(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3950498号
(P3950498)
(45)発行日　平成19年8月1日(2007.8.1)　(24)登録日　平成19年4月27日(2007.4.27)

(51)Int.Cl.			F I		
G06T	7/60	(2006.01)	G06T	7/60	200C
G06K	9/20	(2006.01)	G06K	9/20	340J

請求項の数 11　(全 25 頁)

(21)出願番号	特願平8-221835
(22)出願日	平成8年8月6日(1996.8.6)
(65)公開番号	特開平9-171557
(43)公開日	平成9年6月30日(1997.6.30)
審査請求日	平成15年8月1日(2003.8.1)
(31)優先権主張番号	08/514,250
(32)優先日	平成7年8月11日(1995.8.11)
(33)優先権主張国	米国(US)

(73)特許権者　000001007
キヤノン株式会社
東京都大田区下丸子３丁目３０番２号
(74)代理人　100125254
弁理士　別役　重尚
(74)代理人　100118278
弁理士　村松　聡
(74)代理人　100138922
弁理士　後藤　夏紀
(74)代理人　100136858
弁理士　池田　浩
(74)代理人　100135633
弁理士　二宮　浩康

最終頁に続く

(54)【発明の名称】イメージ処理方法及び装置

(57)【特許請求の範囲】

【請求項１】

入力されたイメージ内に含まれる連結成分に基づいて特定される領域の外接矩形を得る外接矩形取得工程と、

前記入力されたイメージ内に含まれる連結成分に基づいて特定される領域の輪郭情報を得る輪郭情報取得工程と、

前記外接矩形取得工程で得た複数の外接矩形のうち、重複する外接矩形を判別する重複判別工程と、

前記重複判別工程で重複すると判別された外接矩形が、当該外接矩形に対応する所望の連結成分のイメージデータ抽出に影響するかどうか判別する影響判別工程と、

前記重複判別工程で重複すると判別された外接矩形に関して、前記輪郭情報に基づいて複数の矩形に分解する分解工程と、

前記分解工程で複数の矩形に分解された場合には、当該分解された複数の矩形に基づいてイメージデータを抽出し、前記分解工程で複数の矩形に分解されなかった場合には、当該分解されなかった外接矩形に基づいてイメージデータを抽出する抽出工程と、

前記抽出工程で抽出されたイメージデータを処理する処理工程とを有し、

前記分解工程では、前記影響判別工程で影響すると判別された場合、前記重複判別工程で重複すると判別された外接矩形を前記輪郭情報に基づいて複数の矩形に分解し、前記影響判別工程で影響しないと判別された場合、前記重複判別工程で重複すると判別された外接矩形の分解を行わないことを特徴とするイメージ処理方法。

【請求項２】
前記外接矩形取得工程では、各外接矩形の左上角と右下角の座標を求める工程を含み、前記重複判別工程では各外接矩形の座標同士を比較することにより、重複する外接矩形を判別することを特徴とする請求項１に記載のイメージ処理方法。
【請求項３】
前記輪郭情報取得工程では、前記入力されたイメージ内に含まれる連結成分に基づいて特定される領域のタイプに応じて、前記輪郭情報を得ることを特徴とする請求項１に記載のイメージ処理方法。
【請求項４】
前記領域のタイプがテキストタイプである場合、テキスト連結成分の頂部、底部、左側及び右側の縁に基づいて、前記輪郭情報を得ることを特徴とする請求項３に記載のイメージ処理方法。
【請求項５】
前記領域のタイプが内部成分を持たない非テキストタイプである場合、連結成分の輪郭を連結することにより、前記輪郭情報を得ることを特徴とする請求項３に記載のイメージ処理方法。
【請求項６】
前記領域のタイプが内部成分を有する非テキストタイプである場合、外側の輪郭と、内部の白輪郭とに基づいて、前記輪郭情報を得ることを特徴とする請求項３に記載のイメージ処理方法。
【請求項７】
前記影響判別工程で影響すると判別された場合の前記重複すると判別された外接矩形が、テキストタイプ領域の外接矩形と非テキストタイプ領域の外接矩形とである場合に、前記分解工程では、前記非テキストタイプ領域の輪郭情報と前記テキストタイプ領域の連結成分とに基づいて、複数の矩形に分解することを特徴とする請求項１に記載のイメージ処理方法。
【請求項８】
前記影響判別工程で影響すると判別された場合の前記重複すると判別された外接矩形が、テキストタイプ領域の外接矩形と非テキストタイプ領域の外接矩形とである場合に、前記分解工程では、前記非テキストタイプ領域の輪郭情報と前記テキストタイプ領域の輪郭情報とに基づいて、複数の矩形に分解することを特徴とする請求項１に記載のイメージ処理方法。
【請求項９】
前記影響判別工程で影響すると判別された場合の前記重複すると判別された外接矩形が内部成分を持たない非テキストタイプ領域の外接矩形同士である場合、前記分解工程では、前記非テキストタイプ領域の輪郭情報に基づいて、複数の矩形に分解することを特徴とする請求項１に記載のイメージ処理方法。
【請求項１０】
前記影響判別工程で影響すると判別された場合の前記重複すると判別された外接矩形が内部成分を有する非テキストタイプ領域である場合、前記分解工程では、前記内部成分の輪郭情報と、非テキストタイプ領域内部の白輪郭の輪郭情報とに基づいて、複数の矩形に分解することを特徴とする請求項１に記載のイメージ処理方法。
【請求項１１】
入力されたイメージ内に含まれる連結成分に基づいて特定される領域の外接矩形を得る外接矩形取得手段と、
前記入力されたイメージ内に含まれる連結成分に基づいて特定される領域の輪郭情報を得る輪郭情報取得手段と、
前記外接矩形取得手段で得た複数の外接矩形のうち、重複する外接矩形を判別する重複判別手段と、
前記重複判別手段で重複すると判別された外接矩形が、当該外接矩形に対応する所望の

