としている．



(a)横罫線のペアを求める　　(b)短い縦罫線を除去



(c)誤罫線除去領域の設定

図 3-25 罫線ペアに基づく誤罫線の除去

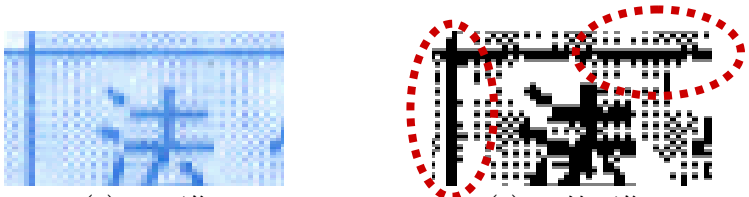
3.4.3. 罫線の形状判定

既開発の方式に加え，罫線の直線性に着目して誤抽出候補か否かを判定する方式を開発した．判定基準は罫線の側面が凸凹ならば罫線ではないと判定するもので，罫線モデルの条件②と A)に該当する．これは従来技術[16]でも用いていた判定基準である．今回は同じ条件を別の手法により実現した．従来技術[16]の方法は，罫線の側面で白画素と黒画素が隣接している座標点を追跡し，方向が変化した回数によって凸凹度を測るものであった．しかし罫線の背景に模様やノイズが多い場合，罫線の側面がその影響で凸凹になることがあるため，一画素ごとに局所的な方向を見る方法では凸凹度の判定が不安定になる（**図 3-26**）．

開発方式では，まず罫線の側面にエッジ点の候補範囲を設定し，その範囲内の座標点を罫線方向に探索する．探索パスは直線であるほど良く，パスの方向が斜めであればペナルティを加算する．探索パスが途切れた場合はそこを線分の終点として，次の座標点から探索を再開する．以上により，罫線候補の側面から複数の線分が抽出される．

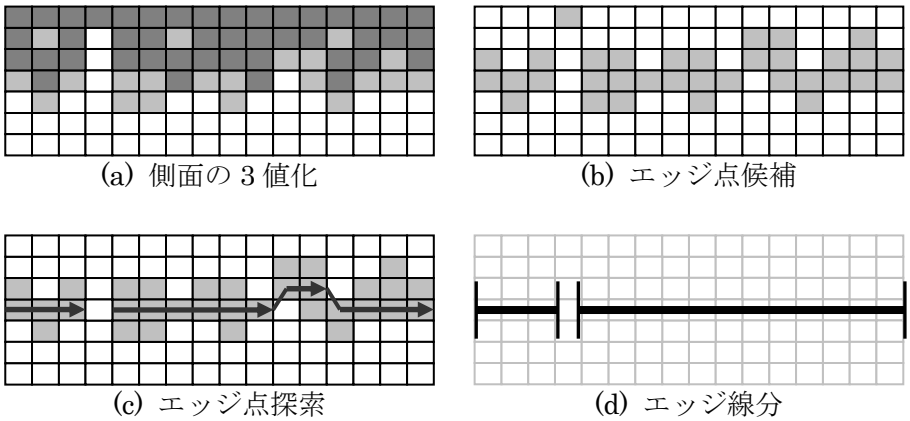
具体的な処理の流れを**図 3-27** に示す．まず罫線候補の側面を画素の明度値によって三値化する（**図 3-27(a)**）．これは黒画素／中間画素／白画素に分けるもので，罫線方向と直交する方向に，黒画素の端点から中間画素の端点までの範囲がエッジ点の候補となる（**図 3-27(b)**）．三値化の閾値は二値化閾値 T に対して $T \pm w$（w は定数）とする．次にエッジ点の候補を DP 探索してエッジ線分を求める（**図 3-27(c)(d)**）．DP の遷移パスは罫線と直交方向に±1 画素の移動のみ許すもので，斜めの遷移の数をペナルティとして最適パスを求める．

エッジ線分は一定の閾値より短いものを削除し，残った線分の長さの総計が罫線候補長の 70%に満たない場合，罫線候補をノイズとして削除する．その結果，**図 3-28** のように不正な罫線候補が削減できる．



(a) 元画像　(b) 二値画像

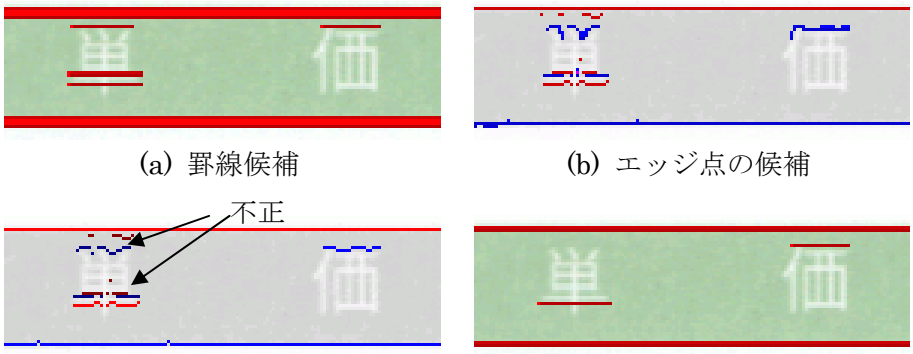
図 3-26　罫線側面に凹凸が生ずる例



(a) 側面の 3 値化　(b) エッジ点候補

(c) エッジ点探索　(d) エッジ線分

図 3-27　罫線の側面エッジ点探索



(a) 罫線候補　(b) エッジ点の候補

(c)エッジ点探索　(d) 補正後の罫線候補

図 3-28　側面エッジ形状による誤抽出罫線の削除

3.5.　評価実験

3.5.1.　開発技術のポイント

開発方式の有効性を確かめるために，実帳票画像を用いた評価実験を行なった．本章では複数の開発技術について記述したので，開発方式の有効性を評価するため，まず開発技術の中で評価すべきポイントを整理する．

開発した罫線抽出技術は，従来技術ではラン抽出のみを用いて罫線を抽出していたのに対して，エッジ情報も併用することによって，テクスチャ領域などの境界罫線も同時に抽出することが可能になったというものである．従って，評価すべき第一のポイントは「ラン抽出とエッジ抽出の併用」の効果の検証である．

続いて，実帳票では文字パターンなどのような罫線以外からの罫線の誤抽出が問題となることから，罫線の付加誤りを防ぐ技術を開発した．具体的には，文字領域を塗り潰すことによって文字画像を消去する（「文字消去」）技術と，長い罫線のペアに挟まれた短い罫線を削除する（「罫線ペアによる誤罫線除去」）技術を開発した．評価の第二のポイントは，これらのノイズ除去技術の有効性の検証である．

更に，罫線の形状を詳細に分析することによって，更なる罫線抽出の高精度化を実現する技術を開発した．具体的には，罫線抽出用二値化の閾値補正と，罫線側面の凹凸判定による罫線形状の評価である．これらを含めて「罫線の詳細判定」として，その有効性を評価する．

以上で述べた評価ポイントに注目して，実帳票画像を用いた罫線抽出の精度比較を行なった．

3.5.2.　精度評価実験

精度評価に用いた評価画像は，一般カラー帳票 50 画像と，伝票類（納品書等）からなる 41 画像の 2 セットである．第一のセットの罫線数は 3289 本（一画像あたり約 65.8 本），第二のセットは 1813 本（一画像あたり約 44.2 本）である．テクスチャ境界罫線を含む画像は 14 画像あり，122 本が含まれている．評価に用いた画像の一部を**図 3-29** に示す．

罫線抽出精度の評価指標には再現率と適合率を用いた．再現率は正解罫線の中で抽出に成功した罫線の割合であり，適合率は抽出された罫線候補の中での正解罫線の割合である．抽出した罫線の始点・終点の座標と，正解罫線の始点・終点の座標を比較して，両者の座標値の差が罫線の長辺方向で 20pixel（約 2.54mm）以内であり，短辺方向で 10pixel（約 1.27mm）以内であれば同一の位

置に抽出できたもの（抽出成功）と数える．

罫線抽出精度の評価値を**表 3-1** に示す．最初の評価ポイントである「ラン抽出とエッジ抽出の併用」の評価については，表中の(a)と(b)を比較すれば良い．再現率は同等か若干向上しており，より多くの正解罫線が抽出できていることが分かる．一方，適合率が大幅に下がっていることから，エッジ抽出を併用することによって，罫線以外からの誤抽出が増加したということも分かる．したがって，ラン抽出とエッジ抽出を併用する場合には，罫線の誤抽出を改善する方策が必須であることが，この結果から読み取れる．ただし一般カラー帳票での再現率が若干低下しているが，これは正しい罫線と誤抽出罫線が結合して一体化してしまうために，抽出された罫線は増えても正解数が減少したためだろうと解釈できる．

続いて，ノイズ除去技術の有効性について，(b)と(c)～(e)を比較する．(b)と(c)を比較すると，文字消去が罫線抽出の精度向上に大きく貢献することが分かる．一方，(b)と(d)を比較すると，再現率が同等か若干下がっているのに対して，適合率が上がっていることから，罫線ペアによる誤罫線除去は，誤抽出の削減には効果があるものの，同時に正解罫線の一部も除去しまっていることが読み取れる．しかしながら，文字消去と罫線ペアによろ誤罫線除去の両方を適用した(e)を見ると，再現率，適合率とも大幅に改善している．特に適合率が(c)，(d)のいずれよりも高いことから，罫線以外からの誤抽出を選択的に削除できていることが分かる．3.4.2 節で述べたように，文字消去処理では完全には消せなかった文字画像から誤抽出された罫線を，罫線ペアを用いた誤罫線除去処理が削除しているのだと考えれば，両者の相性が良いものと解釈することができる．

最後に，(f)について見ると，再現率，適合率ともこれまでの比較対象の中でほぼ最高の値を示していることが分かる．二値化閾値の補正によって罫線の途切れが改善し，再現率が向上したことに加え，罫線の形状を詳細に判定したことにより，文字列などからの罫線の誤抽出を的確に除去できたものと推察する．

表 3-1 罫線抽出精度の比較

	一般カラー帳票		取引伝票	
	再現率	適合率	再現率	適合率
(a) ラン抽出のみ	86.66%	87.59%	85.32%	81.64%
(b) ラン抽出＋エッジ抽出	86.01%	60.05%	90.52%	39.29%
(c) (b)＋文字消去	91.54%	73.88%	89.89%	48.64%
(d) (b)＋罫線ペアによる誤罫線除去	85.59%	68.82%	90.29%	61.59%
(e) (d)＋文字消去＋誤罫線除去	91.00%	82.62%	89.95%	75.43%
(f) (e)＋罫線の詳細判定	92.28%	87.58%	91.47%	89.14%

3.6. まとめ

帳票画像から実線罫線と境界罫線と共にテクスチャ境界罫線を抽出できる罫線抽出方式を開発した．また抽出罫線の種類が増えたことに伴う文字からの罫線誤抽出の増加を防ぐ方式も開発した．開発方式は，ラン抽出にエッジ抽出を併用することにより，一様な模様によるテクスチャ境界の抽出を可能とする．これにより，評価画像を用いた罫線抽出精度評価によれば，ラン抽出のみを用い

た場合よりも多くの罫線を抽出することができるようになった.

一方で，多くの罫線を抽出することにより，罫線以外からの誤抽出も増加することも確認でき，罫線誤抽出を抑制する技術が必要であることも確認できた．それに対して，文字消去，罫線ペアによる誤罫線除去，そして罫線形状判定による誤罫線除去などの技術を開発して，誤抽出罫線を高精度に除去する技術を実現した. また, 二値画像の途切れによる罫線脱落を改善する技術も開発した. 以上の開発技術を組み合わせることにより，記入済み帳票などの文字が多い帳票画像や，テクスチャ境界を含む多様な罫線を含む帳票画像であっても，罫線抽出の精度をより高めることができ，表認識の実用性向上に貢献することができる.

4. セル抽出技術の研究

概要

未知書式の帳票画像から表を構成するセル領域を抽出する方式を提案する．近年，帳票書式の多様化や複雑化により表罫線の抽出はより困難になっている．しかし，従来のセル抽出方式は罫線抽出誤りの影響を受けやすいという問題を持っていた．提案方式は罫線が交差する交点を特徴点として用い，交点を順に追跡して閉領域を求めることによってセル領域を抽出する．また，罫線抽出結果の曖昧さを尤度で表わし，セル候補の最適な組合せを求めることによって全体最適なセル領域の集合を求める．これにより罫線抽出誤りの影響が局所的に限定され，帳票画像の複雑化や劣化に対して頑強なセル抽出が実現できる．44 種類の帳票画像による評価実験によると，適合率を落とさずに，従来方式で 80.21%であったセル抽出精度（再現率）を 83.39%に改善でき，提案方式の有効性が確認できた．

4.1. はじめに

帳票画像認識の主要な目的は，帳票に書かれた文字列を自動認識することにより，データエントリ工数を削減することにある．帳票認識の結果が十分に信頼できる場合，認識結果をそのまま利用できるため大きな工数削減効果が期待できる．しかし，精度が不十分な場合には誤り訂正の手間により，工数削減の効果が限られる．そのため，これまでに実用化された帳票画像認識システムの多くは既知の書式に限定して帳票画像認識の精度を高めるものであった．その一例として郵便振替用紙の認識のような単一書式の帳票認識システムがある．また，帳票照合を組み込むことにより，事前に登録された複数の書式について，帳票認識を可能にする技術も提案されている[1]～[9].

一方で，書式が未知の帳票画像を認識する技術も必要とされている．その一例に銀行の取引伝票の認識がある．銀行は様々な顧客との間に取引関係があり，顧客が定義した書式の伝票を扱うことも多く，書式の種類が数千～数万に及ぶこともある．そのような場合，書式によらない帳票画像認識技術はデータ入力の効率化の大きな助けになる[10], [11]. また, 帳票照合技術を用いる場合でも，照合する書式情報のシステム登録作業を効率化するために未知書式の帳票認識技術が活用できる．更に近年では，個人向けスキャナやデジタルカメラの普及によって様々な種類の文書が電子化されるようになってきており，例えば雑誌や新聞記事のように，個々に書式が異なる文書画像から文字列データを抽出して活用したいという要望も増加している．

帳票文書では，認識したい文字列データのほとんどは表の項目領域（セル領域）の内部に書かれており，文字列を高精度に認識するためにはセル領域の同定が欠かせない．そのため未知書式の帳票画像認識では，表画像からのセル領域認識（セル抽出）技術の高精度化が重要な課題である．
国内で用いられる帳票の大半は表領域が罫線で区切られた罫線表であるため，これまでも罫線情報からセル抽出を行うための技術が数多く発表されている[1]～[8], [12]～[16]．一方，表項目の区切りが明示的に記されていない無罫線表を認識する方式も提案されている[11], [17]が，そのような帳票は非常に少ないため，罫線情報を用いない表認識技術は本稿の対象外とする．

表画像からセル領域を抽出する方式では，画像から矩形領域を直接抽出する方式や，表領域を罫線で分割する方式，罫線が交差する交点を利用する方式が報告されている．画像から矩形領域を直接抽出する方式では，Geometric Hashing によって矩形の中心座標を検出する方式[2]や，エッジ画

像から一般化 Hough 変換によって矩形抽出を行う方式[15], 画像から枠の左上角をフィルタによって抽出し，抽出した角から枠線を辿る方式[3]などが提案されている．表領域を罫線で分割する方式では，表を縦横の平行な罫線で順に分割する方式（**図 4-1**）がよく使われている[12]～[14]．交点を用いる方式では，罫線が交差する交点の形状を 16 種に分類して抽出し，四つの交点の組合せによって矩形領域を抽出する方式が提案されている[16]．交点を抽出する方式には，罫線の交わり方から交点形状を分類する方式[6], [16]や，画像の局所領域の特徴から交点の位置と種類を直接抽出する方式[5]がある．

これらの従来研究は認識対象として比較的単純な表構造を想定しており，複雑な書式の帳票や文字が多い帳票の認識は得意ではない．ところが，近年では PC などで簡単に帳票を作成できることから，利用される帳票の多様化が進んでいる[18]．先に述べたセル抽出方式についてみると，画像から矩形領域を直接抽出する方式には，背景模様や文字列画像から誤った矩形領域が抽出されやすいという問題がある．平行な罫線で表を分割する方式は，L 字型のセルを抽出することができないため，多くのセル領域が未抽出になってしまう（**図 4-2**）．帳票書式に L 字型のセルが存在しない場合でも，罫線が途中で途切れた場合には L 字型の領域が生ずるため，この方式には罫線抽出誤りの影響を受けやすいという問題がある．交点の組合せによってセル領域を抽出する方式[16]は罫線誤りには頑強だが， 4 交点の組合せをセル領域として抽出するため，L 字型のセルが抽出できないという問題は同様である．

筆者らは, これまで主に銀行や証券などの特定業務向け帳票データ入力システムを開発してきた．業務システムでは特にデータの確実な入力が求められる．誤りが多ければシステムが使われなくなり，データエントリ業務の効率化という本来の目標が達成できなくなってしまうため，近年の複雑な書式の帳票でも高精度に認識できる技術が必要とされる．

本章では，帳票画像から抽出した罫線情報を入力として[19], [20]，罫線座標から抽出した交点を順に追跡することによってセル領域を抽出する方式を提案する[21]．本方式は，交点を順に辿って閉領域を求めるため，矩形以外の形状のセルも抽出できるという特徴を持つ．更に，罫線抽出の曖昧さを全体最適化で解決するために，交点ごとに尤度を設定して複数のセル候補を生成し，最後に最適なセルの組合せを求めるというアプローチを取る．以下，第 4.2 節では交点に基づくセル領域抽出方式について述べる．第 4.3 節ではセル候補の組合せ探索によって全体最適なセル集合を求める手法について述べる．第 4.4 節では評価実験によって本手法の有効性について検証する．本章の最終節でまとめと考察を述べる．

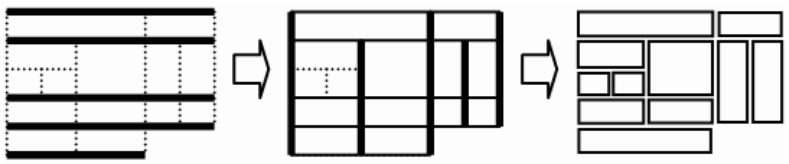


図 4-1　平行な罫線で分割するセル抽出方法

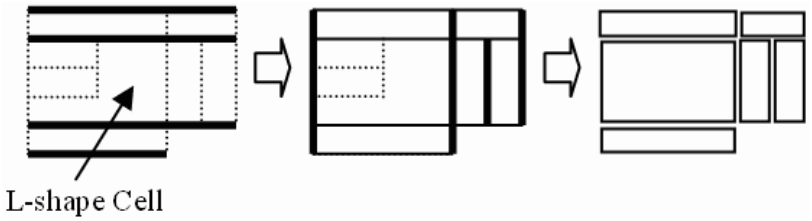


図 4-2　L 字型セルの抽出漏れ

4.2.　交点追跡に基づくセル候補抽出

本章で述べるセル抽出方式は，交点追跡に基づくセル候補抽出と，セル候補の組合せ探索とで構成される．前者のセル候補抽出は，罫線の位置関係を単純化した格子であるグリッドを罫線座標から生成するステップ，グリッド上の格子点に交点を登録するステップ，登録された交点を時計周りに辿ってセル候補を抽出するステップからなる．本節では，これらの各ステップについて順に述べる．後者のセル候補の組合せ探索については次章で述べる．

4.2.1.　グリッド生成と交点登録

セル抽出の第 1 のステップでは，グリッドを生成する．まず，縦罫線では x 座標，横罫線では y 座標が近いものをグループ化し，通し番号を割り当てる．この罫線グループを二次元空間に配置し，グリッドを作成する．例えば，**図 4-3(a)**のような罫線情報が得られた場合，**図 4-3(b)**のように縦罫線が 4 グループ，横罫線が 6 グループに分類され，4×6 のグリッドを生成する．

次に，第 2 のステップでは，グリッドの格子点に交点属性を登録する（例えば，**図 4-3(c)**）．交点属性は罫線が接続するパターンを形状で分類したもので，**図 4-4(a)**のように，T 字型，L 字型，I 字型，＋型の 11 種類を用いる．交点属性の種類を求めるためには，まず，グリッドの格子点において，上下左右の方向に罫線が接続しているかどうかを表す線方向属性（**図 4-4(b)**）を求める．続いて，線方向属性の組合せによって交点属性の種類を定める．例えば，**図 4-4(a)**のうちで ID 番号が i_1の交点属性は，格子点の上，右，下の 3 方向に罫線が接続する形状を示しており，**図 4-4(b)**の線方向属性の ID=l_1,l_2,l_4を組み合せたものに相当する．

更にこのステップでは，**図 4-5** に示すような不正な状態にある線方向属性の削除も行う．そのために，線方向属性の不正な状態を**図 4-6** のように分類する．第一の分類は，隣接する格子点が逆方向の正対する線方向属性を持たない状態であり，これを孤立線方向と呼ぶ（**図 4-6(a)**）．第二の分類は，一つの格子点に登録された線方向が一つだけの状態であり，これを単独線方向と呼ぶ（**図 4-6(b)**）．不正な線方向属性を削除すると新たに別の不正な線方向属性が生成されることがあるので，グリッド中の全ての格子点において孤立線方向と単独線方向を削除するためには，削除すべき線方向属性が無くなるまで繰り返し削除を行う必要がある．

なお，この削除操作は，グラフ理論において次数が 1 の頂点を削除する操作に相当する．セル領域は閉路を構成するが，次数 1 の頂点を削除しても閉路は失われないため，削除操作はセル抽出には影響しない．

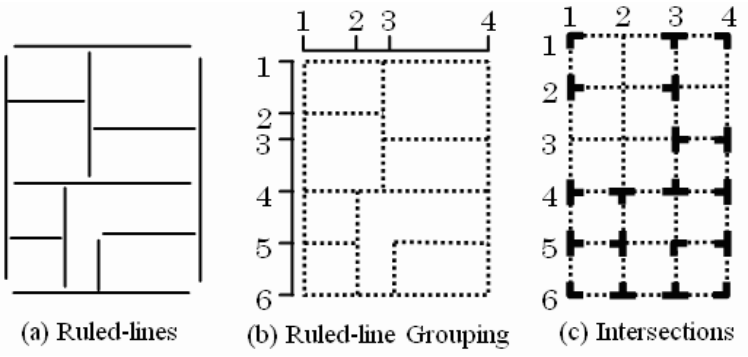


図 4-3　交点登録までの処理の流れ

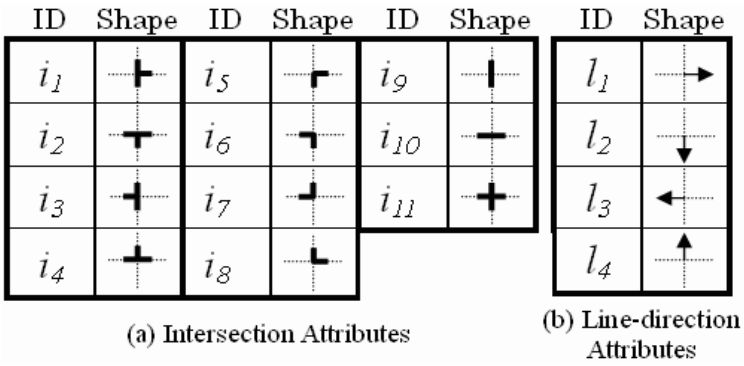


図 4-4　交点属性と線方向属性

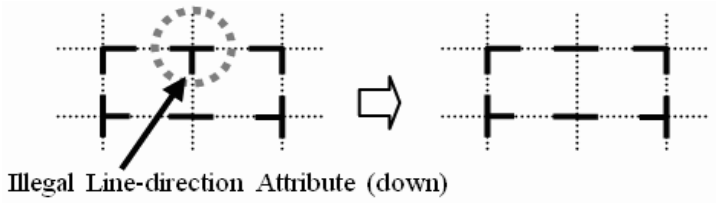


図 4-5　不正な線方向属性の削除

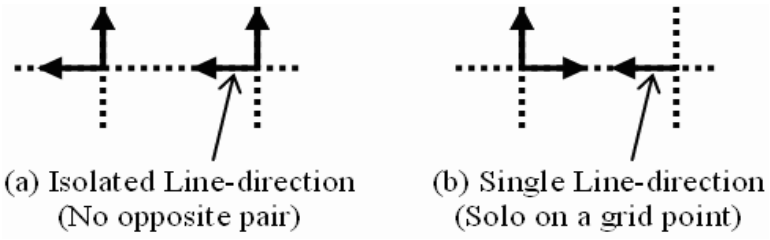


図 4-6　不正な線方向属性の種類

4.2.2. 交点追跡によるセル領域抽出

セル領域抽出は，グリッド上に登録された交点を時計周りに追跡することによって行う．セルの左上角を示す交点属性 i_1, i_2, i_5, i_{11} のそれぞれを始点として時計回りに交点を辿り，始点に戻った時点でセル領域を一つ抽出する．この経路を交点追跡パスと呼ぶ．例えば，図 4-7 に示すグリッドには点線の円で示すように左上角の交点が 4 箇所あり，これらを始点として時計回りに交点を辿り，四つのセルを抽出する．なお，我々は，表として不自然な形状のセルが抽出されるのを防ぐため，セルの頂点数に上限を設けている．

交点追跡を容易にするため，各交点には交点追跡パスが通過できる方向を交点通過方向（**図 4-8**）として登録しておく．例えば，**図 4-9** に交点 ID=i_1 の交点と交点通過方向の関係を示す．交点追跡パスは，各交点を時計回りに辿るので，交点 ID=i$_1$ を三つの方向から通過する可能性がある．それぞれの通過方向を表したのが交点通過方向 ID=p_1,p_4,p_{10} である．

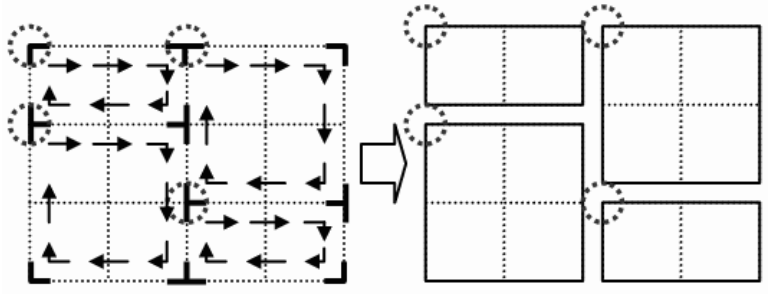


図 4-7　セル候補生成

ID	path direction	ID	path direction	ID	path direction
p_1		p_5		p_9	
p_2		p_6		p_{10}	
p_3		p_7		p_{11}	
p_4		p_8		p_{12}	

図 4-8　交点通過方向属性

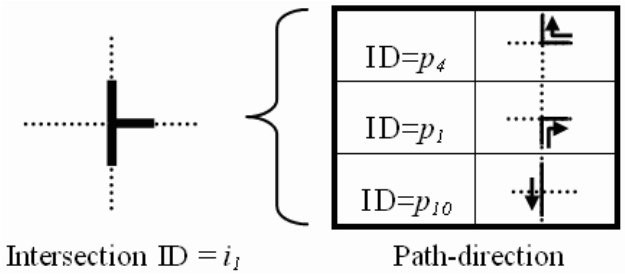


図 4-9　交点と交点通過方向の関係

4.3. 複数セル候補の組合せ探索

全ての罫線が正しく抽出できていれば，4.2 節で述べた方式によって高精度なセル抽出が可能である．しかし罫線抽出の結果は一般に曖昧性を含んでおり，誤って確定した罫線に基づいてセル抽出を行っても正しい結果を得ることはできない．そこで，罫線抽出結果の曖昧性を複数のセル候補の尤度で表わし，セル候補の最適な組合せを求めることによって，全体最適性を考慮したセル抽出方式を提案する．本節では，交点属性に尤度を導入して複数の尤度付きセル候補を生成する方法と，セル候補から最適なセル集合を求める組合せ探索手法について述べる．

4.3.1. 複数セル候補を用いたセル抽出

罫線抽出結果が誤りを含む可能性がある場合，交点属性の誤り可能性を考慮して複数のセル候補を生成する．**図 4-10** は複数セル候補生成の例である．**図 4-10(a)(b)**は，縦罫線が交点付近でかすれており，横罫線に達しているか否か決定できない状態を示す．この場合，交点追跡パスは**図 4-10(c)** のように 2 種類のルートを通り，二つのセル候補を生成する（**図 4-10(d)**）．

次に，様々な位置で生成された複数のセル候補の中から最適なセル候補の組合せ集合（最適セル集合）を求める．最適セル集合を求める問題は，得られたセル候補の集合が与えられた時に，表領域をいくつかのセル候補で敷き詰める問題（パネル敷き詰め問題）として定式化できる（**図 4-11**）．この時，敷き詰められたセル候補は互いに重なってはならず，また隙間はできるだけ少なくする．更に，**図 4-11(a)**において(A)と(B)が同じ領域を占めているように，異なったセル候補の組合せが全く同じ領域を占める場合は，その中から最適な組合せを選択する必要がある．そのため，セル候補の組合せごとに尤度値を設定して優劣比較を行う．

$$P(X)=\prod_{i=1}^{N}P(X_i) \qquad \text{(式 4-1)}$$

$$\begin{aligned} L(X) &= \log(P(X)) \\ &= \sum_{i=1}^{N}\log(P(X_i)) \qquad \text{(式 4-2)} \\ &= \sum_{i=1}^{N}Lc(X_i) \end{aligned}$$

セル集合 X を構成するセル候補を $X=X_1,X_2,\ldots,X_N$ とすると，セル集合が正解である確率 $P(X)$は**式 4-1** で計算できる．我々は，計算の簡略化や計算精度の利点から，**式 4-2** のような対数確率を尤度として用いる．

最適セル集合は，**式 4-2** の尤度が最大になるようなセル候補の組合せである．しかし，**式 4-1** および**式 4-2** を用いた場合，セル候補の数が多いほど多くの値が加算され，相対的に評価値が下がるため，セル候補数 Nが小さなセル集合が優先して選択されるという問題が生ずる．そこで，セル候補の対数確率 $Lc(X_i)$の平均値をセル集合の尤度値とした**式 4-3** を用いることで，セル候補数による偏りを補正する．

$$L'(X)=\frac{1}{N}\sum_{i=1}^{N}Lc(X_i) \qquad \text{(式 4-3)}$$

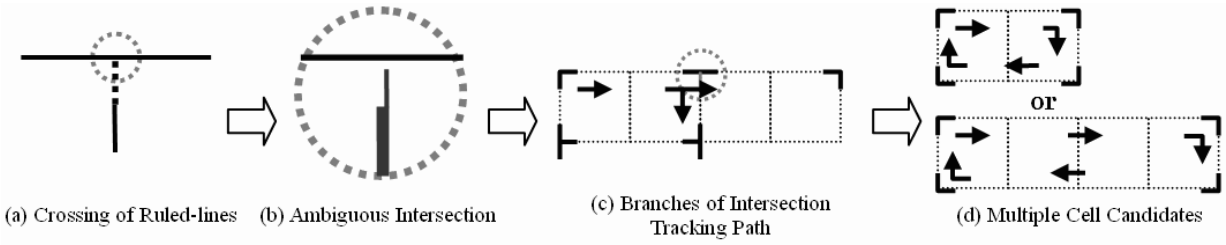


図 4-10　複数セル候補生成

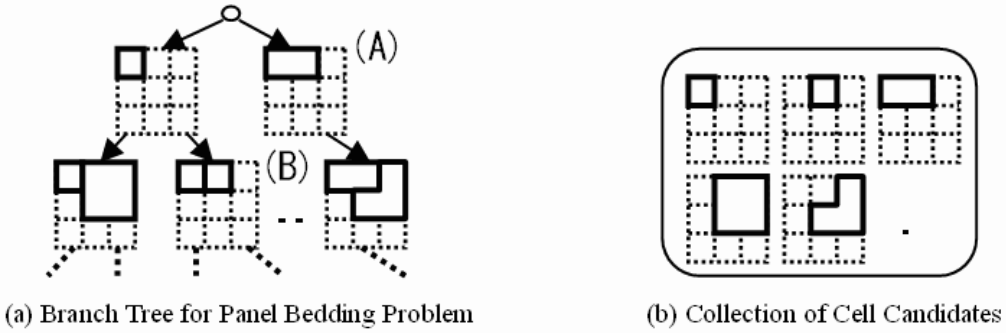


図 4-11　セル候補の組合せ探索

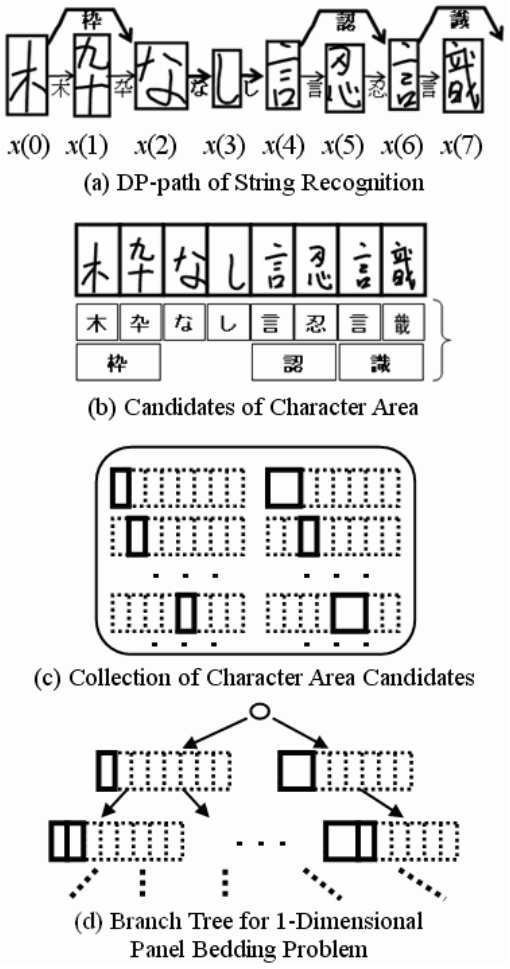


図 4-12　文字列認識における組合せ探索

尤度値に基づいて全体最適な組合せを求める問題は，文字列認識ラティスからの DP マッチング[22]～[25]に類似した問題だと解釈できる．文字列認識の DP パスは**図 4-12(a)**および**図 4-12(b)**のように表せるが，**図 4-12(b)**を**図 4-12(c)**のように文字領域候補の集合として表わすと，**図 4-12** は**図 4-11** が一次元の場合に相当することが分かる．

文字列認識では，DP パス上の文字領域候補を $X=X_1,X_2,...,X_N$ とし，文字候補 $C=C_1,C_2,...,C_N$ が得られた場合に，その DP パスが正解である確率 $P(C,X)$ を，**式 4-4** で評価することができる．

$$\begin{aligned} L(C,X) &= \log(P(C,X)) \\ &= a\sum_{i=2}^{N}\log(P(C_i \mid C_{i-1})) \\ &\quad + b\sum_{i=1}^{N}\log(P(C_i \mid X_i)) \\ &\quad + c\sum_{i=1}^{N}\log(P(X_i)) \\ &\quad + Np \end{aligned} \qquad \text{(式 4-4)}$$

第 1 項は言語情報のバイグラム確率，第 2 項は文字認識確率，第 3 項は文字領域候補が選ばれる確率である．a～c は各項の重み定数である．第 4 項は長い文字列ほど確率値が低くなる傾向を避けるための補正項で，文献[26]の単語挿入ペナルティを文字列長の補正に適用した[23]．なお，文字列長の補正には，各項を DP パス長で割って平均値を求めた式（**式 4-5**）を用いる方法も提案されている[25]．我々が**式 4-3** で対数確率の平均値を用いたのは，この方法に基づいている．

式 4-5 を，最適セル集合を求めるための評価値と考えた場合，セル領域そのものが認識結果なので，第 2 項は省略できる．第 1 項の言語情報確率は，例えば同じ高さや類似形状のセルが隣接しやすいなど，セルの相互関係の評価に対応させることを示唆する．各項の重みを表す定数 c は，項が一つの場合は省略できるので，**式 4-5** は**式 4-6** のように変形でき，**式 4-3** と同等の評価式が得られる．

$$\begin{aligned} L(C,X) &= a\frac{1}{N-1}\sum_{i=2}^{N}\log(P(C_i \mid C_{i-1})) \\ &\quad + b\frac{1}{N}\sum_{i=1}^{N}\log(P(C_i \mid X_i)) \\ &\quad + c\frac{1}{N}\sum_{i=1}^{N}\log(P(X_i)) \end{aligned} \qquad \text{(式 4-5)}$$

$$\begin{aligned} L(X) &= c\frac{1}{N}\sum_{i=1}^{N}\log(P(X_i)) \\ &= \frac{1}{N}\sum_{i=1}^{N}Lc(X_i) \end{aligned} \qquad \text{(式 4-6)}$$

4.3.2. セル候補尤度

本節では，**式 4-3** に示したセル候補尤度 $Lc(Xi)$を近似的に求める計算方法を述べる．セル候補尤度は，交点追跡パスが通過した交点通過方向の尤度から算出し，交点通過方向の尤度は交点通過方向を構成する線方向の尤度から求める．ここでは，線方向尤度，交点通過方向尤度，セル候補尤度の求め方を順に述べる．

線方向尤度は，格子点に罫線が接続しているか否かを表す値である．例えば，**図 4-13(a)**のように罫線の端がグリッドの格子点 Xから dだけ離れている場合に，格子点 Yとの間隔 L を用いて，**式 4-7** で線方向尤度 Sp を定義する．これは，罫線が格子点 Xに接続していれば Sp は 1.0 であり，罫線端と格子点との距離 d が大きいほど尤度 Sp が小さくなるように設定した計算式である．d が L の 1/3 より大きい場合は罫線が格子点に接続していないものとして尤度 Sp を 0 とする．

線方向尤度は格子点ごとに 4 方向に対して求め，それぞれ $Sp(u)$（上方向），$Sp(r)$（右方向），$Sp(d)$（下方向），$Sp(l)$（左方向）と表記する（**図 4-13(b)**）．

$$\begin{aligned} &if\,(d > L/3) \qquad Sp = 0.0 \\ &else \quad if\,(d > 0) \qquad Sp = 1 - 3 \times d / L \\ &else \qquad Sp = 1.0 \end{aligned} \qquad \text{(式 4-7)}$$

続いて，交点通過方向尤度の計算方法について述べる．交点追跡パスが交点を通過する際の進路は，右折（rt），直進（st），左折（lt）の 3 種類の可能性がある（**図 4-13(c)**）．**図 4-13(c)**では，交点追跡パスは下方向から通過しているので，交点追跡パスの右折は，交点の下方向と右方向に罫線が接続していることを示す．そこで，右折の尤度 $Se(rt)$は，$Sp(d)$と $Sp(r)$の積で計算する．一方，直進の尤度 $Se(st)$は，下方向と上方向に罫線が接続しており，同時に右方向に罫線が接続していない場合の尤度なので，$Sp(d)$と $Sp(u)$の積と，更に右方向の接続を否定するため $1\text{-}Sp(r)$を掛け合わせた値を尤度とする．左折の尤度 $Se(lt)$についても同様に定義した計算式を**式 4-8** に示す．

$$\begin{aligned} Se(rt) &= Sp(d) \times Sp(r) \\ Se(st) &= Sp(d) \times Sp(u) \times \{1 - Sp(r)\} \\ Se(lt) &= Sp(d) \times Sp(l) \times \{1 - Sp(r)\} \times \{1 - Sp(u)\} \end{aligned} \qquad \text{(式 4-8)}$$

式 4-8 において，右折の尤度 $Se(rt)$が $Sp(u)$と $Sp(l)$の項を持たないのは，交点追跡パスが右折する場合には上方向と左方向の罫線の有無が尤度に影響しないためである．直進の尤度 $Se(st)$についても同様である．

式 4-8 は交点追跡パスが交点を通過する際の全ての選択肢の尤度を示すので，尤度の総和が 1.0 になるよう正規化する．交点追跡パスの進路を m（m=rt,st,lt）で表すと，交点(i,j)を交点追跡パスが通過する際の交点通過方向尤度 $Le(i,j,m)$は**式 4-9** で定義できる．

セル候補尤度は，交点追跡パスが辿った交点通過方向尤度 $Le(i,j,m)$の平均値を 0～100 に正規化した値とする．セル候補尤度の計算式を**式 4-10** に示す（Mは交点の数，m_kはパス上の k 番目の交点での m）．

$$Le(i, j, m) = \frac{Se(m)}{Se(rt) + Se(st) + Se(lt)} \qquad \text{(式 4-9)}$$

$$Lc(i, j) = \frac{100}{M} \times \sum_{k}^{M} Le(i, j, m_k) \qquad \text{(式 4-10)}$$

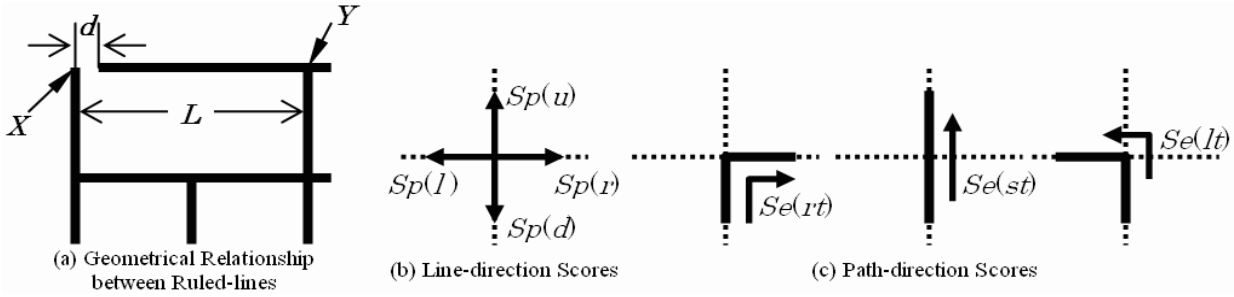


図 4-13　線方向尤度と交点通過方向尤度

4.3.3. セル候補の組合せ探索

セル候補の集合から最適セル集合を求める方法について述べる．DP マッチングを利用するためには，その問題が多段決定過程であり，次に示す条件を満たす必要がある[27].

条件 1)　第 i+1 段階の状態 $s(i+1)$は，第 i 段階の状態 $s(i)$と，i+1 段階での操作 $d(i+1)$によってのみ一意に決定され，i-1 段階以前の状態や操作には影響されない.

条件 2)　n段階全体での評価関数 fの値は, 各段階の状態と操作の関数の和として与えられる.

文字列認識（**図 4-12**）の例では，$x(i)$を終点とする最適パスは $x(0)$から $x(i-1)$までの最適パスと$x(i)$から求めることができるため，DP マッチングによる解決が可能である．一方，セル候補の組合せ探索の場合には，パネル敷き詰め問題としての条件（セルの重なりは不可，セル間の隙間は極力少なくする）による制約があるため，セル候補を終点とする最適パスはそれ以前に採用した複数のセル候補の形状に依存し，上記の条件 1 を満たさない.

そこで我々は, 表の一行を埋めるセル候補の組合せパターン (行内セルパターン) を全て生成し, その行内セルパターンを状態と考えることで DP マッチングを適用する. **図 4-14** のように表のグリッドの一行を全て埋めるセルの組合せを考えると, セルで埋められた行よりも上側の Area1 と下側の Area2 は完全に分離しており，それぞれの領域内でのセルの組合せは独立に評価できる．つまり N行で構成される表において，第 i 行目を埋めるセルの組合せが M_i通り作られる場合，DP マッチングの第 i 段階の状態が M_i通り存在するものと考え，状態 $s(i,m)$ $\{i=1...N,\ m=1...M_i\}$までの最適パスを求める問題と考えることができる.

セル候補の組合せ探索の手順を**図 4-15** に示す．**図 4-14** の Area1 と Area2 の境界となる行を探索境界（search boundary）と呼び，先ず一行目を探索境界とする．次に，探索境界を一行ずつ進めながら，探索境界の行を全て埋める行内セルパターンを生成する．ここで，行内セルパターンの下端の形状が同じものがあれば，それらのうちセル候補尤度が最も高いものを残してそれ以外は削除する．同様の処理を最後の行まで続ける.

上記のアルゴリズムは，分枝限定法[28]の枠組みで解釈することもできる．分枝限定法は問題を独立した部分問題に分割する操作（分枝操作）と，部分問題を終端する操作（限定操作）からなる．提案方式は，表の行ごとにセルの組合せを作成する分枝操作と，その中で同一の下端形状を持つ組合せ同士の尤度を比較(優越テスト)する限定操作により構成されるものだと捉えることができる.

図 4-14　行内セルパターンの例

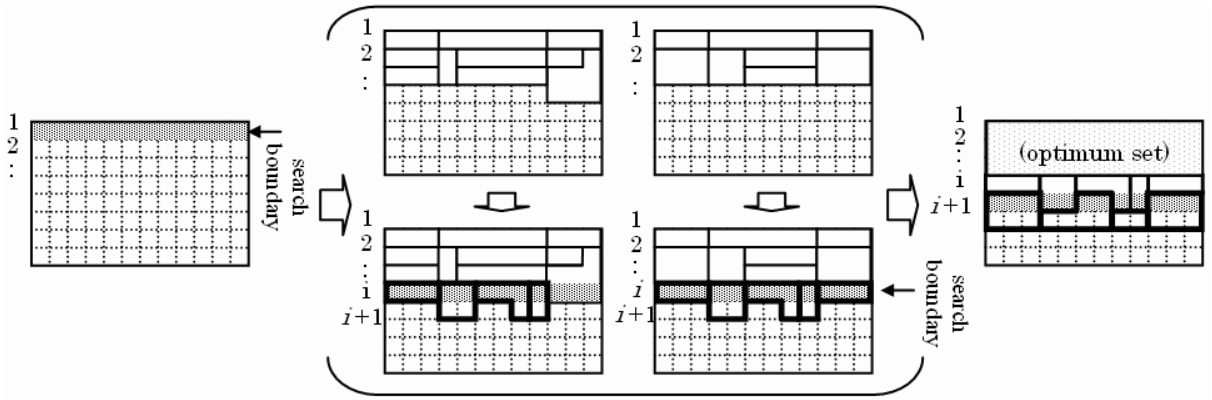


図 4-15　セル候補の組み合わせ探索の手順

4.3.4.　探索領域の限定

前節で述べたセル候補の組合せ探索は，各行を埋めるセル候補の組合せを全て生成するため，曖昧な罫線が多いと処理量が増大する危険性がある．しかし，組合せ探索が必要なのはセル候補が一意に決定していない領域だけなので,探索領域を限定することによって処理量の増大を抑制できる．例えば，**図 4-16** のように一部の罫線だけが曖昧な場合には，その罫線を含まない範囲のセルは一意に確定できる．したがって，図中の領域 A と領域 B のみに処理を限定すればよい．

探索領域を求める手順を**図 4-17** に示す．全てのセル候補が占める領域について，グリッドの単位格子ごとにセル数をカウントする．格子上に重なっているセルが二つ以上である領域を求め，最後にその領域を含むセル候補の領域を全て含む範囲を求める．これによりセルが一意に決まっていない領域を求めることができ，組合せ探索の対象領域が限定できる．

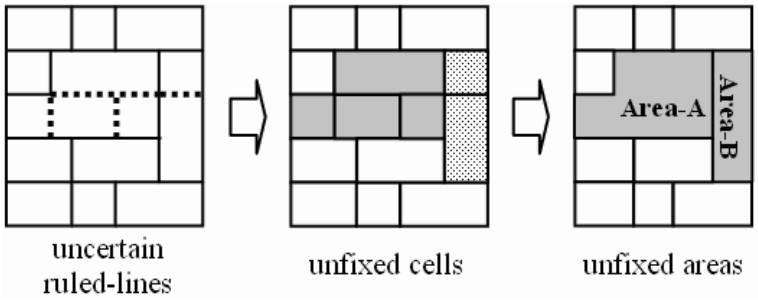


図 4-16　対象領域の限定

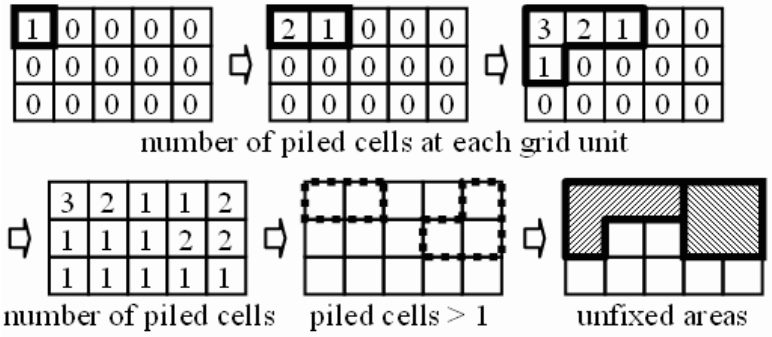


図 4-17　対象領域の生成

4.4.　評価実験

4.4.1.　評価画像と評価指標

実際の帳票画像に本方式を適用して，その有効性を検証した．評価に用いた画像は，44 種類の紙帳票をスキャナで読み込んだ解像度 200dpi のカラー画像である．内訳は，表項目に文字列データが記入されていない未記入帳票（fill-in form）が 12 画像と，データ記入済の帳票が 32 画像であり，記入済帳票は見積書（estimate sheet）が 15 画像と納品書（invoice）が 17 画像から構成される．概して，未記入帳票や見積書は画質が良いが，反対に納品書は別プリンタによる印刷（ポストプリント）や押印などがあったり，また，複数人の手を経ていたりして，画質が悪化しやすい．評価に用いた帳票画像の例を**図 4-18** に示す．

セル抽出精度の評価指標には再現率と適合率を用いた．再現率は正解セルの中で抽出に成功したセルの割合であり，適合率は抽出されたセル候補の中での正解セルの割合である．セル領域の外接枠の頂点座標を比較対象として，正解セルと抽出セルの頂点が全て 20pixel（約 2.54mm）以内であれば同一の位置に抽出できたもの（抽出成功）と数える．

処理時間の測定は，Windows XP SP3 を搭載した PC 上でのセル抽出関数の経過時間を実測することにより行った．ハードウェア環境は，Intel Pentium-D CPU（3.20GHz クロック）＋4GB メモリである．

なお，セルの頂点数の上限は 10 に設定した．

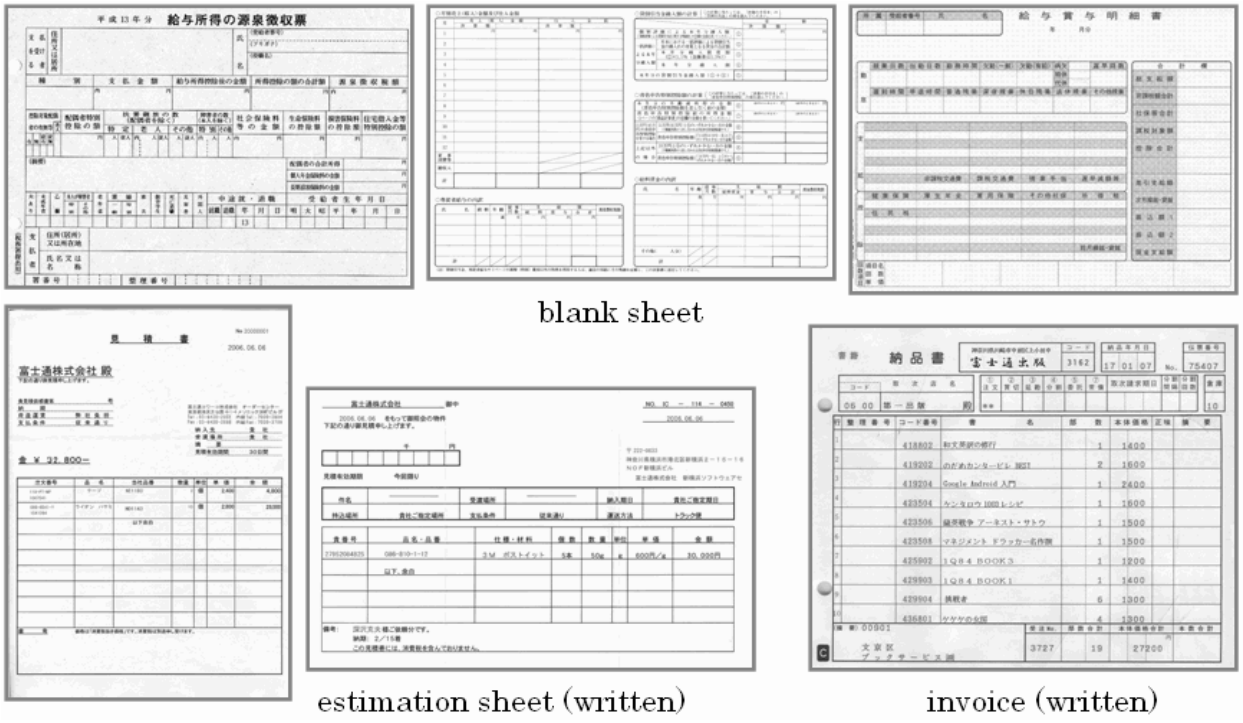


図 4-18　帳票画像の例

4.4.2. 従来方式との比較

従来方式と比較した結果を**表 4-1** に示す．従来方式としては平行な罫線で表を分割する方式（**図 4-1**）[12]～[14]を用いた．提案方式は，従来方式と比較して，適合率を維持しながら再現率を高めた．その性能は，未記入帳票，記入済見積書，記入済納品書の画質を反映している．一部の帳票（記入済見積書）では適合率が若干低下しているが，これは提案方式が複雑な形状のセルまで抽出する能力を持つことから，従来は抽出されなかったセル領域が抽出され，付加誤りが増加したものと解釈できる．

図 4-19 は L 字セルを含む表画像の例である．左から順に入力画像，従来方式の結果，提案方式の結果を示す．表の上部にある記入項目は，従来方式では脱落していたが，提案方式では抽出できている．記入項目を含む大きなセル（「右づめでご記入ください」とある）は L 字型の閉領域であり，従来方式では大きな長方形のセルとして抽出されるために内部の複数の記入項目が抽出できなかった．**図 4-19(c)**により，提案方式ではその問題が解決されたことが分かる．

表 4-1　セル抽出の精度

	従来方式		提案方式	
	再現率	適合率	再現率	適合率
未記入帳票	82.81%	93.04%	88.97%	93.04%
見積書（記入済）	87.06%	93.82%	87.85%	92.41%
納品書（記入済）	71.93%	77.56%	75.29%	79.01%
total	80.21%	87.47%	83.39%	87.52%

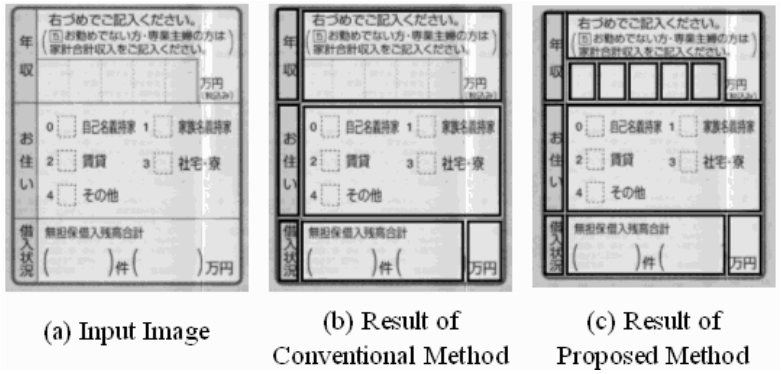


図 4-19　L 字セル認識結果の例

4.4.3. 提案方式の有効性

提案方式の第一の特徴は交点追跡によるセル候補の抽出であり，第二の特徴は複数セル候補の抽出と全体最適化の採用である．そこで，それぞれの特徴の有効性を検証するための比較実験を行った．

第一の特徴である交点追跡の有効性を確認するため，正解罫線を入力としてセル抽出を行った結果を**表 4-2** に示す．入力に誤りが無ければセル抽出率は 100%が期待できるが，再現率が 99.10%といくつかの誤りが生じている．誤りは，例えば**図 4-20** のようなケースで発生する．今回，不適切なセルを削除するためにセルの頂点数の上限を 10 に設定しているが，**図 4-20** では交点を追跡する途中で頂点数が上限に達してしまっている．この問題を解決する対策としては，頂点数の上限を大きくする，既に通過した交点を逆に通過する際に通過した交点の数をリセットする，などが考えられる．しかし，頂点数の上限を増やすと，例えば**図 4-21** のような不適切な形状のセル（頂点数は**図 4-20** と同じ）が候補に残る可能性が生ずるので，セル候補の適切さを再判定する機能を新たに導入する必要がある．

次に，第二の特徴である複数セル候補の抽出と全体最適化の効果を検証するため，線方向尤度を 0 または 100 として，セル候補を一つに確定するようにした結果との比較を**表 4-3** に示す．未記入帳票と記入済見積書の適合率で若干の低下が認められるが，記入済納品書では提案方法を用いることによって再現率，適合率とも格段に向上している．再現率の向上は 15 ポイント以上である．記入済納品書は他の 2 種の帳票に比べて認識精度が低く，認識が困難な画像であることから，複数セル候補の全体最適化は，罫線抽出の結果に曖昧性が高い場合に効果が高いと考えられる．一方，未記入帳票や記入済見積書での適合率における若干の低下は，提案方式で尤度が低いセルまで抽出してしまったことが原因だと考えられる．

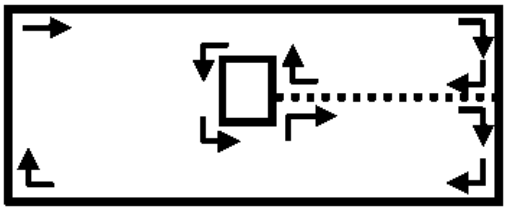


図 4-20　セルが脱落するケース

図 4-21　不適切な形状のセルの例

表 4-2　正解罫線によるセル抽出の精度

	再現率	適合率
未記入帳票	99.34%	99.80%
見積書（記入済）	99.01%	99.11%
納品書（記入済）	99.09%	99.02%
total	99.13%	99.26%

表 4-3　セル抽出の精度（単一セル候補との比較）

	単一セル候補		複数セル候補（提案方式）	
	再現率	適合率	再現率	適合率
未記入帳票	88.97%	93.11%	88.97%	93.04%
見積書（記入済）	87.85%	94.68%	87.85%	92.41%
納品書（記入済）	60.06%	75.29%	75.29%	79.01%
total	77.65%	88.11%	83.39%	87.52%

4.4.4. 処理時間の比較

セル抽出の処理時間を**表 4-4** に示す．セル抽出は一般に画像処理や文字認識に比べると処理時間は短く，従来方式でも 8.32msec とごく僅かだが，提案方式では更に短い 2.77msec まで短縮される．

従来方式では最初からセルを一意に確定するのに対し，提案方式では複数のセル候補の可能性を考慮するため，提案方式の方が処理時間は大きいものと予想していたが，予想に反した測定結果が得られた．その理由としては，従来方式では平行な罫線のペアを求める際に個々の罫線の座標を全て相互比較しているが，提案方式ではグリッドを用いることで罫線座標を比較する回数が削減できたことが考えられる．実際，座標の比較回数を評価画像でカウントすると，従来方式では一帳票あたり平均で約 68000 回であったものが，提案方式では約 2900 回に削減されていた．

表４－４には，セル候補の探索領域を限定することによる処理時間の削減効果も示した．探索領域の限定を行わない場合（領域非限定）では処理時間が 4.64msec なのに対し，領域限定を行うことで 2.77msec と約 60%に削減できている．

表 4-4　セル抽出の処理時間

	従来方式	提案方式	
		領域非限定	領域限定
未記入帳票	11.56msec	6.09msec	3.67msec
見積書（記入済）	2.10msec	1.97msec	1.16msec
納品書（記入済）	11.90msec	6.13msec	3.64msec
total	8.32msec	4.64msec	2.77msec

4.5. まとめ

未知フォーマットの帳票画像から表を構成するセル領域を抽出する方式を提案した．提案方式では，罫線が交差する交点を求め，交点を順に辿って閉領域を抽出する．また罫線抽出結果は一般に誤りを含むため，曖昧な罫線に対して尤度値を導入し，複数の尤度付きセル候補を生成する．更に，セル候補の最適な組合せを DP によって求め，全体最適なセル領域を決定する．44 種類の紙帳票をスキャナで読み込んだ 200dpi のカラー帳票画像に対して，従来方式と比較して再現率が 3.18%向上し，適合率は同程度の値が得られた．

従来の多くの表認識技術が抽出対象を矩形セルに限定していたのに対し，提案方式では交点追跡を用いることによって非矩形の領域も抽出できるようになった．また，複数のセル候補から組合せ探索によって全体最適なセル集合を求めるというアプローチは，表認識においてはこれまでに無いものである．評価実験により，罫線抽出結果の曖昧性が高い場合に全体最適化に基づく提案方式が特に有効であることが確認できた．

提案方式の課題としては，再現率が従来方式より向上したのに対し，適合率が同等精度に留まっている点が挙げられる．その理由は，提案方式はこれまで抽出できなかったセルまで抽出する能力があり，付加誤りが増加する場合があるためだと考えられる．今後は，セル候補尤度の算出法や全体最適化の評価値の改善によって付加誤りを抑制できるよう方式改善を進める．例えば，今回は省略したセル候補の相互関係の評価値（**式 4-4** の第 1 項）の導入や，不適切なセル（**図 4-21**）を判定する基準の適用などの改善案が考えられる．

5. 文字抽出用二値化の研究

概要

本章では，多値文書画像から，文字の二値画像を解像度に寄らずに高精度に抽出する二値化方式を提案する．文字抽出用二値化は，入力画像から文字の近傍領域を抽出し，文字近傍ごとに文字画素と背景を分離する．既開発の方式では近傍内の二値化に Niblack 二値化を用いており，低解像度画像で細い線が途切れるという問題があった．提案方式では，二値化閾値を後処理で補正して画素の脱落を抑制する．更に，高解像度画像では大津二値化を併用することで更に高精度な二値化が実現できる．150～600dpi の各解像度の画像で文字認識評価実験を行い，本方式の有効性を確認した．

5.1. はじめに

近年，業務データの電子化の進展に伴い，紙書類がスキャナ等で電子化保存されるケースが増えている．また電子書籍ブームにより，個人ユーザが書籍を裁断してスキャンし，PDF 等の形式で自前の電子書籍を作成する，いわゆる「自炊」も一般化している．

大量の文書画像を活用する場合，目的とする文書を効率よく見つけるため，OCR 等でテキストを抽出し，キーワード検索を可能にする技術が不可欠である．しかし従来のような特定業務向け定型文書の認識に比べ，近年は様々なデザイン・画質の文書画像が認識対象となっており，文字認識はより困難にな状況にある．更に OCR が利用される環境も多様化が進んでいる．iPhone 等のスマートフォンの普及により，デジタルカメラで文書画像を取得するケースも増え，画像のボケや照度不足による画質の劣化が文書画像認識が解決すべき重要な課題となっている．

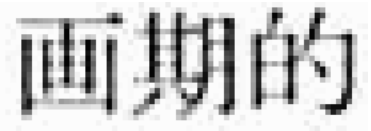
我々は，これまでに劣化カラー画像を対象に高精度な文字認識を実現する技術を開発してきた．その一つのアプローチは劣化画像に強い文字抽出用二値化技術の開発[1][2]であり，もう一つは劣化画像文字認識技術の開発[3][4]である．前者の二値化技術は，文字近傍領域を抽出し，各領域内で高解像度化した画像を二値化することで低解像度画像でも高品質な文字画像を抽出する．後者の文字認識技術は，多値画像を二値化せずに直接認識することで，二値化による情報脱落を避けて高精度な文字認識を実現する[3]．更に，高品質な画像では二値画像を用いた方が高い精度が得られることから，二値／多値の認識方式を組み合わせて画質によらず高精度な文字認識を行う技術[4]も開発した．

大量の文書画像を認識する場合は計算時間も重要である．我々が開発した多値画像認識は文字カテゴリごとに固有空間を用いるため，二値画像の認識に比べて計算量が大きいという問題がある．そこで今回，我々は前者の二値化技術を改良することによる劣化画像認識の高精度化を検討した．

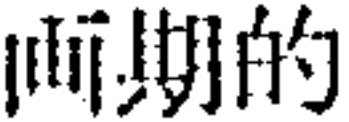
低解像度画像では文字あたりの画素数が少ないため，画質の劣化が文字認識の精度低下に影響しやすい．例えば画像のボケや量子化誤差が発生すると，本来は単一色である文字画素に明度の差が生じ，二値画像に途切れが生ずる（**図 5-1(a)(b)**）．一方，高解像度画像では画像のボケが文字サイズに対して僅かなので二値画像への影響は少ない（**図 5-2(c)(d)**）．薄い画素の途切れは，単に背景との明度差が小さい（コントラストが低い）ためだけでなく，薄い画素の近くに濃い画素があることによる二値化閾値の変化も関係する．これは薄い罫線の近くに濃い文字がある場合に罫線が途切れる現象として文献[5][6]で分析したもので，同様の現象が文字画素においても生じている．**図 5-1(a)**の場合，文字の横画が縦画よりも薄いため，横画の画素が相対的に背景クラスに分類された

のが途切れの主要因の一つである．

本稿では，二値化閾値を補正して文字の途切れを改善する方法について述べる．これは文献[5][6]で罫線抽出に対して適用した手法を文字画像に応用したものである．更に，本手法が特に低解像度画像に有効であることから，解像度に応じて処理を切り替える方法についても検討した．第 5.2 節では我々が開発したテキスト抽出用二値化[1][2]の概要を記す．第 5.3 節で二値画像が途切れる原因と，それを改善するための閾値補正について述べる．第 5.4 節で評価実験を行い，本方式と解像度の関係について検証する．



(a) 低解像度の文字画像



(b) 低解像度の文字画像(a)を Niblack 法で二値化した画像

(c) 高解像度の文字画像

(d) 高解像度の文字画像(c)を Niblack 法で二値化した画像

図 5-1 文字画像の二値化

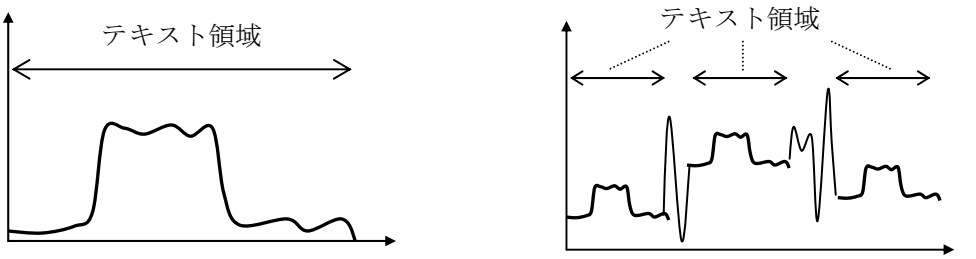
5.2. テキスト抽出用二値化

5.2.1. テキスト領域抽出と 2 クラスモデル

テキスト抽出用二値化について概観する．大津[7]，Niblack[8]等の標準的な二値化手法は入力画像を画素ごとの値に応じて 2 群に分類するのみだが，テキスト抽出用二値化は文字を構成する画素を抽出することが目的である．そのため，二値化の対象となるテキスト領域を分離抽出する処理と，テキスト領域内を二値化する処理から構成される．

テキスト領域抽出は，処理対象画像によって様々な手法が提案されている．業務用帳票のように単純な書式の文書画像では，先ず標準的な手法で二値画像を生成した後に，文字とそれ以外と分離するというアプローチが可能である．一方で，図や写真が多用された雑誌文書や，情景画像からのテキスト抽出では，エッジ情報やテクスチャ情報，色クラスタリング等を用いてテキストの近傍を抽出する技術が必要となる[9]～[11]．また，多くの手法では一様な文字色が仮定されているが，複数色の文字を各色の部分要素の組合せで抽出する手法も提案されている[12]．

抽出されたテキスト領域から二値画像を生成する方法は，一般の文書画像や情景内の看板抽出のような場合，テキスト領域内を平坦な背景と一様色の文字からなる 2 クラスモデルで表せるものとみなし，Niblack 等の標準手法を適用することができる（**図 5-2**）．



(a) 2 クラスで表せる画像　　(b) 区分的に 2 クラスで表せる画像

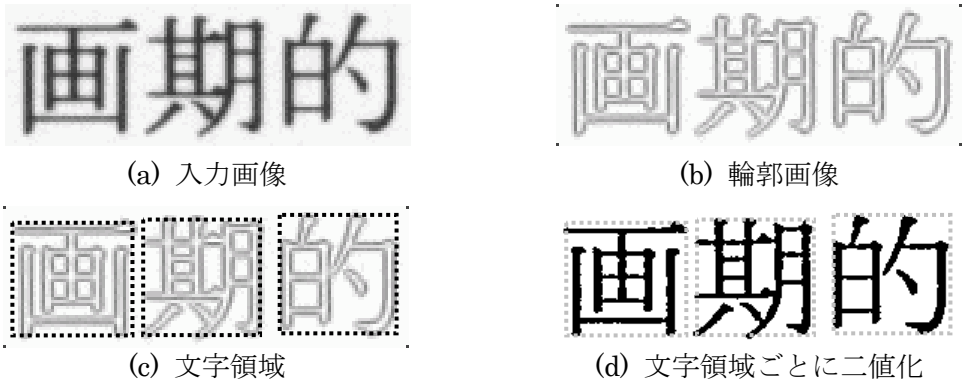
図 5-2 区分的 2 クラス領域

5.2.2. テキスト抽出用二値化の流れ

本研究で提案する二値化手法は, 藤本らが開発した文字抽出用二値化[1][2]を改良したものである. そこで，まず藤本らの手法を概観する．二値化の対象は一般文書画像である．図や写真の存在も想定するが，文字領域は**図 5-2(b)**のように区分的な 2 クラスモデルで表せるものとする.

本手法は，(a)Sobel 画像を大津二値化して輪郭画像を生成，(b)輪郭画像を連結成分（CC: Connected Component）に分離，(c)文字らしい CC の重なり統合によって文字領域を抽出，(d)文字領域を二値化，という手順を行う（**図 5-3**）.（各処理の詳細は文献[1]を参照）

処理対象の定義により，文字領域は平坦な背景の上に文字の画素が配置された 2 層構造だとみなして二値化する．ここで我々は Niblack 二値化[8]を用いているが，領域内の背景を平坦だとみなせば，大津二値化[7]のような大域的二値化を用いることもできる.



(a) 入力画像　　(b) 輪郭画像

(c) 文字領域　　(d) 文字領域ごとに二値化

図 5-3 文字抽出用二値化の流れ

5.3. 閾値補正による途切れ改善

5.3.1. 二値画像の途切れの原因

前記, 藤本らの二値化手法には, 低解像度画像で文字線に途切れが発生するという問題がある（**図 5-1(b)**）．その原因は，低解像度画像での画素の量子化誤差により細い線が薄くなる（**図 5-4**）ことにあるが，更に Niblack 等の局所的二値化に特有の問題も生じる.

Niblack 二値化は, 画素ごとにその周辺の $w \times w$ の局所領域の画素値から二値化閾値を求める（**式 5-1～5-3**）．局所領域の平均値 m が基準になるため，薄い画素でも背景よりも相対的に濃い画素であれば黒画素として抽出できる．したがって，部分的に画素が薄くなった文字の二値化には，大津などの大域的二値化よりも Niblack 等の局所的二値化の方が向いている.

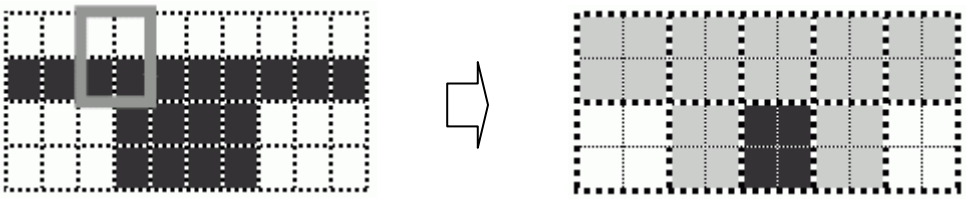
$$m = \sum_{i}^{w}\sum_{j}^{w} I(i,j)/w^2 \qquad \text{(式 5-1)}$$

$$\delta^2 = \sum_{i}^{w}\sum_{j}^{w} \{I(i,j)-m\}^2/w^2 \qquad \text{(式 5-2)}$$

$$T = m + k\delta \qquad \text{(式 5-3)}$$

しかし，対象画素の近くに濃い画素が存在し，それが局所領域の中に含まれた場合に問題が生ずる．例えば**図 5-5** は局所領域内での画素値の頻度分布を示している．左図ではテキストと背景の極大値の間に閾値があるため正しく二値化されるが，右図では近くにある黒い画素の影響で閾値が左にシフトするため，テキストの画素が背景側に分類される．これにより，薄い画素の途切れが発生する．

図 5-1(a)(b)の途切れ現象を見ると，薄い横画の近くには濃い縦画が必ず存在するため，ほとんどの横画が失われている．これは明朝体フォントで頻繁に発生する現象である．



(a) 高解像度画像　　(b) 低解像度画像

図 5-4 量子化誤差による低解像度画像のボケ

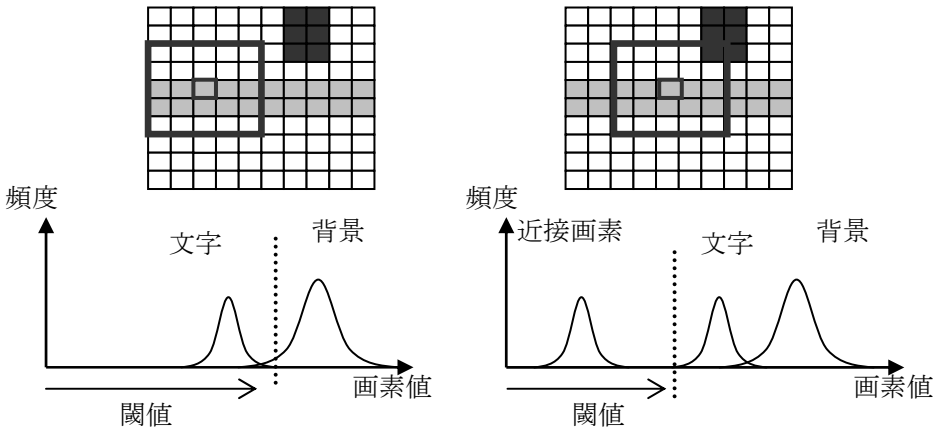


図 5-5 近接画素の閾値への影響

5.3.2. 閾値補正による途切れの改善

近接画素による画素の途切れを改善するため，シフトした二値化閾値を適正な値に補正する手法を適用した．処理の流れを**図 5-6** に示す．Niblack 二値化（Niblack thresholding）では，入力画像 $I(i,j)$の画素ごとに二値化閾値 $T(i,j)$を求める．大津二値化（Otsu thresholding）では，二値化対象の文字領域で大津閾値（T_{otsu}）を求める．背景判定（Background determination）では，初期閾値を用いて局所領域内を 2 クラス分類し，クラス平均の差分により平坦領域か変動領域かを判定する[13]．ここで平坦領域と判定された画素では大津閾値による二値化を行い，変動領域では Niblack 閾値を用いる．

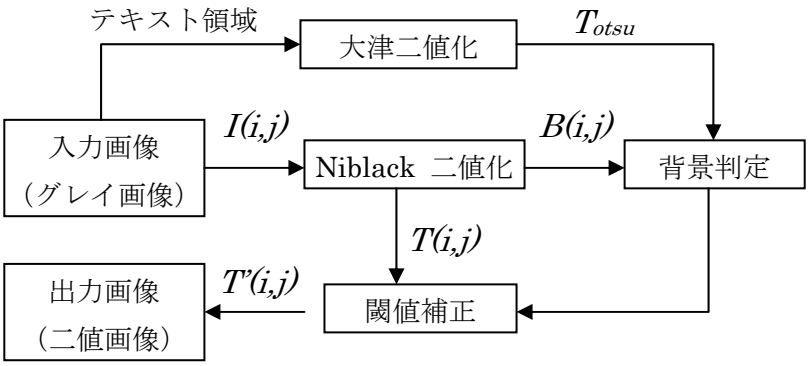


図 5-6 閾値補正を適用した二値化

閾値補正（Threshold correction）では，先に求めた Niblack 閾値をその周辺閾値によって補正する．近くに濃い画素がある場合の閾値は値が大きい方にシフトしているが，**図 5-7** のように局所領域をスライドすると，対象画素を含み，かつ濃い画素を含まない領域を設定することができる．この場合の閾値を新たな補正閾値とすれば途切れは解消する．そこで**式 5-4** のように，閾値 $T(i,j)$をスライド幅 Sの中での最小値に変換する．

$$T'(i,j)=\min_{k,l=-S}^{S}\{T(i+k,j+l)\} \qquad \text{(式 5-4)}$$

式 5-4 は**図 5-8(b)**のように対象画素の周辺領域をスキャンして，その範囲での最小閾値を求める処理を表す．しかし，閾値補正とはシフトした閾値を元に戻すためのものなので，対象画素と画素値が全く異なる座標での閾値を流用するのは適切とは言えない．

対象画素と二値化閾値が同等であるべき画素は，対象画素の周辺で同等の値を持つ画素である．つまり**図 5-8(c)**のように，薄いグレイの画素の範囲内での最小値を求めた方がより適切だと考えられる．そこで，**式 5-4** に定数 eの条件を追加して**式 5-5** のように修正する．

$$T'(i,j)=\min_{k,l=-S}^{S}\{T(i+k,j+l)\}$$
$$when\ |I(i+k,j+l)-I(i,j)|<e \qquad \text{(式 5-5)}$$

これを一次元の模式図で表すと**図 5-8(d)(e)**のようになる．**図 5-8(d)**において，ステップエッジの近くで閾値を表す点線が画素値を超えている部分が画素が途切れた位置である．ここで，$\pm S$の範囲内で画素値が同じレベルの部分の閾値が変換され，途切れが解消された様子が**図 5-8(e)**に示されている．

閾値補正を適用すると，**図 5-1(a)**の画像は**図 5-9** のように二値化できる．まだ必ずしも奇麗な二値画像ではないが，大きな途切れが改善され，文字認識が可能な程度の画像が得られている．

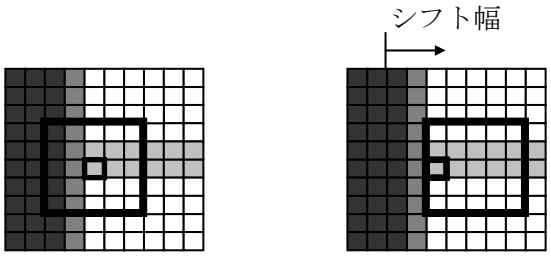
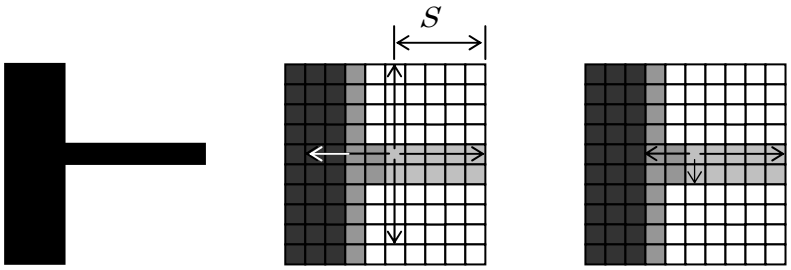
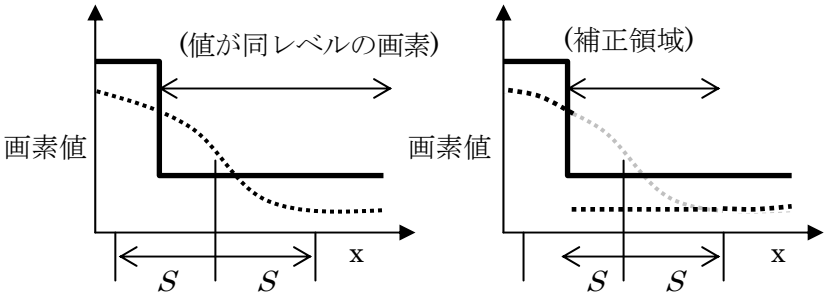


図 5-7 局所領域の移動



(a) 理想的な二値画像　(b) 閾値補正領域（**式 5-4** より）　(c) 閾値補正領域（**式 5-5** より）



(d) 閾値の補正　(e) 閾値補正領域

図 5-8 二値化閾値の補正

(a) 低解像度テキスト画像　(b) 閾値補正を適用して画像(a)を二値化した画像

図 5-9 文字画像の二値化（改善後）

5.4. 評価実験

5.4.1. 大津二値化・Niblack二値化との比較

提案方式は，文字領域内を Niblack 法で二値化した場合の文字の途切れを改善する方法である．その効果を検証するため，文字認識実験によって精度の比較を行った．5.2.2 節にて，文字領域内の二値化には大津二値化を用いても良いと述べたように，ここでは従来法として Niblack 二値化を用いた場合と，大津二値化を用いた場合，そして Niblack に閾値補正を適用した場合（提案方式）の比較を行う．

評価画像に 63 種類の文書画像を用い，テキスト領域の座標をあらかじめ与えて，その領域内を二値化して文字列認識した．文書画像は我々が評価用に収集したもので，内訳を**表 5-1** に示す．解像度は 150dpi～600dpi の 4 種類である．

表 5-2 によれば，従来法（Niblack）に対して，閾値補正を用いた場合（Corrected）ではほとんどの解像度で認識精度が改善しているが，その中でも特に低解像度での改善幅が大きい．これは，量子化誤差による文字画素の脱落が低解像度でより発生しているためである．高解像度でも僅かに精度向上しているが，その幅は小さい．一方，大津二値化を用いる（Otsu）と高解像度での精度は高いが，低解像度では Niblack にやや劣る．大津二値化の精度が高い理由については 5.4.3 節で考察する．

文字の途切れは明朝体フォントのように横画が縦画に比べて細いような場合に頻発する．そこで，評価画像の中から明朝体が主に使われている JEITA 標準画像のみを用いて評価した結果を**表 5-3** に示す．明朝体が多い文書で閾値補正による改善効果が大きいことから，提案手法が文字の途切れが目立つ画像で特に有効であることが分かる．

表 5-1 評価画像の内訳

Catalog	6	Office document	5
Comics	5	Pamphlet	5
JEITA 標準画像	17	Slide（ppt）	5
Magazine	6	Thesis	6
Newspaper	8	合計	**63**

表 5-2 文字認識精度の比較

	Niblack 二値化		大津二値化		提案方式	
	再現率	適合率	再現率	適合率	再現率	適合率
150dpi	82.9%	85.3%	82.1%	84.3%	83.4%	86.0%
200dpi	90.5%	91.7%	90.0%	90.4%	90.7%	92.2%
300dpi	92.7%	94.3%	94.4%	94.8%	92.4%	94.5%
600dpi	93.9%	93.9%	94.4%	94.8%	94.1%	93.8%

表 5-3 文字認識精度の比較（主に明朝体）

	Niblack 二値化		大津二値化		提案方式	
	再現率	適合率	再現率	適合率	再現率	適合率
150dpi	94.5%	93.1%	90.8%	87.6%	95.9%	95.2%
200dpi	96.8%	95.1%	94.3%	90.5%	97.9%	97.1%
300dpi	98.1%	97.5%	97.9%	97.1%	98.6%	98.4%
600dpi	98.2%	98.0%	98.0%	97.6%	98.8%	98.6%

5.4.2. ストローク幅を用いた改善

表 5-2 を見ると，150,200dpi では閾値補正を用い，300,600dpi では大津二値化を用いれば，画像に寄らずに高い精度が得られることが分かる．しかし，文字の途切れは画素数が少ない小さな文字で起きやすいのであって，入力画像の解像度に直接依存するわけではない（低解像度画像では小さい文字の割合が多いため，結果的に閾値補正がより有効）と解釈できる．したがって，画像の解像度ではなく，文字サイズによって方式を切り替えればより効果的だと考えられる．

二値化の段階では対象画素が属する文字のサイズは不明である．そこで，文字領域内を大津閾値で二値化して仮の二値画像を作り，ストローク幅を推定して，ストローク幅に応じて大津二値化と閾値補正を切り替える．ストローク幅は，仮の二値画像を縦横にスキャンして，ランの長さの平均値に基づいて算出する．

表 5-4 に実験結果を示す．定数 S に対して，ストローク幅が S 以上の場合に大津二値化を用い，S 未満の場合は閾値補正を用いた．この結果を見ると，低解像度画像では，ストローク幅が 1～2 の場合に閾値補正を行う（S=3）場合に認識精度が若干向上することが分かる．一方，高解像度では S=2 の方が良く，評価画像の解像度に依存した結果が出ている．これは評価画像での文字サイズの分布の影響ではないかと考えている．

表 5-5 は，**表 5-3** と同様に JEITA 標準画像のみで集計した結果である．元々，大津二値化での精度が高くないため，定数 S が大きいほど精度は高いが，S=3 で値がほぼ安定していることから，閾値補正の効果はストローク幅が 1～2 の場合に特に顕著であることが分かる．

表 5-2, **表 5-3** と**表 5-4**, **表 5-5** を比較すると，画像の解像度で方式を切り替えた場合（150,200dpi は Niblack+閾値補正，300,600dpi は大津）に比べて，ストローク幅を用いた方が精度のピーク値は高いことが分かる．しかしその差は僅かであり，画像の種類によってパラメータの最適値も異なるので，ストローク幅を用いた方法が常に優れているとまでは言えない．

表 5-4 ストローク幅（S）による改善

S	2		3		4		5	
	再現率	適合率	再現率	適合率	再現率	適合率	再現率	適合率
150dpi	83.1%	85.8%	83.4%	86.2%	83.4%	86.1%	83.4%	86.0%
200dpi	90.8%	91.7%	91.1%	92.5%	90.9%	92.4%	90.8%	92.3%
300dpi	94.6%	94.8%	93.2%	94.9%	92.9%	94.8%	92.7%	94.8%
600dpi	94.5%	94.8%	94.5%	94.7%	94.5%	94.8%	95.1%	94.8%

表 5-5 ストローク幅（S）による改善（主に明朝体）

S	2		3		4		5	
	再現率	適合率	再現率	適合率	再現率	適合率	再現率	適合率
150dpi	93.6%	92.1%	95.9%	95.3%	95.9%	95.2%	95.9%	95.2%
200dpi	95.8%	93.2%	97.8%	97.0%	97.9%	97.1%	97.9%	97.0%
300dpi	98.0%	97.2%	98.6%	98.4%	98.7%	98.5%	98.7%	98.5%
600dpi	98.2%	97.7%	98.2%	97.7%	98.2%	97.7%	98.3%	97.7%

5.4.3. 考察

高解像度画像で大津二値化の精度が高い理由を考察する．Niblack 二値化は局所的な変動を抽出できる半面，微小なノイズまで拾ってしまうという問題がある．一方，大津二値化では領域内で閾値が一定なので，局所的なノイズの影響を受けにくい．これは劣化画像の文字特徴を抽出できないという問題の裏返しであるが，高解像度画像では小さな文字が少ないため，大津二値化の良い面が勝るのだと思われる．

ストローク幅の推定値が小さければ，文字中で途切れやすい箇所が多いため，閾値補正がより有効に働くというのは予想通りである．しかしストローク幅の推定値がどの程度妥当かについては議論の余地がある．我々の実験では，文字領域を大津二値化で仮に二値化し，縦横のランの長さの頻度分布を用いてストローク幅を推定したが，そもそも文字の途切れが発生した二値画像を用いて推定しているため，ストローク幅の推定値も途切れの影響を受けていると考えられる．また，文字領域の中には太いストロークも存在するので，適切な推定値を求めるには更なる分析が必要である．

5.5. まとめ

本報告では，多値文書画像から，文字の二値画像を解像度に寄らずに高精度に抽出する二値化方式について述べた．低解像度画像では，二値画像における文字の途切れがしばしば見られ，文字認識精度が低下する原因となっている．我々はその原因の一つとして局所的二値化における閾値の変動に着目し，二値化閾値を近傍の閾値によって補正することによって文字の途切れを改善する手法を提案した．また，高解像度画像では文字領域の二値化に大津二値化を用いた方が高精度である場合があることから，解像度によって二値化の方式を切り替える方法が有効であることを示した．更に，二値化方式の切り替えを入力画像の解像度によって行うのではなく，文字領域におけるストローク幅を参照して切り替える方式も提案した．

閾値補正が文字の途切れ改善に有効であることは，実画像による目視確認と，文字認識精度の比較によって示すことができたと考えている．また，画像の解像度によって二値化の方式を切り替える方法の有効性も示すことができた．ストローク幅を用いる方法については，可能性は示したものの改善効果は不十分で，更なる検討が必要である．

今後は，各手法の有効性を文書画像の種類ごとに調査し，誤り原因を分析することで，より有効な方式を確立したいと考えている．具体的には，文字領域を二値化する際に，大津二値化が Niblack 二値化よりも高精度である場合の原因調査（大津の優位点，Niblack の欠点）や，ストローク幅推定の精度向上などを行い，どのような種類の文書画像に対しても有効な二値化方式の実現を目指す．

6. 帳票画像認識技術の実用化

概要

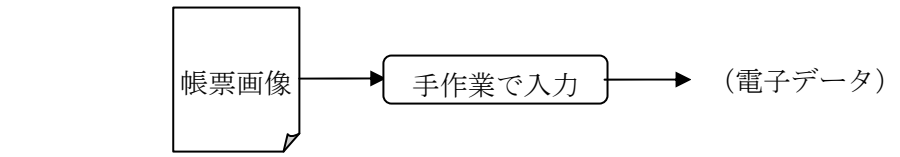
本章では，本研究において開発した帳票画像認識技術の実用化について述べる．本研究で開発した技術は，帳票画像認識を用いた製品やサービスに適用されると共に，帳票に限らず一般文書画像の認識精度向上のためにも用いられる．帳票画像認識技術を実用化するためには，単に認識精度を改善する以外にも実用化のための工夫が必要となる．本章では，帳票画像認識と一般文書画像認識のそれぞれについて，実用化のために考慮しなければならないポイントについて述べる．

6.1. 帳票画像認識への適用

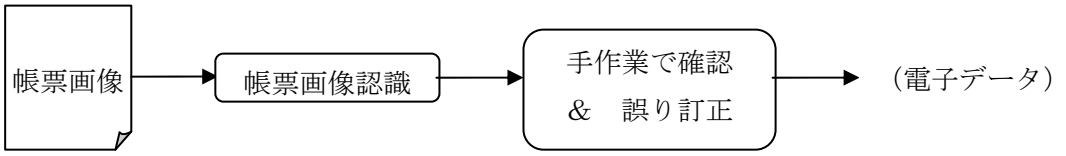
本節では，帳票画像認識技術を実用化するために必要とされる工夫（考慮すべき点）について述べる．帳票画像認識技術は，主に業務用帳票のデータ入力作業の効率化のために用いられるが，業務用データには誤りが少ないことが求められる．帳票画像認識には認識誤りが含まれるため，入力データの誤りを発見し，修正する手段が必要とされる．また，データ入力作業は専門の業者に依頼することが多く，データエントリ業務を専門に行う業者も多数存在するが，帳票を扱う担当部署の外部に帳票データが渡される場合，帳票データの内容が漏洩するリスクが生じるため，セキュリティ面での配慮が必要とされる．このように，帳票画像認識技術を実際に用いるために必要な誤り訂正手段と，セキュリティ対策について述べる．

6.1.1. オペレータによる誤り訂正

従来，帳票に書かれたデータを電子データに変換する作業は，**図 6-1(a)**に示すように手作業で行われていた．作業効率を向上するために帳票画像認識技術を用いた場合でも，専門業者は従来と同等の精度（入力誤りの割合）を維持しなければならないため，認識結果を確認して誤り訂正を行う必要がある．したがって，帳票画像認識を用いた場合でも作業工数がゼロになるわけではなく，**図 6-1(b)**のようにオペレータによる目視確認と誤り訂正の作業は残る．この場合，帳票画像認識は予備入力の手段として位置づけられる．



(a) 従来のデータ入力作業



(b) 帳票画像認識を用いたデータ入力作業

図 6-1 帳票画像認識技術のデータ入力作業への適用

図 6-1(b)のように帳票画像認識を予備入力の手段として用いると，オペレータは認識結果を全て目で確認しなければならず，オペレータに大きな負荷がかかる．また認識誤りはよく似た形状の文字に誤ることが多いため，オペレータが目で確認しても誤りを見逃す可能性が高い．そのため，多くの OCR ソフトウェアでは，**図 6-2** のように確信度が低い文字（認識結果に自信が無い文字）の表示方法を変える（例えば色やフォントを変える）などにより，誤認識の可能性が高い文字を示すものが多い．（文字認識の結果の確信度を求める方法については，例えば文献[1]などのように事後確率を求める方法が提案されている）

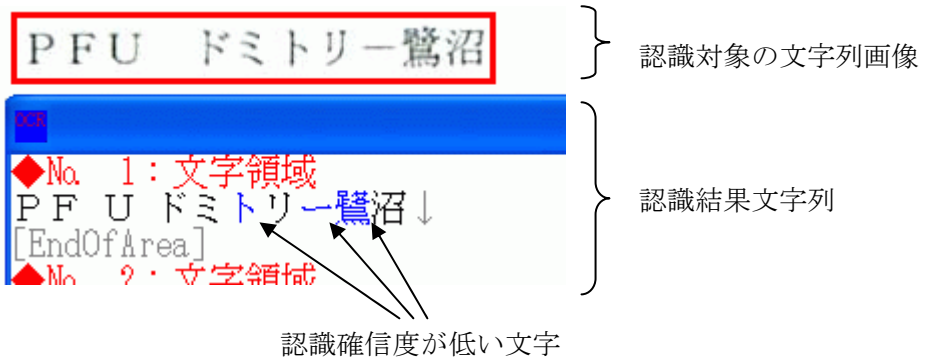


図 6-2 認識確信度による表示形態

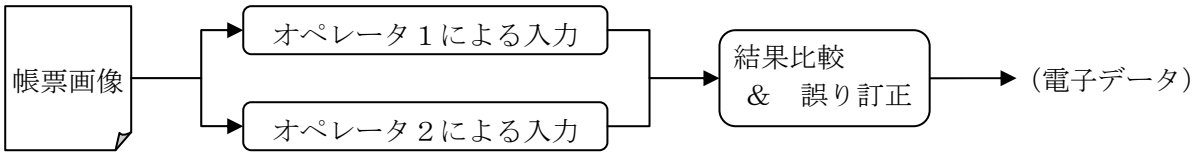
6.1.2. 認識結果の二重チェック

図 6-1 のようにオペレータが最終結果を確認する方法は，オペレータの体調や能力によって結果の品質が左右されるため，専門業者の作業としてはあまり適切ではない．そこで，帳票画像認識技術を用いてデータ品質を高める工夫も必要とされる．

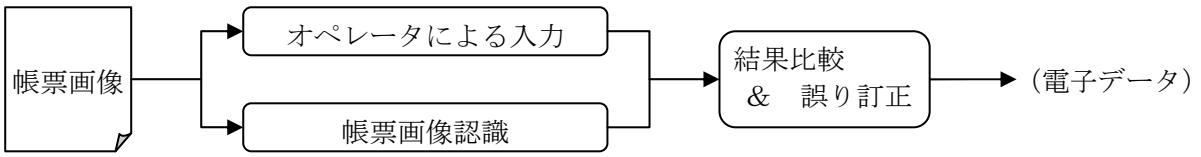
図 6-3 は複数の入力手段による入力データを比較して，相違点を抽出することによって誤りを削減する方法である．オペレータが手作業で入力する場合でも，**図 6-3(a)**のように二人のオペレータが同じデータを独立に入力して結果を比較するということは一般に行われていたが，**図 6-3(b)**のように一人のオペレータを帳票画像認識技術に置き換えれば，作業工数を半分に削減でき，データ入力業務のコスト削減が実現できる．

このような二重チェックが入力データの品質向上に貢献するのは，それぞれの入力手段での誤り傾向が異なるためである．仮に**図 6-3(c)**のように両方とも帳票画像認識に置き換えた場合，同じ認識技術を用いれば誤り傾向も全く同じになるため，データ品質は向上しない．一方で２種類の認識技術の結果が無相関であれば，それぞれの認識誤り確率が p_1, p_2だとすれば，認識誤りがチェックに引っ掛からない確率（つまり最終結果における誤り確率）p は，$p= p_1 \times p_2$で計算できることとなり，入力データの正解率は大幅な改善が期待できる．このように，複数の独立した入力データを統合して誤りを改善する手法は，複数識別器を統合して認識精度を改善する”Multi Expert System”と同じ原理に基づいており，これまでに多くの研究成果が発表されている[1]-[6].

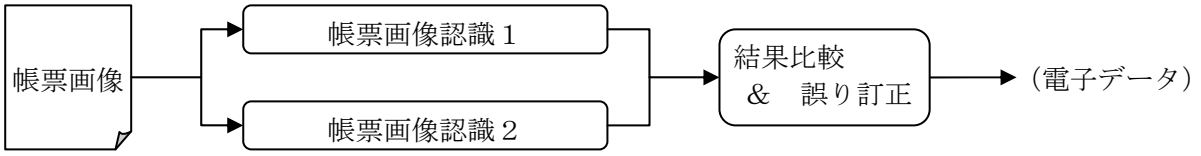
通常，オペレータはキーボードと仮名漢字変換システムを用いて文字列データを入力するので，キーボード上で近くにある文字や，音が似ている文字への誤りが生じやすいが，帳票画像認識は文字の形状が良く似た文字に誤りやすいので，オペレータと帳票画像認識は誤り傾向が異なり，**図 6-3(b)**のような構成は入力データの品質向上に適している．



(a) 従来のデータ入力作業での二重チェック



(b) 帳票画像認識を用いたデータ入力作業での二重チェック



(c) 複数の帳票画像認識を用いた二重チェック

図 6-3 入力データの二重チェックによるデータ品質向上

6.1.3. 入力データの漏洩防止

帳票のデータ入力を外部の専門業者や社内の専門部署に依頼すると，帳票を扱っている担当部署の外部に帳票データが渡されることになる．近年では個人情報保護法などに見られるように，個人情報の漏洩を防止するために情報を必要以上に転送しないようにする必要があり，担当部署の外部に帳票データが渡るのは望ましくない．

しかしながら，紙帳票本体や帳票画像が無ければデータ入力はできないため，帳票画像を部分画像に分解して，それぞれを別のオペレータが入力することによって，帳票内の各項目の関連がオペレータに分からないようにするという対策が取られることが多い．**図 6-4** と**図 6-5** に部分画像によるデータ入力の様子を示す．先ず，帳票に書かれた文字列データを各領域ごとの部分画像に分割する．データ入力を行うオペレータは，それぞれ自分に割り振られた部分画像しか見ることができないため，例えば**図 6-5** のような画像が渡された場合も，「田中宏」という人物が「東京都港区」に住む「会社員」である，というような関連を知ることはできない．個々の文字列そのものには秘密性はほとんど無いため，各領域のデータ間の関係が分からなければ情報漏洩のリスクを大幅に軽減することができる．

このような仕組みはデータ入力を人手で行うか，画像認識によって行うかとは関係なく適用可能である．前節までに述べたような認識結果の確認・修正や，二重チェックによる誤り訂正は，データ入力の処理において適用できる．

また，帳票画像を部分画像に分割する方法についても特に限定は無いが，例えば帳票画像をレイアウト解析してテキスト領域を抽出して，各テキスト領域ごとに部分画像に分割するというような構成を採用すれば，レイアウト解析技術がデータ入力のセキュリティ向上に貢献できるということになる．

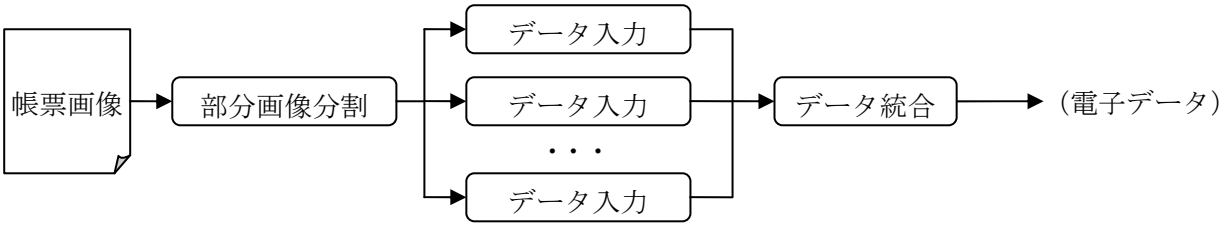


図 6-4 帳票データの分割入力

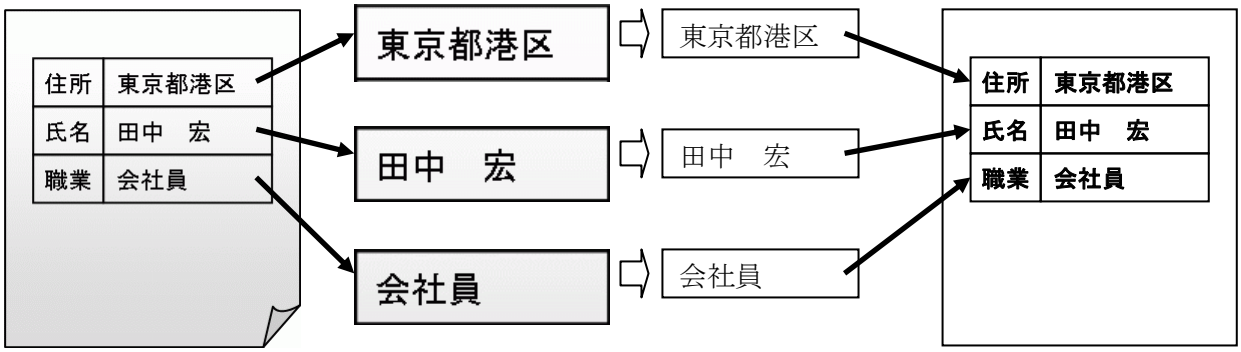


図 6-5 部分画像による分割入力

6.2. 一般文書画像認識への適用

本研究で開発した技術は，帳票画像認識の高精度化を実現するための技術だが，帳票以外の一般文書画像の認識精度向上にも貢献する．本節では，一般文書画像認識を実用化するために必要とされる工夫（考慮すべき点）について述べる．

6.2.1. 帳票画像認識と一般文書画像認識の違い

帳票画像認識は主に業務用帳票のデータ入力作業で用いられるため，認識結果の正確さが求められる．例えば請求書の金額が一桁間違って入力されるなど，業務データに誤りがあると影響が大きいため，先ず十分な精度が要求され，その上で効率が改善できれば採用が可能となる．一方，一般文書画像認識の場合も精度は要求されるものの，業務で利用される帳票データに比べれば誤りに対する許容度は高い．

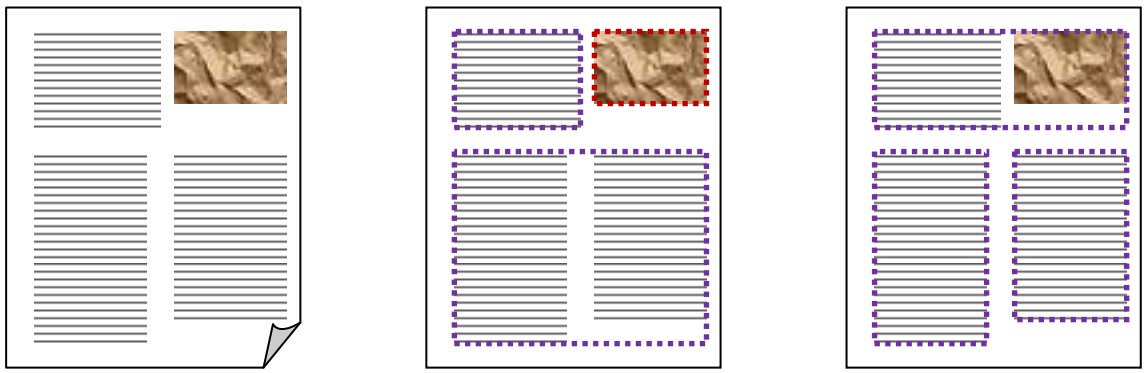
近年，一般文書画像認識の主な用途は二つある．第一は文書画像を認識して電子文書を再構成するという用途であり，第二は認識結果を検索用テキストとして用いる用途である．前者には，例えば文書画像を認識して Word 文書を作成するソフトや，ePub 形式で認識結果を出力するソフトが存在し，紙文書を電子文書に変換して編集したり，認識した文書を音声読み上げで出力するなどの

用途に用いられる．後者は，例えば文書画像を PDF ファイルに変換して，認識したテキストを PDF ファイルの透明テキストとして添付することによって，文書中のデータをテキスト検索で探せるようにする機能である．現在，スキャナ製品のほとんどはスキャンと同時に文書を認識して，テキスト付 PDF ファイルを生成する機能を有している．いずれの用途も，文書中に数文字の誤りがあっても大きな問題にはならないことが多い．Word 文書に変換して再利用する際は，ユーザがデータを編集する際に用いるため，若干の誤りがあってもその場で修正すれば良い．検索用のテキストの場合は，あいまい検索の機能を用いれば若干の誤りを含むテキストでも検索することができる．

6.2.2. 電子文書の再構成

文書画像を認識して，Word などの電子文書を再構成する機能は，従来からパッケージ型 OCR ソフトウェア（スキャナ等に添付されるのではなく単独で市販されているソフトウェアのこと）の多くが持っている機能である．例えば筆者が勤務している富士通グループからは，「文書 OCR for Word」，「文書 OCR for Excel」などのソフトウェアが販売されていた（2002 年に販売終了）．近年では「e.Typist v.14.0」（メディアドライブ株式会社：2012 年 4 月発売）や，「ABBYY FineReader 11」（ABBYY 社：2011 年 12 月発売）などが代表的な OCR ソフトウェアとして知られている．これらのソフトウェアは Word 文書だけでなく，電子書籍の標準形式である EPUB 形式[7]で電子ファイルを出力する機能も有する．

文書認識の結果を電子文書として出力すると，ユーザは僅かな認識誤りにも気付くため，ユーザの満足度は認識精度に大きく依存する．したがって，認識誤りは業務利用の場合ほど深刻な問題ではないとはいえ，ソフトウェアの使い勝手を左右する重要な要件となる．また，変換した Word 文書を再利用する場合には，テキスト領域，図表領域などの部分領域ごとに編集することが多く，レイアウト解析の精度がより重要である（例えば**図 6-6** のようにテキスト領域に誤りがあると，文章の順序がおかしくなるなど，文書データの利用に支障が大きい）．



(a) 文書画像　(b) テキスト領域の誤統合　(c) テキストと図の誤統合

図 6-6 レイアウト解析の誤り

EPUB 形式で出力された電子文書は，PC 上の電子書籍リーダや，タブレット型 PC などの電子書籍端末で読むことができる．また，EPUB の最新仕様である EPUB3 は，デジタル録音図書の国際標準規格である DAISY が持つ全ての機能を包含するように設計されており，EPUB3 形式の電子

文書は音声読み上げの入力データとしても用いられる．EPUB 文書を電子書籍端末で読む場合，端末の表示サイズに合わせて表示レイアウトを変更する，リフロー（reflow）形式で文書が表示されるため，レイアウト解析の正確さは非常に重要である．また EPUB 文書を音声で読み上げる場合も，正しい順序でテキストを並べる必要があるため，正確なレイアウト解析と読み順（リーディングオーダー）の設定は重要である．

このように，文書画像認識の結果を Word や EPUB などの電子文書で利用する場合には，レイアウト解析の正確さが非常に重要である．本研究ではレイアウト解析の精度向上については特に検討していないが，一般文書画像を認識して電子文書として活用することを考えるのであれば，レイアウト解析についての研究も深める必要がある．

6.2.3. 検索用テキストの付与

文書画像を認識して取得したテキストを PDF ファイルなどに添付して検索用のテキストとして用いる場合，これまでに見てきた用途に比べて認識誤りに対する許容度は高い．近年のテキスト検索は，部分的に一致しない文字があってもあいまい検索によって良く似たテキストを見つけることができるため，誤認識テキストが検索精度に及ぼす影響はさほど大きくないことが知られている．

また，文字列認識の途中経過である認識候補ラティスを保持しておき，ラティス中から検索キーワードに近い部分文字列を探すという方法もある[9]-[11]. **図 6-7** は文献[9]に例示されていた認識候補ラティスである．このように，文字認識の 2 位以下の候補や，複数の文字境界候補を含んだ情報が電子文書に保持されていれば，検索のヒット率をより高めることができる．そのため，文字認識の正解率が若干低くても，テキスト検索という目的では実用になることが多い．

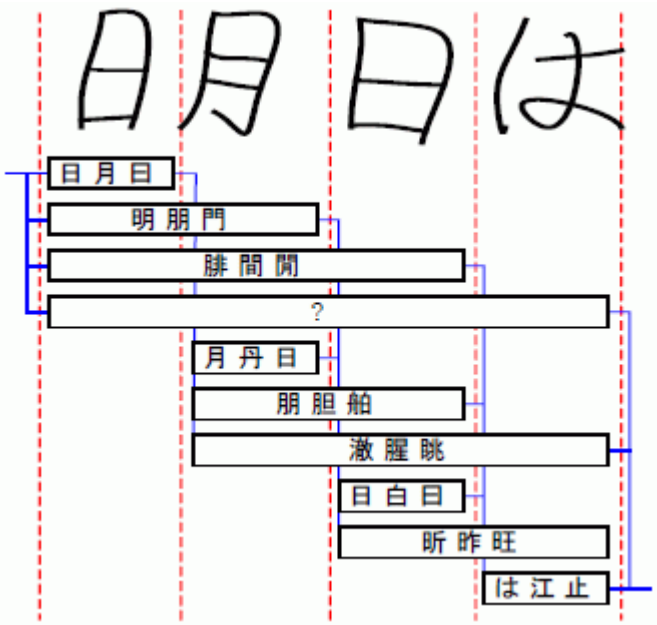


図 6-7 認識候補ラティス（文献[9]の図 3 より）

ただし，文書画像を電子化して用いる場合に最もよく使われる PDF ファイルには，このような認識候補ラティスを保持して検索に用いるという仕様は含まれていない．そのため電子文書として PDF ファイルを用いる場合は, 認識候補ラティスを用いた検索技術を用いることはできない. また, OCR ソフトウェアの中には，文書画像認識の結果としては通常のテキストしか出力せず，認識候補ラティスは OCR ライブラリの内部でしか利用できないようなものも存在する．

このように，文書画像認識の結果をテキスト検索で用いる場合には，認識精度が若干低くても実用化するための技術は存在している．しかし，その技術を十分に発揮するためには，認識候補ラティスのような特殊なデータ形式を利用できるためのファイル仕様や環境が必要である. 例えば PDF ファイルの場合，非公式な付加情報をファイルに追加できる仕様は存在するので，認識候補ラティスを独自形式でファイルに保存することによって検索精度をより高められる可能性がある．したがって，文書画像認識によるテキスト検索の課題は，高度な検索技術を活用できる仕様やソフトウェア実行環境の整備だと言える．

6.3. 開発技術の実用化状況について

本章では, 本研究で開発した技術を実用化するために考慮すべき点について述べている. そこで, 参考までに開発した技術がこれまでにどのような製品，サービスに適用されたかについて，簡単に紹介しておく．

著者は現時点で株式会社富士通研究所に所属しており，研究成果が富士通グループの製品に適用される立場にいる．本研究の成果も，その多くは富士通製品に搭載され，製品（主にイメージスキャナ）やソリューションの一部として実用化されている．商談などに適用された事例については業務上の規約により明らかにできない部分もあるが，ほとんどの場合，パッケージ製品へ組み込まれたものが商談などに適用されるため，適用の有無は公開された範囲で記述することができる．

業務向け用途での実用化の例としては，1997 年に発売開始した帳票処理アプリケーション「AutoEntry」に開発技術が適用されている[12]．発売は 1997 年で，当初から Press Release には「帳票レイアウト認識技術」が採用されているとの記述があるが，本研究で開発した技術は，搭載技術の高精度化に貢献している．製品の内部情報となるため，どの技術要素がいつ頃に搭載されたか，などの詳細については残念ながら明らかにはできないが，帳票画像認識技術が主に流通業，卸業，金融業など，帳票を扱う業務全般に適用され，業務改善に貢献しているという点については明言できる．

一般文書画像認識としての実用化例としては，富士通グループのイメージスキャナ「fi シリーズ」（業務用），「ScanSnap シリーズ」（個人用）にテキスト検索用キーワードを付与する技術として用いられている．帳票に限らず様々な文書がスキャンされるため，本研究で開発した表認識や二値化技術のみではなく，多彩な文書画像に対応した画像処理技術や，高精度なレイアウト解析技術も必要とされ，今後も更なる研究開発が必要となる用途である．このような汎用スキャナの市場は，近年大きく成長し続けており（付録Ｂ参照），企業の業績だけでなく，世の中への貢献という点で非常に重要な技術分野だと言える．

6.4. まとめ

本章では，本研究で開発した帳票画像認識技術を実用化する際に，開発技術の他に考慮すべきポイントについて述べた．開発技術は帳票画像認識のための技術だが，それは一般文書画像認識の精度改善にも貢献するため，帳票画像認識と，一般文書画像認識のそれぞれについて記した．

帳票画像認識は，主に業務データを入力するために用いられるため，入力データに誤りが少ないことが最も重要である．現在の帳票画像認識は十分な精度が得られているとは限らないため，業務で利用するためには人手による確認作業が必要とされる．その作業の負荷を軽減し，また確認作業の品質を向上させるための技術が実用化のために必要であることを本章で述べた．

続いて，一般文書画像認識の主な用途として，文書画像を認識して電子文書を再構成するという用途と，認識結果を検索用テキストとして用いるという用途があることを示した．前者の用途の場合，電子文書の編集・再利用や，音声読み上げなどに利用することを考えると，レイアウト解析技術の高度化が求められることを記した．また後者については，あいまい検索の精度を高めるための技術が既に提案されているが，その技術を有効活用するためのファイル仕様やプログラム実行環境の整備が必要であることを記した．

ここに記したように，開発技術を実用化するためには，技術の用途に応じて新たな工夫が必要とされることがある．単に認識精度向上のための技術を開発するだけでなく，それが活用される状況を考慮して，更なる検討を進める必要があるように思われる．

7. 結論

概要

本章では，本研究の成果についてまとめ，本論文を総括する．

7.1. 本研究の成果

本研究では，文書画像認識技術の主要な適用先である，帳票画像からのデータ入力作業の効率化を実現するために，帳票画像認識の高精度化を実現するための開発技術について述べた．第 2 章の背景で述べたように，表領域が文書の大半を占める帳票画像は，一般的な文書画像の中ではレイアウトが比較的シンプルである．そのため，帳票画像認識の高精度化のためには，レイアウト解析よりも，表画像認識とテキスト認識の高精度化がより重要である．本研究では，このような視点に基づいて，三つの要素技術の高精度化に取り組んだ．

第一は，表を構成する罫線を高精度に抽出する，罫線抽出技術の高精度化である．第二は，抽出した罫線に基づいて，表を構成するセル領域を抽出する，セル抽出技術の高精度化である．第三は，テキスト認識の精度を改善するために，その前処理である文字画像の二値化技術の改良である．これらの開発技術の詳細については第 3 章から第 5 章において記述し，また個々に評価実験によって有効性の検証を行った．

罫線抽出技術の研究（第 3 章）においては，近年増加する多様なデザインの帳票画像から罫線を抽出する技術を開発した．近年の帳票画像では，典型的な実線罫線だけでなく，点線による罫線や，一様色で塗られた領域の境界が罫線として用いられたもの（境界罫線），一様な模様領域の境界による罫線（テクスチャ境界罫線）などのように様々な種類の罫線が用いられる．また多くの色が使われた多彩な文書画像や，デジタルカメラで取得した低品質の画像のように，罫線抽出が困難な画像が増加している．このような状況に対応するために，様々な種類の罫線を高い精度で認識できる罫線抽出技術を開発した．本技術では，罫線抽出誤りの主な原因として罫線画素の途切れによる罫線脱落誤りと，文字と罫線の混同による罫線付加誤りを改善する方法を提案し，罫線抽出の精度向上を実現した．

セル抽出技術の研究（第 4 章）においては，複雑な構造を持つ表画像から，表を構成するセル領域を抽出する技術を開発した．従来の主なセル抽出技術ではセル領域の同定が困難であった，非矩形のセル領域を含む表であっても，また罫線抽出の結果に誤りが生じたためにセル領域が不自然な形状となってしまった場合でも，罫線が交差する交点の情報を用いることによってセル領域を頑強に抽出することができる．また，罫線抽出の結果に曖昧性が残る場合があるため，セル領域も複数の候補を生成して曖昧性を保持し，最終的に組み合わせ探索アルゴリズムを用いて最適なセル集合を生成するという，新しいアプローチによるセル抽出技術を開発した．

文字抽出用二値化技術の研究（第 5 章）においては，二値画像を対象とした文字認識アルゴリズムを後段で用いることを前提として，多値の文字画像を二値化する技術を改善することによって，文字認識の精度向上を実現する技術を開発した．二値化画像の品質低下による文字認識誤りのうち，特に目立つのは低解像度画像において文字の一部が途切れてしまう現象による誤りである．本研究では，文字の途切れが局所的二値化の閾値変動に起因するものであることを示し，第 3 章で述べた罫線画素の途切れ対策に類似した閾値補正手法によって，文字の途切れが解消できることを示した．

また，文字のストローク幅に応じて大域的二値化（大津二値化）と局所的二値化（Niblack 二値化）を併用することによって，更に二値化の精度が改善できることを示した．本章では，二値化の品質を表す指標として文字認識の精度を基準として文字認識精度を用い，文字認識の精度評価によって本技術の有効性を示した．

これら三つの要素技術によって，帳票画像認識の高精度化が実現できる．ただし，開発技術を実用化するためには，技術の用途を踏まえた更なる工夫も必要となる．第 6 章において，帳票画像認識技術の実用化に伴う様々な工夫（周辺技術や考慮すべき点）についての考察を行った．

7.2. 本論文の結論

本論文では，紙帳票に書かれた文字列データを計算機に効率良く入力するために必要な，帳票画像認識技術の高精度化について述べた．近年の帳票はユーザにとって読みやすいデザインが用いられることが多く，計算機による自動認識にかかる負荷が増大する傾向にある．しかし，この傾向を逆に理解すれば，帳票画像認識の技術が高度化すればユーザにとっての利便性が向上するという因果関係がより明確になりつつある，という状況を示しているのだとも解釈できる．帳票画像認識は依然として大量に使用されている紙帳票を電子化して業務で活用するために重要な技術であり，認識精度の向上が業務の効率化に直結する．本研究が目指す，多彩かつ多様な帳票画像を高精度で認識する技術は，業務の効率化とユーザにとっての使いやすさを両立するために重要な技術である．

近年では，文書画像認識（OCR）技術はデータ入力作業のためだけでなく，紙文書をスキャンして電子化する際に検索用キーワードを自動付与する目的で利用されるなどのように，用途が拡大しつつある．第 2 章でも述べたように，帳票以外の一般文書の中でも表は頻繁に使われており，本研究で検討した表認識技術の高精度化や，テキスト認識のための二値化技術の高精度化は，一般文書の認識においても重要な技術である．例えば，雑誌，新聞，パンフレットなどのように複雑な文書を扱うことができるレイアウト解析技術と組み合わせれば，様々な種類の紙文書がスキャンされる一般ユーザ向けスキャナなどの市場においても，本研究の成果は有効に活用することができる．

今後は，開発した技術の更なるブラッシュアップを進めると共に，帳票画像だけに留まらず様々な文書画像に対しても開発技術を適用したいと考えている．そのためには，本研究では詳しく検討しなかったレイアウト解析技術についても調査を進め，複雑なレイアウトの文書でも解析可能な方式を採用する必要がある．実際には，第 2 章で述べたように多くのレイアウト解析手法はマンハッタンレイアウトなどの単純な構造の文書を主に扱うアルゴリズムとなっており，必要があれば新たなレイアウト解析技術を開発しなければならない. そのためには, どのような文書を解析したいか, という利用シーンの検討から始めて最適なアプローチの選択を行うべきだと考える．

現在，筆者は一般企業において文書画像認識技術を製品に適用する業務を行っているが，一般ユーザを対象とすると，実に様々な種類の文書が電子化されていることを実感する．インターネット等での製品レビューを見ると，文字認識の精度が低く使い勝手が悪いという意見（苦情）も散見する．文書画像認識の研究者としては忸怩たる想いは否定できない．一般ユーザ向け製品に近い業務に携わっているという立場の利点を活かし，ユーザの声に真摯に耳を傾けることにより，紙文書と電子文書を繋ぐ本当の架け橋となる技術開発を進めていきたいと考えている．

謝辞

筆者が手書き文字認識の研究を行っていた 10 年以上前から継続的に多くのご指導と叱咤をいただき，本研究の推進を支えてくださった，東京農工大学，中川正樹教授に深く感謝いたします．また中川研究室に在籍中に，学内での研究生活において様々なご助力を頂いた，桜美林大学，耒代誠仁専任講師（当時，東京農工大学准教授）に感謝いたします．

研究を進めるにあたり，株式会社富士通研究所の文書メディア認識グループのメンバー各位には様々な形でお世話になりました．特に，本研究を始めるにあたって研究テーマの設定や方向付けにおいて多大なご助力を頂いた，藤本克仁所長付，堀田悦伸主任研究員に感謝いたします．更に，本研究の推進にあたっては，筆者が過去に在籍していたペン入力インタフェースグループにおいて習得した知識，ノウハウが大きな支えとなっておりました．当時の上司である石垣一司主席研究員，同僚であった中島健次研究員，秋山勝彦研究員を始め，多くの元同僚，業務上のパートナーに深く感謝いたします．

最後に，これまで私を支えてくださった両親，兄，そして妻に感謝いたします．

本当にありがとうございました．

本研究に関する発表

論文誌（査読あり）

[1] 田中　宏，武部浩明，藤本克仁，中川正樹，“交点追跡と全体最適化に基づく罫線抽出誤りに頑強な表項目セル抽出，”信学論 D, Vol.J94-D, No.7, pp.1113-1124 (2011.7)　（第四章）

[2] 田中　宏，藤井勇作，堀田悦伸，中川正樹，“対象知識を利用した文書画像の二値化，”信学論 D, 投稿中　（第三章，第五章）

国際会議（査読あり）

[1] H. Tanaka, “Threshold Correction of Document Image Binarization for Ruled-line Extraction,” Proc. 10th ICDAR, pp.541-545, Barcelona, Spain, Jul. 2009.　（第三章）

[2] H. Tanaka, Y. Fujii, and Y. Hotta, “Threshold Correction of Document Image Binarization for Text Extraction,” Proc. 6th VISAPP, Vilamoura, Portugal, Mar. 2011.　（第五章）

[3] H. Tanaka, H. Takebe, and Y. Hotta, “Robust Cell Extraction Method for Form Documents based on Intersection Searching and Global Optimization,” Proc. 11th ICDAR, pp.354-358, Beijing, China, Sept. 2011　（第四章）

国際会議（招待講演）

[1] H. Tanaka, “An Overview of Document Image Recognition - Layout and Logical Structure Analysis -,” Proc. WEIMS’09 (The Workshop on E-Inclusion in Mathematics and Science 2009), pp.6-9, Fukuoka, Japan, Dec. 2009.　（第二章）

その他（査読なし）

[1] 田中　宏，武部浩明，藤本克仁，“複数セル候補の組み合わせ探索に基づく帳票画像からのセル抽出，”信学技報，PRMU2005-185，Feb. 2006.　（第四章）

[2] 田中　宏，中島健次，武部浩明，藤本克仁，“テクスチャ領域を含む帳票画像からの罫線抽出，”信学技報，PRMU2006-264，Mar. 2007.　（第三章）

[3] 田中　宏，藤井勇作，武部浩明，藤本克仁，“二値化閾値の補正と罫線形状判定による罫線抽出の高精度化，”信学技報，PRMU2007-216，Feb. 2008.　（第三章）

[4] 田中　宏，藤井勇作，堀田悦伸，“二値化閾値の補正による低解像度画像に頑強な文字抽出用二値化，”信学技報，PRMU2010-254，Mar. 2011.　（第五章）

本研究に関する特許

登録済

[1] 田中　宏, 中島健次, 武部浩明, 藤本克仁，“表認識装置、及びコンピュータプログラム”，国内，第 4, 628, 278 号，2010 年 11 月 19 日登録　（第四章）

[2] Hiroshi Tanaka，“Table data processing method and apparatus”，中国，ZL200610171447.0，2010 年 5 月 19 日登録　（第四章）

出願中

[1] 田中　宏，“表データ処理方法及び装置”，国内，特開 2008-046812，2008 年 2 月 28 日公開　（第四章）

[2] 田中　宏，中島健次，皆川明洋，武部浩明，藤本克仁，“表認識プログラム、表認識方法および表認識装置”，国内，特開 2008-198157，2008 年 8 月 28 日公開　（第三章）

[3] 田中　宏，藤井勇作，藤本克仁，“画像処理プログラム、画像処理装置、および画像処理方法”，国内，特開 2009-105768，2009 年 5 月 14 日公開　（第三章，第五章）

参考文献

第二章　（背景技術と課題）

[1] H. Tanaka, “An Overview of Document Image Recognition - Layout and Logical Structure Analysis -,” Proc. WEIMS’09 (The Workshop on E-Inclusion in Mathematics and Science 2009), pp.6-9, Fukuoka, Japan, Dec. 2009.

[2] M. Agrawal, and D. Doermann, “Voronoi++: A Dynamic Page Segmentation approach based on Voronoi and Docstrum features,” Proc. 10th ICDAR, pp.1011-1015, Barcelona, Spain, Jul. 2009.

[3] K.Y. Wong, R.G. Casey, and F.M. Wahl, “Document Analysis System,” IBM Journal of Research and Development, 26-6, pp.647-656, Nov. 1982.

[4] G.Nagy, S. Seth, and M. Viswanathan, “A Prototype Document Analysis System for Technical Journals,” Computer, 25-7, pp.10-22, Jul. 1992.

[5] H.S. Baird, “Background structure in document images,” Advances in Structural and Syntactic Pattern Recognition, pp.17-34, World Scientific, 1994

[6] T.M. Breuel, “Two geometric algorithms for layout analysis,” Proc. 5th DAS, pp.188-199, Princeton, USA, Aug. 2002.

[7] L.A. Fletcher, and R. Kasturi, “A Robust Algorithm for Text String Separation from Mixed Text/Graphic Images,” IEEE PAMI vol.10 No.6, pp.910-918, Nov. 1988

[8] L. O’Gorman, “The Document Spectrum for Page Layout Analysis,” IEEE PAMI vol.15 No.11, pp.1162-1173, Nov.1993.

[9] A. Antonacopoulos, and R.T. Ritchings, “Flecible Page Segmentation Using the Background,” Proc. 12th ICPR vol.2, pp.339-344, Jerusalem, Israel, Oct. 1994.

[10] K. Kise, M. Iwata, and K. Matsumoto, “On the Application of Voronoi Diagrams to Page Segmentation,” Proc. DLIA99, IV-c, Bangalore, India, Sept. 1999

[11] 黄瀬浩一，佐藤昭則，“一般図形ボロノイ図を用いた文書画像の領域分割，”信学技報 PRMU96-181, Mar. 1997

[12] 石谷康人，“データ駆動型処理と概念駆動型処理の相互作用による文書画像レイアウト解析，” 情処論 42 巻 11 号, pp.2711-2723, Nov. 2001

[13] 武部浩明，小澤憲秋，藤本克仁，勝山裕，直井聡，“仮説検証に基づく再帰的テキストブロック抽出手法，”信学全大講論集 2004 年, D-12-42, pp.208, Mar. 2004

[14] 武部浩明，小澤憲秋，藤本克仁，勝山裕，直井聡，“複数の処理結果統合によるテキストブロック抽出手法，” 信学全大講論集 2005 年, D-12-65, pp.215, Mar. 2005

[15] F. Deckert, B. Seidler, M. Ebbecke, and M. Gillmann, “Table Content Understanding in smartFIX,” 11th ICDAR, pp.488-492, Beijing, China, Sept. 2011.

[16] F.Cesarini, M.Gori, S.Marinai, and G.Soda, “INFORMys: A Flexible Invoice-Like Form-Reader System,” IEEE Trans. PAMI, vol.20, no.7, pp. 730-745, July 1998.

[17] D.W.Embley, D.Lopresti, and G.Nagy, “Notes on Contemporary Table Recognition,” Proc. 7th. DAS, pp. 164-175, Feb. 2006.

[18] F. Shafait, and R. Smith, “Table Detection in Heterogeneous Documents,” Proc. 9th DAS, pp.65-72, Boston, USA, Jun. 2010.

[19] 中野康明，藤澤浩道，国崎　修，岡田邦弘，花野井歳弘，“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, Mar. 1986

[20] R.O. Duda, and P.E. Hart, “Use of the Hough Transformation to Detect Lines and Curves in Pictures,” Communications of the ACM, Vol.15, No.1, pp.11-15, Jan. 1972.

[21] S.W. Lam, L. Javanbakht, and S.N. Srihari, “Anatomy of a FormReader,” Proc. 2nd ICDAR, pp.506-509, Tsukuba, Japan, Oct. 1993.

[22] 直井　聡，矢吹眞紀，浅川敦子，堀田悦伸，“ＧＩＭ法による枠接触文字の高品位分離，” 信学技報 PRU93-25, Jul. 1993

[23] 小原敦子，藤本克仁，直井　聡，“非接触入力による濃淡画像からの罫線抽出方式，”信学全大

情報・システムソサイエティ大会講演論文集, pp.206, Sep. 2000

[24]浅野三恵子，下辻成佳，“セル構造を用いた帳票識別，” 信学技報，PRU95-61，July 1995.

[25]J.Yuan, L.Xu, and C-Y.Suen, “Form Items Extraction By Model Matching,” Proc. 1st. ICDAR, pp. 210-218, Sept. 1991.

[26]児島治彦，清末悌之，秋山照雄，"複雑な構造を持つ表の認識に関する基礎検討，" 情処学第37回全大，6W-8，pp. 1660-1661，Oct. 1988.

[27]長谷博行，辻　正博，園田浩一郎，米田政明，酒井　充，“汎用を目指した自動文書画像認識システムー要素抽出技術の問題点と検討ー，” 信学技報，PRU94-33，Sep. 1994.

[28]駱 琴，渡辺豊英，杉江　昇，“多種帳票文書の構造認識，” 信学論(D-II), vol. J76-D-II, no.10, pp.2165-2176, Aug. 1993.

[29]新庄　広，高橋寿一，古川直広，“DP マッチングを用いた帳票枠構造照合方式，” 信学技報 PRMU2002-228，Mar. 2003.

[30]H.Shinjo, E.Hadano, K.Marukawa, Y.Shima, and H.Sako, “A Recursive Analysis for Form Cell Recognition,” Proc. 6th. ICDAR, pp. 694-698, Sept. 2001.

[31]寺沢憲吾，長崎健，川嶋稔夫，“固有空間法と DTW による古文書ワードスポッティング，”信学論(D), vol.J89-D, no.8, pp.1829-1839, Aug. 2006.

[32]村瀬　洋，若原　徹，梅田三千雄，“候補文字ラティス法による枠無し筆記文字列のオンライン認識，” 信学論(D), vol. J68-D, no.4, pp. 765-772, Apr. 1985.

[33]T.Matsui, I.Yamashita, and T.Wakahara, “The Results of the First IPTP Character Recognition Competition and Studies on Multi-Expert Recognition for Handwritten Numerals,” IEICE Trans. Inf.&Sys., vol.E77-D, No.7, pp.801-809, Jul. 1994.

[34]T.Tsutsumida, T.Matsui, and T.Noumi: "Results of IPTP Character Recognition Competitions and Studies on Multi-expert System for Handprinted Numeral Recognition," IEICE Trans. Inf.&Sys., vol.E79-D, No.5, pp.429-435, May 1996.

[35]岡　隆一, “セル特徴を用いた手書き漢字の認識，”信学論(D), vol.J66-D, no.1, pp.17-24, Jan. 1983.

[36]鶴岡信治，栗田昌徳，原田智夫，木村文隆，三宅康二，“加重方向指数ヒストグラム法による手書き漢字・ひらがな認識，” 信学論(D), vol.J70-D, no.7, pp.1390-1397, Jul. 1987.

[37]萩田紀博, 内藤誠一郎, 増田　功, “外郭方向寄与度特徴による手書き漢字の識別，” 信学論(D), vol.J66-D, no.10, pp. 1185-1192, Oct. 1983.

[38]若林哲史，鶴岡信治，木村文隆，三宅康二，“手書き数字認識における特徴選択に関する考察，”信学論(D-II), vol.J78-D-II, no.11, pp.1627-1638, Nov. 1995.

[39]鈴木雅人，大町真一郎，加藤　寧，阿曽弘具，根元義章，“混合マハラノビス識別関数による高精度な類似文字識別手法，”信学論(D-II), vol.J80-D-II, no.10, pp.2752-2760, Oct. 1997.

[40]藤澤祥治，澤　和弘，若林哲史，木村文隆，三宅康二，“濃度こう配の方向と曲率を用いた手書き数字認識（その２），”信学技報 PRMU97-228, Feb. 1998.

[41]J. Sun, Y. Hotta, Y. Katsuyama, and S. Naoi, “Camera based Degraded Text Recognition Using Grayscale Feature,” Proc. 8th ICDAR, pp.182-186, Seoul, Korea, Aug. 2005.

[42]Y. Hotta, J. Sun, Y. Katsuyama, and S. Naoi, “Robust Chinese Character Recognition by Selection of Binary-Based and Grayscale-Based Classifier,” Proc. 7th DAS, pp.553-563, Nelson, New Zealand, Feb. 2006.

[43]N. Otsu, “A Threshold Selection Method from Gray-level Histograms”, IEEE Trans. Systems, Man, and Cybernetics, vol.9, no.1, pp. 62-66, 1979.

[44]J. Kittler, and J. Illingworth, “Minimum error thresholding,” Pattern Recognition, Vol.19, Issue 1, pp.41-47, 1986.

[45]O.D.Trier and A.K.Jain, "Goal-Directed Evaluation of Bainarization Methods," IEEE Trans. PAMI. vol.17, no.12, pp.1191-1201, Dec. 1995.

[46]W. Niblack, “*An Introduction to Digital Image Processing*,” Prentice Hall, Englewood Cliffs, N.J., pp.

115-116, 1986.

[47]J.Sauvola, T.Seppanen, S.Haapakoski and M.Pietikainen, "Adaptive Document Binarization," Proc. 4th. ICDAR, pp.147-152, Ulm, Germany, Aug.1997.

[48]H.Kamada, and K.Fujimoto, “High-speed, High-accuracy Binarization Method for Recognizing Text in Images of Low Spatial Resolutions,” Proc. 5th ICDAR, pp.139-142, Bangalore, India, Sept. 1999.

[49]藤本克仁, 鎌田　洋, “低解像度カラー文書画像から高品質な文字画像を抽出する二値化方式,” 信学技報 PRMU99-89, Oct. 1999.

[50]T. Kasar, and AG Ramakrishnan, “COCOCLUST: Contour-based Color Clustering for Robust Binarization of Colored Text,” Proc. The Third CBDAR, pp.11-17, Balcerona, Spain, Jul. 2009.

第三章　（罫線抽出技術の研究）

[1] 中野康明，藤澤浩道，国崎　修，岡田邦弘，花野井歳弘，“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, Mar. 1986.

[2] 成瀬博之，渡辺豊英，駱　琴，杉江　昇，“枠罫線情報を用いた帳票文書の構造認識,” 信学論(D-II), vol.J75-D-II, no.8, pp.1372-1385, Aug. 1992.

[3] 広瀬克昌，明　偉，馬場口登，北橋忠宏，“レイアウト解析と文字認識に基づく文書画像のメディアコンバージョン,” 信学技報 PRMU99-224, Feb. 2000.

[4] 長谷博行，辻　正博，園田浩一郎，米田政明，酒井　充，“汎用を目指した自動文書画像認識システムー要素抽出技術の問題点と検討ー,” 信学技報 PRU94-33, Sep. 1994.

[5] R.M.Haralick, K.S.Shanmugam, N.Distein, “Texture Fertures for Image Classification”, IEEE Trans. SMC Vol.3, pp.610-621, Mar. 1973.

[6] J. Canny; “A Computational Approach to Edge Detection,” IEEE Trans. PAMI Vol.8, No. 6, Nov. 1986

[7] 小原敦子，藤本克仁，直井　聡，“非接触入力による濃淡画像からの罫線抽出方式,” 信学全大 情報・システム, pp.206, Sep. 2000.

[8] W.Niblack, “An Introduction to Digital Image Processing,” Englewood Cliffs, N.J.: Prentice Hall, pp.115-116, 1986.

[9] H-C.Lee: “Detecting Boundaries in a Vector Field,” IEEE Trans. SP Vol.39, No.5, pp.1181-1194, May 1991.

[10]原島　博，小田島薫，鹿喰善明，宮川　洋，“ε-分離非線形ディジタルフィルタとその応用,” 信学論(A), vol.J65-A, no.4, pp.297-304, Apr. 1982.

[11]石寺永記，荒井祐之，土屋雅彦，宮内裕子，高橋信一，栗田正一，“主観的輪郭の形成に関する視覚情報処理モデル,” 信学論(D-II), vol.J76-D-II, no.4, pp.873-880, Apr. 1993.

[12]藤本克仁, 鎌田　洋, “低解像度カラー文書画像から高品質な文字画像を抽出する二値化方式,” 信学技報 PRMU99-89, Oct. 1999.

[13]O.D.Trier and A.K.Jain, “Goal-Directed Evaluation of Bainarization Methods,” IEEE Trans. PAMI. vol.17, no.12, pp.1191-1201, Dec. 1995.

[14]J.Sauvola, T.Seppanen, S.Haapakoski and M.Pietikainen, “Adaptive Document Binarization,” Proc. 4th. ICDAR, pp.147-152, Ulm, Germany, Aug.1997.

[15]L.Eikvil, T.Taxt and K.Moen, “A Fast Adaptive Method for Binarization of Document Images,” Proc. 1st. ICDAR, pp.435-443, Saint-Malo, France, Sep. 1991.

[16]小原敦子，直井　聡，江口真一，勝又　裕，“ラベル画像の階層的矩形表現を用いた罫線抽出方式,” 信学全大 情報・システム, No.2, pp.240, Mar. 1998.

第四章　（セル抽出技術の研究）

[1] 石谷康人，“モデルマッチングによる表形式文書の理解,” 信学技報，PRU94-34，Sep. 1994.

[2] 浅野三恵子，下辻成佳，“セル構造を用いた帳票識別,” 信学技報，PRU95-61，Jul. 1995.

[3] 成瀬博之，渡辺豊英，駱　琴，杉江　昇，“枠罫線情報を用いた帳票文書の構造認識，” 信学論(D-II), vol. J75-D-II, no.8, pp. 1372-1385, Aug. 1992.

[4] 駱　琴，渡辺豊英，杉江　昇，“帳票文書の構造認識のための書式構造知識の自動獲得，” 信学論(D-II), vol. J76-D-II, no.3, pp. 534-546, Mar. 1993.

[5] 平野　敬，岡田康裕，依田文夫，“ロバストなモデル照合に基づく FAX 送信された一般帳票の読取り，” 信学論(D-II), vol. J85-D-II, no.9, pp.1371-1381, Sep. 2002.

[6] 新庄　広，高橋寿一，古川直広，“DP マッチングを用いた帳票枠構造照合方式，” 信学技報，PRMU2002-228，Mar. 2003.

[7] S.W.Lam, L.Javanbakht, and S.N.Srihari, “Anatomy of a FormReader,” Proc. 2nd ICDAR, pp. 506-509, Oct. 1993.

[8] K-C.Fan, Y-K.Wang, and M-L.Chang, “Form Document Identification Using Line Structure Based Features,” Proc. 6th. ICDAR, pp. 704-708, Sep. 2001.

[9] L.A.D.Hutchison, and W.A.Barrett, “Fast Registration of Tabular Document Images Using the Fourier-Mellin Transform,” Proc. 1st.DIAL, pp. 253-267, Jan. 2004.

[10] 皆川明洋，藤井勇作，武部浩明，藤本克仁，“確率伝搬法を用いた帳票の論理構造認識に関する一方式，” 信学技報，PRMU2006-107，Oct. 2006.

[11] D.W.Embley, D.Lopresti, and G.Nagy, “Notes on Contemporary Table Recognition,” Proc. 7th. DAS, pp. 164-175, Feb. 2006

[12] 児島治彦，清末悌之，秋山照雄，“複雑な構造を持つ表の認識に関する基礎検討，” 情処学第 37 回全大，6W-8，pp. 1660-1661，Oct. 1988.

[13] 長谷博行，辻　正博，園田浩一郎，米田政明，酒井　充，“汎用を目指した自動文書画像認識システムー要素抽出技術の問題点と検討ー，” 信学技報，PRU94-33，Sep. 1994.

[14] 駱 琴, 渡辺豊英, 杉江　昇, “多種帳票文書の構造認識，” 信学論(D-II), vol. J76-D-II, no.10, Aug. 1993.

[15] J.Yuan, L.Xu, and C-Y.Suen, “Form Items Extraction By Model Matching,” Proc. 1st. ICDAR, pp. 210-218, Sep. 1991.

[16] H.Shinjo, E.Hadano, K.Marukawa, Y.Shima, and H.Sako, “A Recursive Analysis for Form Cell Recognition,” Proc. 6th. ICDAR, pp. 694-698, Sep. 2001.

[17] F.Cesarini, M.Gori, S.Marinai, and G.Soda, “INFORMys: A Flexible Invoice-Like Form-Reader System,” IEEE Trans. PAMI, vol.20, no.7, pp. 730-745, Jul. 1998.

[18] （社）電子情報技術産業協会，“OCR 関連装置／ソフトの出荷動向，” 入力装置に関する調査報告書，IS-09-情端-3，pp.45-46，Jun 2009.

[19] 田中　宏，中島健次，武部浩明，藤本克仁，“テクスチャ領域を含む帳票画像からの罫線抽出，” 信学技報，PRMU2006-264，Mar. 2007.

[20] 田中　宏，藤井勇作，武部浩明，藤本克仁，“二値化閾値の補正と罫線形状判定による罫線抽出の高精度化，” 信学技報，PRMU2007-216，Feb. 2008.

[21] 田中　宏，武部浩明，藤本克仁，“複数セル候補の組み合わせ探索に基づく帳票画像からのセル抽出，” 信学技報，PRMU2005-185，Feb. 2006.

[22] 村瀬　洋，若原　徹，梅田三千雄，“候補文字ラティス法による枠無し筆記文字列のオンライン認識，” 信学論(D), vol. J68-D, no.4, pp. 765-772, Apr. 1985.

[23] 仙田修司，濱中雅彦，山田敬嗣，“切り出し・認識・言語の確信度を統合した枠なしオンライン文字列認識手法，” 信学技報，PRMU98-138，Dec. 1998.

[24] 福島貴弘，中川正樹，“確率モデルに基づくオンライン枠なし手書き文字列認識，” 信学技報，PRMU98-139，Dec. 1998.

[25] 田中　宏，秋山勝彦，石垣一司，“階層遅延セグメンテーションを用いた実時間枠なしオンライン手書き文字列認識，” 信学技報，PRMU2001-264，Mar. 2002.

[26]小川厚徳, 武田一哉, 板倉文忠,“文長を考慮した言語モデルの検討,” 情処研資, 97-SLP-16-5, May 1997.

[27]鍋島一郎,“関数方程式による定式化と解法,” 動的計画法，pp.59-60，森北出版株式会社，東京，2005.

[28]長尾智晴,“分枝限定法,” 最適化アルゴリズム，pp.71-85，株式会社昭晃堂，東京，2000.

第五章　（文字抽出用二値化の研究）

[1] 藤本克仁, 鎌田　洋,“低解像度カラー文書画像から高品質な文字画像を抽出する二値化方式,” 信学技報 PRMU99-89, Oct. 1999.

[2] H.Kamada, and K.Fujimoto, “High-speed, High-accuracy Binarization Method for Recognizing Text in Images of Low Spatial Resolutions,” Proc. 5th ICDAR, pp.139-142, Bangalore, India, Sep. 1999.

[3] J.Sun, Y.Hotta, Y.Katsuyama, and S.Naoi, “Low Resolution Character Recognition by Dual Eigenspace and Synthetic Degraded Patterns,” Proc. 1st ACM Workshop on Hardcopy Document Processing, pp.15-22, Washington, DC, USA, Nov. 2004.

[4] Y.Hotta, J.Sun, Y.Katsuyama, and S.Naoi, “Robust Chinese Character Recognition by Selection of Binary-Based and Grayscale-Based Classifier,” Proc. 7th DAS, pp.553-563, Nelson, Newzealand, Feb. 2006.

[5] 田中　宏，藤井勇作，武部浩明，藤本克仁,“二値化閾値の補正と罫線形状判定による罫線抽出の高精度化,” 信学技報 PRMU2007-216, Feb. 2008.

[6] H.Tanaka, “Threshold Correction of Document Image Binarization for Ruled-line Extraction,” Proc. 10th ICDAR, pp.541-545, Barcelona, Spain, Jul. 2009.

[7] N. Otsu, “A Threshold Selection Method from Gray-level Histograms”, IEEE Trans. Systems, Man, and Cybernetics, vol.9, no.1, pp. 62-66, Jan. 1979.

[8] W. Niblack, “*An Introduction to Digital Image Processing*,” Prentice Hall, Englewood Cliffs, N.J., pp. 115-116, 1986.

[9] J.Ohya, A.Shio, and S.Akamatsu, “Recognizing Characters in Scene Images,” IEEE Trans. on PAMI, vol.16, No.2, pp.214-220, Feb. 1994.

[10]芦田和毅, 永井弘樹, 岡本正行, 宮尾秀俊, 山本博章,“情景画像からの文字抽出,” 信学論(D-II), vol.J88-D-II, no.9, pp.1817-1824, Sep. 2005.

[11]松田友輔，大町真一郎，阿曽弘具,“2 値化とエッジ抽出による情景画像からの高精度文字列検出,” 信学論(D), vol.J93-D, no.3, pp.336-344, Mar. 2010.

[12]H.Takebe, Y.Katsuyama, and S.Naoi, “Character String Extraction from Newspaper Headlines with a Background Design by Recognizing a Combination of Connected Components,” Proc. Document Recognition and Retrieval VI, pp.22-29, San Jose, CA, Jan. 1999.

[13]L. Eikvil, T. Taxt, and K. Moen, “A Fast Adaptive Method for Binarization of Document Images,” Proc. 1st ICDAR, pp. 435-443, Saint-Malo, France, Sep. 1991.

第六章　（帳票画像認識技術の実用化）

[1] O.Velek, S.Jaegar, M.Nakagawa, “Accuulated-Recognition-Rate Normalization for Combining Multiple On/Off-Line Japanese Character Classifiers Tested on a Large Database,” 4th. Intl. Workshop on Multiple Classifier Systems (MCS2003), pp.196-205, Guildford, UK, Jun 2003.

[2] H.Tabaru, Y.Nakano, “A Printed Japanese Character Recognition System Using a Majority Logic,” Proc. 3rd DAS, pp.185-189, Nagano, Japan, Nov. 1998.

[3] 佐伯胖,“きめ方の論理―社会的決定理論への招待,” 東京大学出版会, Apr. 1980, ISBN4-13-043017-3.

[4] C.Y.Suen, and Y.S.Huang, "A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals," IEEE PAMI vol.17, No.1, pp.90-94, Jan. 1995.

[5] T.K.Ho, J.J. Hull, and S.N.Srihari, "Decision Combination in Multiple Classifier Systems," IEEE PAMI

vol.16 No.1, pp.66-75, Jan. 1994.

[6] H.Tanaka, K.Nakajima, K.Ishigaki, K.Akiyama, and M.Nakagawa, "Hybrid Pen-Input Character Recognition System Based on Integration of Online-Offline Recognition," Proc. 5th ICDAR, pp.209-212, Sep. 1999.

[7] EPUB 形式の仕様は「国際電子出版フォーラム (International Digital Publishing Forum, IDPF)」が規格化し，普及促進している．2012 年現在，“http://idpf.org/epub” にて仕様や改定状況を知ることができる．最新仕様は ver.3 であり，EPUB3 と呼ばれる．

[8] DAISY 形式の仕様は「国際 DAISY コンソーシアム（DAISY Consorsium）」が規格化している．DAISY コンソーシアムの Web ページ “http://www.daisy.org/” にて仕様や改定状況を知ることができる．最新仕様は ver.4 であり，DAISY4 と呼ばれる．

[9] 織田英人，耒代誠仁，小沼元輝，中川正樹“筆記枠無しオンライン手書き文字列検索,” 信学技報 PRMU2003-239, Feb. 2004.

[10] 丸川勝美，藤澤浩道，嶋　好博，“認識機能の出力あいまい性を許容した情報検索手法の一検討,” 信学論(D-II), vol. J79-D-II, no.5, pp. 785-794, May. 1996.

[11] 太田　学，高須淳宏，安達　淳，“認識誤りを含む和文テキストにおける全文検索手法,”情処論, vol.39, no.3, pp.625-635, Mar. 1998.

[12] “イメージ処理パッケージ「AutoENTRY（オートエントリー）」新発売,”富士通 Press Release 1997-0249，Nov. 1997.（2012 年現在，“http://pr.fujitsu.com/jp/news/1997/Nov/11-4.html”にて参照可能）

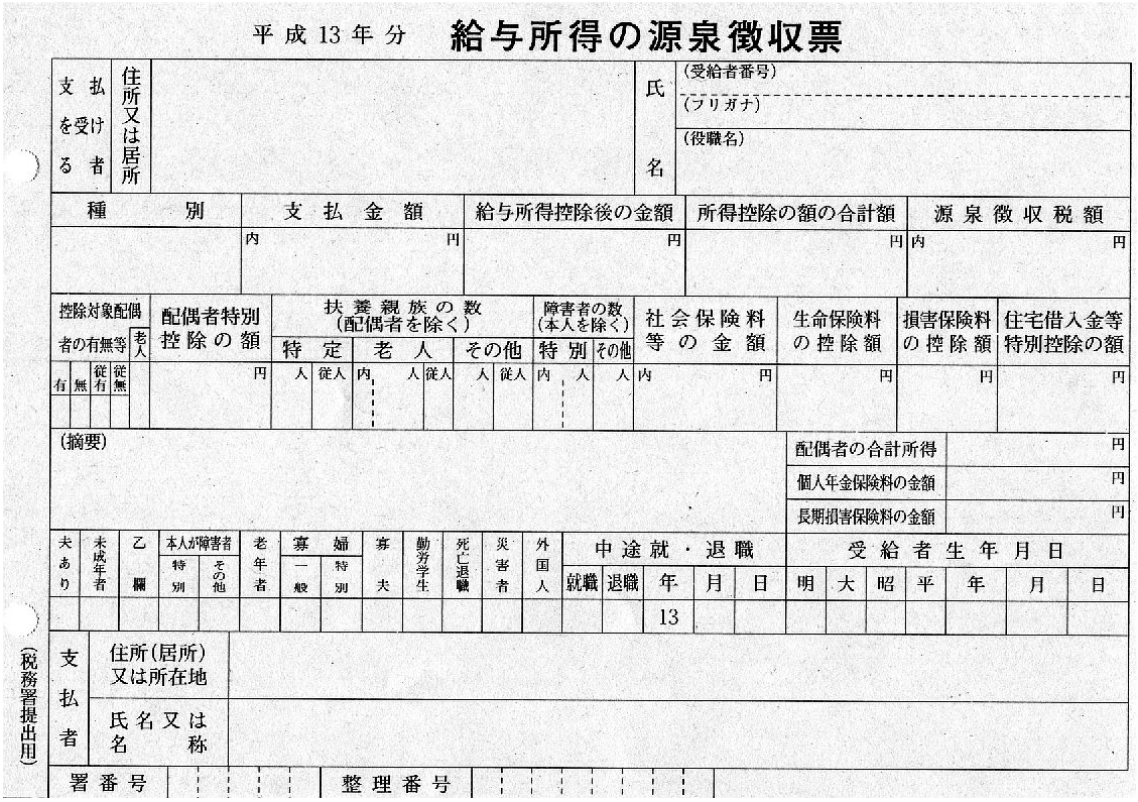
付録Ａ　帳票画像の例

近年，世の中では様々な書式の帳票が用いられるようになっており，特にユーザにとって読みやすいように，図や矢印が多用されるなど，帳票画像認識の立場で見るとより認識が困難な帳票画像が増加している．ここでは，本研究において評価用に用いた帳票画像の中の一部を提示する．

帳票は主に業務で用いられるため，その多くは外部持ち出しが禁止されている．本研究では，筆者が社内業務の一環として利用した帳票も用いたため，全ての帳票画像を公にすることはできない．また，世の中の帳票デザインは様々に変わり得るので，全ての帳票画像を提示することよりも，どのような種類の帳票が使われているかについて紹介することが重要だと考える．

そこで，ここではまず業務用の帳票（「未記入帳票」「見積書」「納品書」）について，実データを別の文字列に置き換えた帳票画像を提示する．これらは表や文字のレイアウトや画質については実際に帳票に近いものとなっているが，記載されている内容は変更してある．続いて，百貨店の店頭で配布されていた「入会申込書」が，近年のカラフルな帳票の例として適当だと考え，それも提示した．これらの帳票も評価用に用いている．

ここに示す帳票画像の例により，帳票画像には様々な書式が存在すること，近年では非常に多彩な帳票が存在することがご理解いただけると思う．なお，一部のサイズが大きな帳票画像は横向きに提示している．


平成 13 年分　給与所得の源泉徴収票

支払を受ける者　住所又は居所　氏名　(受給者番号)　(フリガナ)　(役職名)

種別	支払金額	給与所得控除後の金額	所得控除の額の合計額	源泉徴収税額
	内 円	円	円	内 円

控除対象配偶者の有無等	配偶者特別控除の額	扶養親族の数（配偶者を除く）			障害者の数（本人を除く）		社会保険料等の金額	生命保険料の控除額	損害保険料の控除額	住宅借入金等特別控除の額
有 無 従有 従無 老人	円	特定 人 従人	老人 内 人 従人	その他 人 従人	特別 内 人	その他 人	内 円	円	円	円

(摘要)　配偶者の合計所得 円　個人年金保険料の金額 円　長期損害保険料の金額 円

夫あり	未成年者	乙欄	本人が障害者 特別	本人が障害者 その他	老年者	寡婦 一般	寡婦 特別	寡夫	勤労学生	死亡退職	災害者	外国人	中途就・退職 就職	中途就・退職 退職	中途就・退職 年 月 日	受給者生年月日 明 大 昭 平 年 月 日
															13	

(税務署提出用)　支払者 住所(居所)又は所在地　氏名又は名称　署番号　整理番号

図 A-1 未記入帳票の例(1)

図 A-2 未記入帳票の例(2)：横向き

○月別売上(収入)金額及び仕入金額

月	売上(収入)金額 決算額	売上(収入)金額	仕入金額 決算額	仕入金額
	円	円	円	円
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
家事消費等				
雑収入				
計				

○専従者給与の内訳

氏名	続柄	年齢	従事月数	支給額 給料	支給額 賞与	支給額 合計	源泉徴収税額
		歳	月	円	円	円	円
計							

○貸倒引当金繰入額の計算 (この計算に当たっては、「決算の手引き」の「貸倒引当金」の項を読んでください。)

			金額 決算額	金額
個別評価による本年分繰入額 (「個別評価による貸倒引当金に関する明細書」の⑬欄の金額を書いてください。)		①	円	円
一括評価による本年分繰入額	年末における一括評価による貸倒引当金の繰入れの対象となる貸金の合計額	②		
	本年分繰入限度額 (②×5.5%(金融業は3.3%))	③		
	本年分繰入額	④		
本年分の貸倒引当金繰入額 (①+④)		⑤		

○青色申告特別控除額の計算 (この計算に当たっては、「決算の手引き」の「青色申告特別控除」の項を読んでください。)

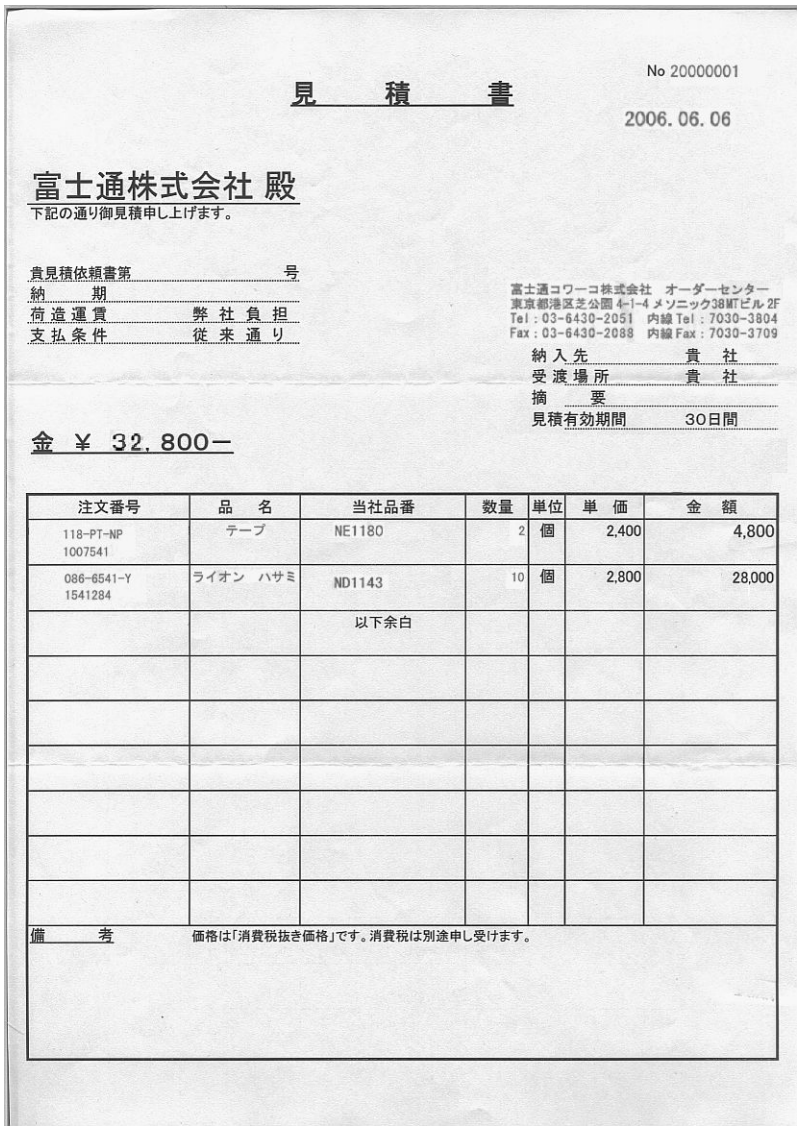
本年分の不動産所得の金額 (青色申告特別控除額を差し引く前の金額)		⑥	(赤字のときは0) 円	(赤字のときは0) 円
青色申告特別控除前の所得金額 (1ページの「損益計算書」の㊱欄の金額を書いてください。)		⑦	(赤字のときは0)	(赤字のときは0)
55万円(45万円)の青色申告特別控除を受ける場合	55万円(45万円)と⑥のいずれか少ない方の金額 (不動産所得から差し引かれる青色申告特別控除額です。)	⑧		
	青色申告特別控除額 (「55万円(45万円)−⑧」と⑦のいずれか少ない方の金額)	⑨		
上記以外の場合	10万円と⑥のいずれか少ない方の金額 (不動産所得から差し引かれる青色申告特別控除額です。)	⑧		
	青色申告特別控除額 (「10万円−⑧」と⑦のいずれか少ない方の金額)	⑨		

○給料賃金の内訳

氏名	年齢	従事月数	支給額 給料賃金	支給額 賞与	支給額 合計	源泉徴収税額
	歳	月	円	円	円	円
その他(人分)						
計						

(注) 貸倒引当金、専従者給与や3ページの割増(特別)償却以外の特典を利用する人は、適宜の用紙にその明細を記載し、この決算書に添付してください。

図 A-3 未記入帳票の例(3):横向き



No 20000001

見　積　書

2006. 06. 06

富士通株式会社 殿

下記の通り御見積申し上げます。

貴見積依頼書第　　号
納　　期
荷造運賃　　弊社負担
支払条件　　従来通り

富士通コワーコ株式会社　オーダーセンター
東京都港区芝公園 4-1-4 メソニック38MTビル 2F
Tel : 03-6430-2051　内線 Tel : 7030-3804
Fax : 03-6430-2088　内線 Fax : 7030-3709

納入先　　貴　社
受渡場所　　貴　社
摘　　要
見積有効期間　　30日間

金 ￥ 32,800－

注文番号	品　名	当社品番	数量	単位	単　価	金　額
118-PT-NP 1007541	テープ	NE1180	2	個	2,400	4,800
086-6541-Y 1541284	ライオン　ハサミ	ND1143	10	個	2,800	28,000
		以下余白				

備　　考　　価格は「消費税抜き価格」です。消費税は別途申し受けます。

図 A-4 見積書の例(1)

富士通株式会社　御中

NO. IC － 114 － 0450

2006.06.06

2006.06.06　をもって御照会の物件
下記の通り御見積申し上げます。

千　円

見積有効期限　今回限り

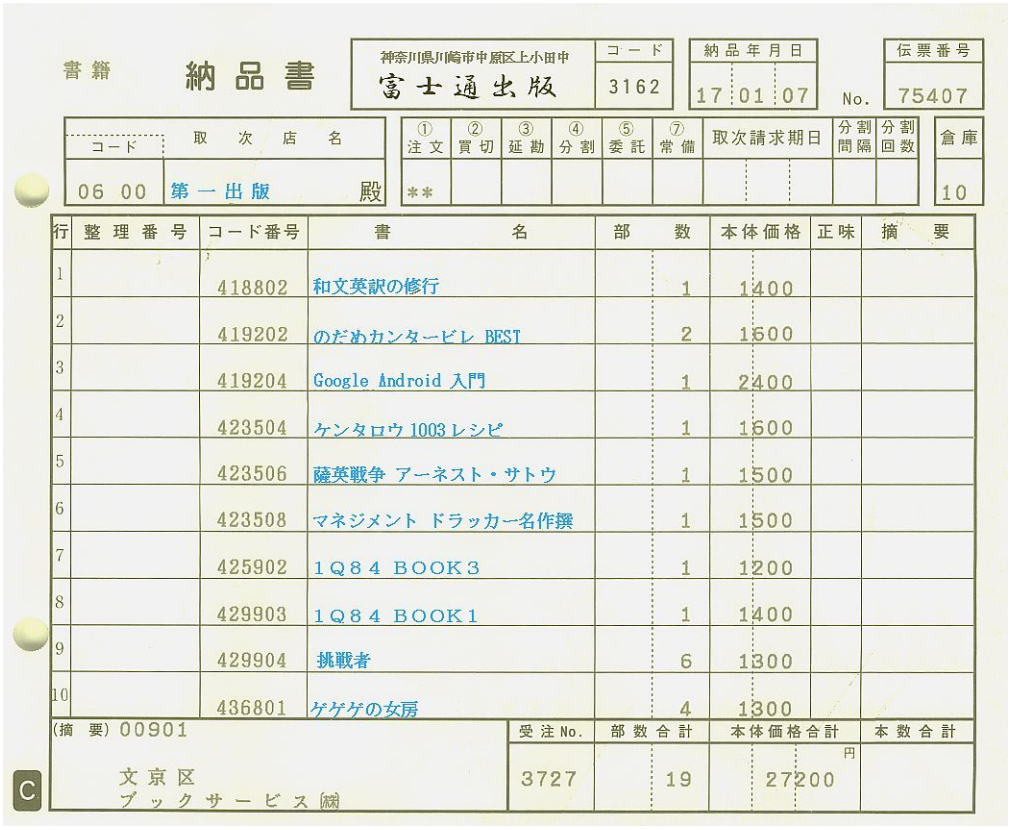
〒222-0033
神奈川県横浜市港北区新横浜２－１５－１６
ＮＯＦ新横浜ビル
富士通株式会社　新横浜ソフトウェアセ

件名		受渡場所		納入期日	貴社ご指定期日
持込場所	貴社ご指定場所	支払条件	従来通り	運送方法	トラック便

貴番号	品名・品番	仕様・材料	個数	数量	単位	単価	金額
27952084825	086-810-1-12	３Ｍ　ポストイット	5本	50g	g	600円／g	30,000円
	以下、余白						

備考：　深沢克夫様ご依頼分です。
納期：２／15着
この見積書には、消費税を含んでおりません。

図 A-5 見積書の例(2)



書籍　納品書

神奈川県川崎市中原区上小田中
富士通出版　コード 3162

納品年月日 17 01 07　No.　伝票番号 75407

コード	取次店名	①注文	②買切	③延勤	④分割	⑤委託	⑦常備	取次請求期日	分割間隔	分割回数	倉庫
06 00	第一出版　殿	**									10

行	整理番号	コード番号	書名	部数	本体価格	正味	摘要
1		418802	和文英訳の修行	1	1400		
2		419202	のだめカンタービレ BEST	2	1600		
3		419204	Google Android 入門	1	2400		
4		423504	ケンタロウ 1003 レシピ	1	1600		
5		423506	薩英戦争 アーネスト・サトウ	1	1500		
6		423508	マネジメント ドラッカー名作撰	1	1500		
7		425902	1Q84 BOOK3	1	1200		
8		429903	1Q84 BOOK1	1	1400		
9		429904	挑戦者	6	1300		
10		436801	ゲゲゲの女房	4	1300		

(摘要) 00901

文京区
ブックサービス㈱

受注No.	部数合計	本体価格合計	本数合計
3727	19	27200 円	

図 A-6 納品書の例

図 A-7 入会申込書の例(1)

2000WI-020331

●この申込書はクレジットカードをお持ちのかたのみご利用いただけます。

FI　@nifty入会申込書　WIN6

私は会員規約および料金体系に同意のうえ、入会を申し込みます。

フリガナ
氏名　印

性別　0.男性　1.女性

生年月日（西暦）　19　年　月　日

ローマ字名（活字体）　（姓）　（名）　※ご記入のない場合は当社で記入します。

フリガナ
住所　〒　都道府県　TEL　－　－

フリガナ
緊急連絡先（携帯・勤務先等）　TEL　－　－

メールアドレス記入欄

あなたの好きなメールアドレスが作れます

ご記入例　taro.123@nifty.comのメールアドレスをご希望の場合は、下記のようにご記入ください。
ティーエーアールオーピリオド
t a r o ピ ① ② ③　@nifty.com

●第1希望～第3希望は全て異なる文字列にしてください。　●メールアドレス登録後の変更はできません。

フリガナ　第1希望
フリガナ　第2希望
フリガナ　第3希望

※使用可能な文字種は、英小文字a～z、数字0～9、記号「-」ハイフン、「_」アンダーバー、「.」ピリオドです。（ピリオドは末尾や連続での使用はできません）カタカナ、ひらがなは登録できません。
※「-」ハイフンの場合は「ハ」、「_」アンダーバーの場合は「ア」、「.」ピリオドの場合は「ピ」を記入してください。また、数字は全て○で囲んでください。例）①②③
※先頭、末尾、文字間にスペースはあけられません。（スペースがあった場合はスペース部分は詰めさせていただきますのでご了承ください）
※英字の筆記体でのご記入はご遠慮ください。
※英字は大文字での登録はできません。（大文字でご記入いただいた場合は小文字での登録となります）
※2文字以上32文字以内でご記入ください。
※先頭の1文字は必ず英文字にしてください。
※他の会員がすでに登録しているメールアドレスと同一の場合は、入会後、再度申請をいただく必要があります。
※コネクションID等との混乱を防ぐため、英文字4字＋数字4桁と英文字3字＋数字5桁の文字列の登録はできません。

料金コース（該当するものに○）

スタンダード料金コース	・お手軽1コース	月々250円（1時間まで）
	・お手軽5コース	月々950円（5時間まで）
	・無制限コース	月々2,000円（時間無制限）
	・デイタイムコース	月々1,200円（6:00～21:00は時間無制限）
	・オープンコース	月々1,200円（時間無制限）
テレコミ料金コース	・テレコミ3コース	月々1,000円（3時間まで）
	・テレコミ10コース	月々2,500円（10時間まで）
ザ・学割 ※学生に限ります	・無制限コース（ザ・学割）	月々1,500円（時間無制限）
	・オープンコース（ザ・学割）	月々600円（時間無制限）

★「オープンコース」はインターネット経由接続（他プロバイダー、LAN環境等）に限り、月々1,200円（ザ・学割は600円）でご利用になれます。
★「テレコミ料金コース」を除き、加入料1,000円が必要です。
★料金コースの選択がない場合は「無制限コース」での登録となります。
★「テレコミ料金コース」は、3ヶ月接続料金無料の対象外です。

学生証（コピー）貼付欄（「ザ・学割」をお申し込みのかたのみ下欄をご記入のうえ、学生証のコピーを必ず貼付してください。）

クレジットカード種別（該当するものに○）　JCB、VISA、UC、NICOS、セゾン、DC、ミリオン、オリコ、JACCS、アメリカン・エキスプレス、ダイナース、CF（JCB／VISA／MasterCard提携のみ）、バンクカード、OMC、イオン、アプラス、ライフ、国内信販、MasterCard

※クレジットカードの名義は入会されるご本人様に限定させていただきます。クレジットカードをご利用にならないかたには「アット・ニフティカードレス会員」制度もあります。P7をご覧ください。

カード番号

カード有効期限　月／　年（西暦）　（例）06月／01年

※カード番号、有効期限は必ずご確認のうえご記入ください。

この冊子が添付されていたお買い上げの製品　メーカー名　富士通　製品　パソコン本体

※記載内容に不備、記入漏れがあった場合、印のない申込書は、ご確認のため一旦ご返送する場合があります。

フリガナ
学校名　学校　学部　学年　卒業予定年月　西暦　年　月（ザ・学割は卒業予定月まで有効）

学校所在地　〒　学校TEL　－　－

※学校の情報は添付する学生証のコピーと同内容をご記入ください。また、お申し込みの際には必ず「学生証（コピー）」を申込書に貼付してください。「学生証（コピー）」は、入会審査以外の目的には使用いたしません。
※弊社では、お客様の個人情報を、@nifty会員規約および@nifty個人情報保護ポリシーに基づき適切に取り扱うとともに、これらの定める範囲内で、サービスの提供や弊社サービスの案内等のために利用させていただいております。@nifty個人情報保護ポリシーは、右記のURLよりご参照いただけます。（http://www.nifty.com/policy/privacy.htm）

図 A-8 入会申込書の例(2)

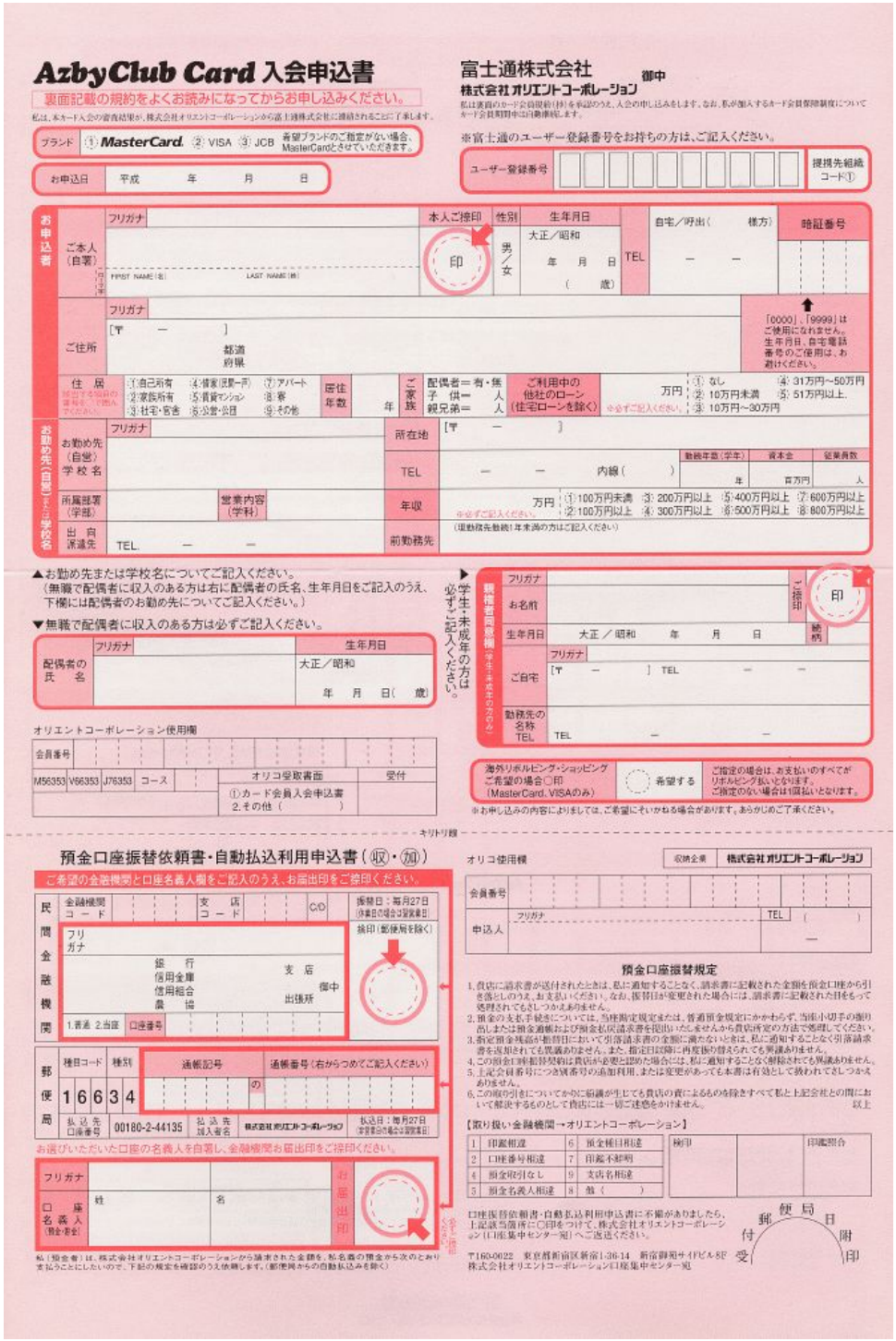


図 A-9 入会申込書の例(3)

付録B　イメージスキャナ市場状況

B.1. 背景

文書画像認識の対象となる画像は，多くの場合，イメージスキャナによって取得される．近年ではデジタルカメラで撮影された画像も利用される頻度が増え，例えばスマートフォン用のアプリケーションの中に「スキャナアプリ」という分類が使われるなど，デジタルカメラで文書画像を取得することも当たり前になっている．しかし，デジタルカメラで取得した文書画像は，スキャナ画像に比べて焦点ボケが生じやすく，手振れの発生や照明の不安定さなど，画質劣化に繋がる要因が多いため，認識精度が要求される用途では未だにイメージスキャナによる文書画像が主流である．

ここでは，文書画像認識技術が活用される主な対象として，イメージスキャナの市場動向について記す．イメージスキャナの市場は，JEITA（(社)電子情報技術産業協会）が毎年継続的に調査を行っており，JEITA のイメージスキャナ専門委員会が，主要メーカー各社から収集したデータに基づいて一年間の装置出荷台数と出荷金額を集計している．JEITA の報告書を年次集計して比較すれば，スキャナ市場の大きな流れを把握することができる．また，大手量販店の販売データを集計した「BCN ランキング」を参照すれば，主にコンシューマ市場の動向を知ることができる．これら二つの市場統計により，スキャナ市場の大きな流れを概観する．

B.2. JEITA市場調査

JEITA の市場調査では，イメージスキャナは「コンシューマ向けスキャナ」「業務用スキャナ」「その他」の三つの分類で集計されている．例えば 2009 年の出荷実績を**表 B-1** に示す．

表B-1　2009 年イメージスキャナ出荷実績

2009 年イメージスキャナ市場		台数（前年比）	金額（前年比）
国内出荷と輸出を合わせた総出荷		276 万台（19%減）	576 億円（24%減）
	コンシューマ向けスキャナ	198 万台（23%減）	124 億円（30%減）
	業務用スキャナ	77 万台（8%減）	425 億円（25%減）
	その他	1.8 万台（119%増）	27 億円（130%増）

それぞれの分類の基準は下記の通りである．

- コンシューマ向けスキャナ　　A３以下／50,000 円以下のフラットベッド
- 業務用スキャナ　　A３以下／50,000 円以下のフラットベッドを除く
- その他　　それ以外

この分類では分かりづらいが，コンシューマ向け，業務向けともＡ３サイズ以下の用紙をスキャン対象としており，A３を超えるサイズの用紙がスキャンできる装置は一般的ではないものとして，その他に分類する．実際には，コンシューマ向けスキャナは「フラットベッド型」であり，業務用スキャナの大半は自動用紙送り機能（ADF）を搭載した「ADF 型」に分類される．

(a) フラットベッド型スキャナの例
（CANON CanoScan LIDE 210）

(b) ADF 型スキャナの例
（PFU ScanSnap S1500）

図 B-1　イメージスキャナ装置の例

従来，ADF 搭載イメージスキャナは大量の用紙を高速に読み取るための装置として，業務用に分類されていた．一方，フラットベッド型は構造的に単純で安価に製造できるため，個人用途向けの製品とみなされていた．しかしながら，PFU が 2001 年から発売している ScanSnap シリーズが，それまで上位機種でのみ用いられていた ADF 機能を安価な個人向け製品に搭載したことから，個人用途でも ADF 付きイメージスキャナが一般化している．そのため「コンシューマ用」「業務用」という分類名は現状とは合っておらず，注意が必要である．

JEITA 調査では，ADF 型スキャナが実際に個人で使われているか，業務で使われているかについての調査は行っていない．また，同じ機種でもどのような用途に用いるかを把握するためには各ユーザに追跡調査を行わなければならず，現実的ではない．更に，出荷台数・金額を調べるメーカー側の調査としては，使用用途まで調査するのは調査目的から外れる．したがって，JEITA 調査で見えるのは，用途に関わらずフラットベッド型スキャナと ADF 型スキャナの市場の変遷のみだと言える．

「その他」に分類される装置は，Ａ３を超えるスキャナの他は，主にフィルムスキャナとスタンド型スキャナである．フィルムスキャナはマイクロフィルムなどのフィルムをスキャンして画像を取得するスキャナであり，スタンドスキャナは卓上台座上の被写体をスキャンする非接触スキャナで，主に銀行窓口で紙帳票や免許証などの画像を撮影する用途で用いられる．

JEITA 市場調査の報告より，2002 年から 2010 年の出荷台数（国内・海外含む）を集計した結果を**表 B-2** に示す．イメージスキャナの総出荷数は 2002 年から減少傾向にあったが，その主要因はフラットベッドスキャナの出荷減少にある．2002 年当時のスキャナの用途についてのデータは無いが，筆者の記憶によれば従来のスキャナの主な用途は写真のデジタル化であったように思われる．専用用紙に印刷された写真をスキャナで読み込んで，PC 上で保存や表示を行うという用途がフラットベッドスキャナの主な用途とすれば，デジタルカメラの普及によって急激に需要が低下するという現象は十分にあり得ることである．

表 B-2　スキャナ出荷台数の変遷（単位：万台）

		出荷台数(万台)			出荷台数(比率)		
	総出荷	FLAT	ADF	その他	FLAT	ADF	その他
2002	662	643	5.7	10.5	97.1%	0.9%	1.6%
2003	511	432	15.2	12.4	84.5%	3.0%	2.4%
2004	437	404	22.3	10.3	92.4%	5.1%	2.4%
2005	387	344	34	7.7	88.9%	8.8%	2.0%
2006	339	289	49	0.8	85.3%	14.5%	0.2%
2007	336	270	65	0.8	80.4%	19.3%	0.2%
2008	342	258	83	0.8	75.4%	24.3%	0.2%
2009	276	198	77	1.8	71.7%	27.9%	0.7%
2010	367	232	134	0.9	63.2%	36.5%	0.2%

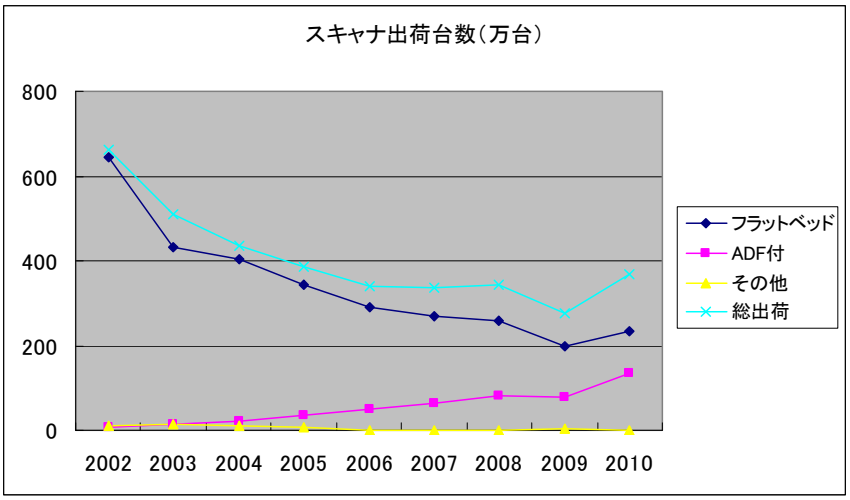


図 B-2　スキャナ出荷台数

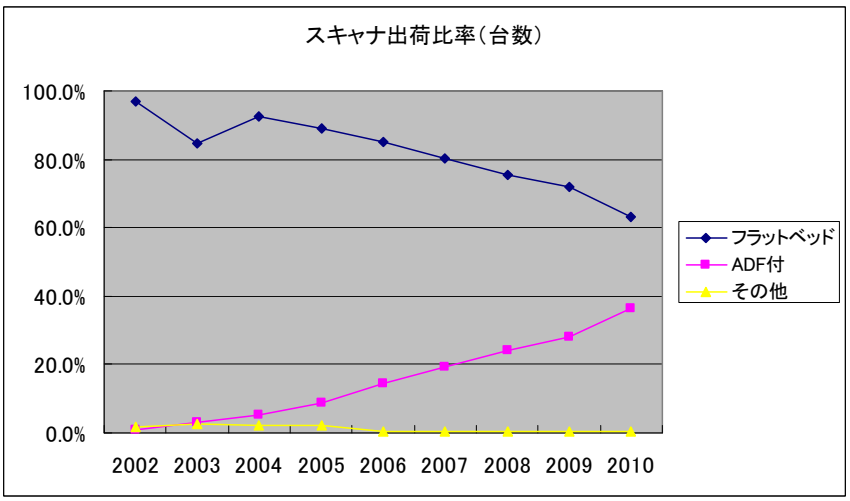


図 B-3　スキャナ出荷台数の比率

表 B-3　スキャナ出荷金額の変遷（単位：億円）

		出荷金額(億円)			出荷金額(比率)		
	総出荷	FLAT	ADF	その他	FLAT	ADF	その他
2002	942	760	96.7	67.9	80.7%	10.3%	7.2%
2003	755	533	138	54.3	70.6%	18.3%	7.2%
2004	720	469	177	43	65.1%	24.6%	6.0%
2005	700	401	234	34	57.3%	33.4%	4.9%
2006	644	246	385	13	38.2%	59.8%	2.0%
2007	773	223	537	13	28.8%	69.5%	1.7%
2008	756	178	567	12	23.5%	75.0%	1.6%
2009	576	124	425	27	21.5%	73.8%	4.7%
2010	818	145	610	45	17.7%	74.6%	5.5%

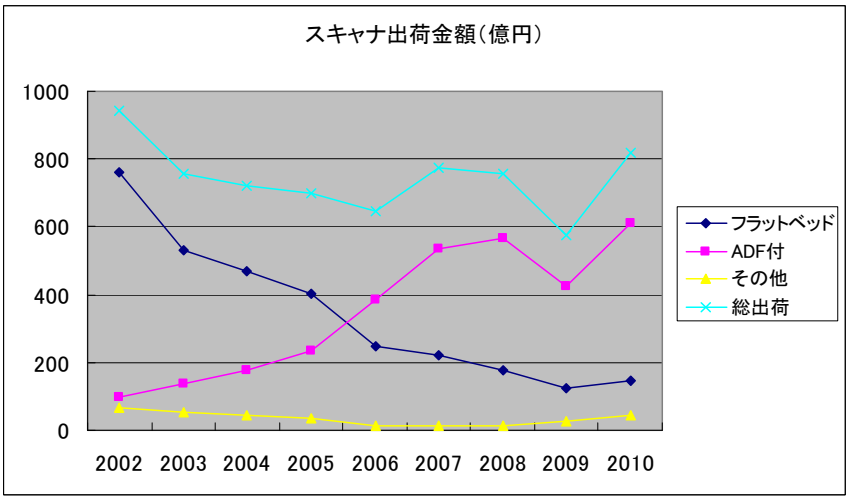


図 B-4　スキャナ出荷金額

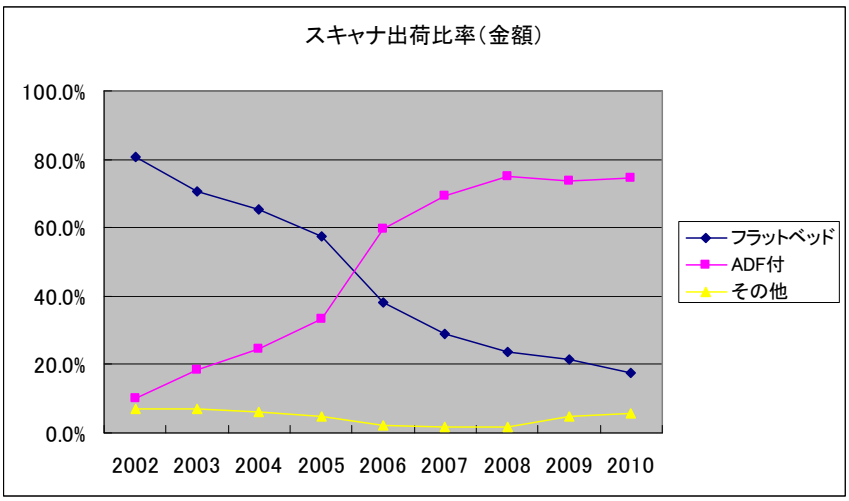


図 B-5　スキャナ出荷金額の比率

それに対して，ADF 型スキャナの出荷台数が 2005 年前後より増加している．**図 B-2** ではさほど増えていないように見えるが，ADF 型スキャナはフラットベッド型に比べて高価であり，**図 B-4** で分かるように，ADF 型スキャナは 2006 年にはフラットベッド型の出荷金額を逆転し，イメージスキャナ市場の体勢を左右する立場となっている．

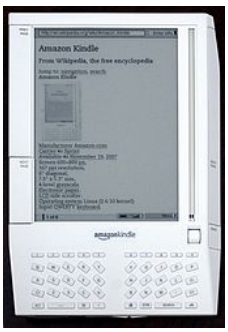
出荷台数と出荷金額のそれぞれについて，イメージスキャナ全体に対する割合も示した（**図 B-3**, **図 B-5**）．出荷台数では単にフラットベッド型が減少して（それでもトータルでは多い），ADF 型が少しずつ迫っているだけのように見えるが，出荷金額を**図 B-5** で見ると，明らかに 2005 年から 2006 年でイメージスキャナ市場が変質している．

次節で詳しく述べるが，近年の ADF 型イメージスキャナ市場を牽引したのは PFU のドキュメントスキャナ ScanSnap シリーズである（製造元は PFU だが，販売時には富士通ブランドで出荷されている．ここでは開発元に敬意を表し，一貫して PFU 製品と記述する）．ScanSnap は，業務用ドキュメントスキャナである fi シリーズの個人向け製品として 2001 年から発売されており，デザインを一新した ScanSnap S500 が発売されて以降，急激に市場シェアを拡大している．その ScanSnap S500 の発売が 2006 年 2 月である．更に，2007 年には Amazon が電子書籍端末 Kindle を発売，2010 年には Apple が iPad を発売と，電子書籍ブームと呼ばれる現象が起きている．

書籍を電子書籍端末で読むには，十分な数の電子書籍が流通していなければならないが，日本国内では電子書籍の流通が進まず（書籍の権利関係が複雑なために，Amazon などの先行業者が参入しずらかったと言われている），それなら紙の本をスキャナで電子化して，自分で電子書籍を作れば良いという先進ユーザが現れ始めた．このように，紙書籍からスキャナで電子書籍を作る作業は，「自炊」という用語が作られるほどに普及し，そのために最適なスキャナとして ScanSnap が市場に選ばれたのだと言える．

世の中で電子書籍端末が話題になったのは Kindle がきっかけではあるが，その直前の 2006 年 4 月には，iRex 社が iLiad という端末を発売して，一部で大きな話題になっていた．また「自炊」という言葉と共に世の中に広く知られた感のある ScanSnap であるが，2001 年に初期版を発売して以来，地道にユーザを増やしている．**図 B-5** は ScanSnap だけのグラフではないが，ADF 型のいわゆるドキュメントスキャナが 2006 年に需要を増やした（グラフの角度が急峻）のは，これらの先行的な実績がこの年に花開き，それ以降のブームに直結したものと思われる．

(a) iRex iLiad (2006 年 4 月)



(b) Amazon Kindle (2007 年 11 月)

(c) Apple iPad (2010 年 4 月)

図 B-6　話題になった電子書籍端末の例

本節での考察をまとめると，JEITA 市場調査の結果から見えるのは，主に写真のスキャンが主目的であったフラットベッド型スキャナ中心のイメージスキャナ市場が，次第に文書画像をスキャンするための ADF 型スキャナに移行していった様子である．その要因分析は定性的なものとなるが，電子書籍ブームや，いわゆる「自炊」ブームなどの流れの中で，紙文書をスキャンして文書イメージを読む（電子書籍，電子書類）というスタイルが世の中に定着していったことが市場変化の大きな要因であるように思われる．

このような市場変化の中では，文書画像をより便利に活用するための文書画像認識技術の重要性はより高まるものと思われる．本論文の第 6 章でも述べたように，本研究の成果である帳票文書画像認識の高精度化のための技術は，今後はむしろ一般文書画像認識への適用がより重要視されてくるように思われる．

B.3. BCNランキングによるコンシューマ市場調査

BCN ランキングとは，全国の量販店の POS データを日次で収集し，製品ジャンルごとに集計した実売データである（BCN ランキング Web サイト “http://bcnranking.jp/” より）．このデータを毎年集計して，ジャンルごとの上位 3 位までのメーカー別順位を公開し，1 位のメーカーを表彰するものが，BCN AWARD である．民間企業による調査結果であるが，集計過程が公開されているなど，比較的信用できるデータである．もちろん，量販店のみの集計データなので，各メーカーの順位をそのまま表しているものではない．あくまで量販店での順位であることを忘れてはならない．

しかし，機械的に集計していることから，市場動向の一面を確実に捉えているのは事実である．ここでは，前節でも述べたように，ADF 型スキャナの成長を支えた PFU スキャナの動向を中心として，ドキュメントスキャナの普及状況について概観する．

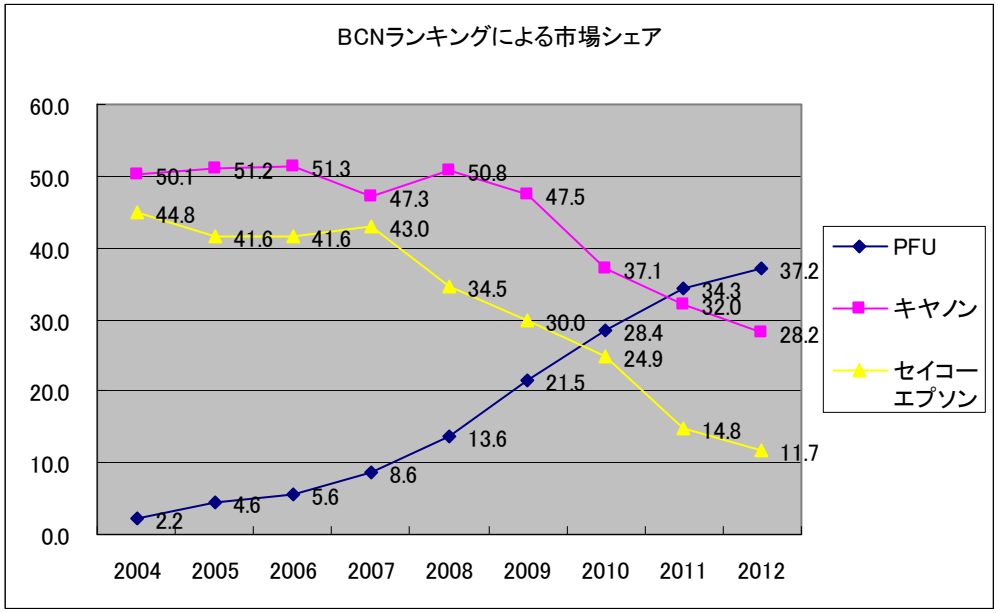


図 B-7　BCN ランキングによるスキャナ部門の市場シェア（単位：%）

2004 年以降の BCN アワードで公開された市場シェアを**図 B-7** に示す．2004 年から現在に至るまで，上位 3 位までのメーカーは変わっていない（キヤノン，PFU，セイコーエプソン）．このグラフで分かるのは，PFU が急激にシェアを拡大し，従来の 2 強であったキヤノンとセイコーエプソンがシェアを減らしている様子である．

BCN ランキングの値は販売台数シェアなので，低価格製品が多いメーカーは有利である．またスキャナ部門はフラットベッド型，ADF 型など全てを含むため，低価格で台数の割合が大きなフラットベッドスキャナは，販売金額の割にシェアを伸ばしやすいという特徴がある．

そのような中でも，PFU がシェア 1 位となっているのは，量販店ではそれだけ ADF 型スキャナに市場がシフトしているためだと言える．というのは，PFU のイメージスキャナ製品の中にはコンシューマ向けフラットベッド型スキャナは存在しないので，PFU のシェアはほぼ全てが ADF 型スキャナだと言えるためである．一方で，セイコーエプソンの主力はフラットベッド型である．特定用途には有効だと思われるが，コンシューマ市場のニーズは確実に低下している．キヤノンには ADF 型スキャナも存在し，市場での評価も高いので，シェア低下の程度は中程度である．

図 B-7 を見て先ず言えるのは，前節でも述べたように ADF 型スキャナの市場を地道に開拓してきた PFU の実績が，電子書籍などのブームとの相乗効果によって圧倒的に市場シェアを奪ったという事実であるが，それよりも大きな市場の流れとして，コンシューマが文書画像をスキャンする頻度が急激に上がっているという点が重要だと思われる．これだけ市場が激変すれば，今後は PFU に対抗するメーカーも増え，これまでのような圧倒的なシェアは維持するのが難しくなるであろう．しかし，ユーザが従来よりも文書画像のスキャンを志向するようになってきているという傾向は，そう簡単には変わらないように思われる．

今後は，ADF 型スキャナからデジタルカメラによる文書画像，スタンド型スキャナなど，様々な装置によって取得された文書画像が利用されるようになると思われる．そのような場合でも，文書画像をより便利に活用するための，文書画像認識技術の重要性はより大きなものになるものと考えられる．