連結成分のイメージデータ抽出に影響するかどうか判別する影響判別手段と、
前記重複判別手段で重複すると判別された外接矩形に関して、前記輪郭情報に基づいて複数の矩形に分解する分解手段と、
前記分解手段で複数の矩形に分解された場合には、当該分解された複数の矩形に基づいてイメージデータを抽出し、前記分解手段で複数の矩形に分解されなかった場合には、当該分解されなかった外接矩形に基づいてイメージデータを抽出する抽出手段と、
前記抽出手段で抽出されたイメージデータを処理する処理手段とを有し、
前記分解手段は、前記影響判別手段で影響すると判別された場合、前記重複判別手段で重複すると判別された外接矩形を前記輪郭情報に基づいて複数の矩形に分解し、前記影響判別手段で影響しないと判別された場合、前記重複判別手段で重複すると判別された外接矩形の分解を行わないことを特徴とするイメージ処理装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書頁のテキスト領域と非テキスト領域を区別するために該文書頁のイメージを分析するブロック選択（または特徴抽出）装置であって、テキスト領域の各グループの周囲と非テキスト領域の各グループの周囲とに外接する矩形を定め、その後で重複する矩形を同定し、それらの矩形の重複する部分を分解することによって該重複する部分の除去を可能にする装置に関する。

【０００２】

【従来の技術】

従来の特徴抽出装置においては、文書頁の異なるタイプのイメージデータは、まず、該イメージデータがテキストタイプであるかまたは非テキストタイプであるかに応じて同定され、次にイメージデータの領域が、種類に応じて一緒にグループ化（または「ブロック化」）される。イメージデータの各ブロックが、その後で、更なる処理のために抽出される。即ち、従来の特徴抽出装置は、ブロック選択ルーチンを実行して、前記イメージデータ内の連結成分を同定し、非テキストタイプの連結された成分からテキストタイプの連結成分を分離し、別々に該テキスト及び非テキスト連結成分を好ましくは矩形のブロック内にグループ化し、そして非テキスト連結成分を更なる分析に付して、表データ、グラフデータ、線画イメージ、ハーフトーンイメージ、フレーム等の非テキスト連結成分の特定の種類を同定する。

【０００３】

一度イメージデータの特徴が抽出されると階層木を、該木の各ノードがブロック選択中に定められたイメージデータの各ブロック矩形グループに１対１対応するように設けて、設定することができる。階層木の各ノードには、属性情報がイメージデータの各ブロック矩形片毎に記憶される。即ち、該属性情報は、イメージ情報のブロックがテキストタイプであるか非テキストタイプであるかについての情報を含み、テキストタイプ情報は、さらに題名領域、見出し領域、テキスト領域、などに分類され、非テキストタイプ情報は、さらに、表情報、線画情報、グラフ情報、ハーフトーンイメージ情報などに分類されていることが可能である。なお、階層木中のノードの位置は、文書頁内のイメージ情報の対応するブロック矩形の位置を暗に記憶している。文書のイメージ情報内の矩形ブロックの場所との組合せで、階層木により、光学式文字認識（ＯＣＲ）、データ及び／またはイメージ圧縮、データルーチング、データ抽出、保存、検索などのようなその後の適切な処理のための情報の各ブロックの抽出が可能になる。例えば、テキストデータとして指定されたイメージデータのブロックは、適切なＯＣＲ処理に付しうるが、ピクチャデータと指定されたイメージデータのブロックは、データ圧縮に付し得るであろう。その結果、多様な異なる文書頁のいずれのイメージデータであれ、任意に入力し、オペレータの介入なしに自動的に処理可能となる。

【０００４】

イメージデータの各ブロックの秩序だった処理のために、多くの従来の特徴抽出装置は、

ブロックが重複していないという仮定に依存していた。そのような仮定は、図２に示した文書などの、非常に多くの文書において正しい。該図から分かるように、代表的な文書頁１は、２コラムフォーマットで配列され、題名領域２、水平な線３、各々テキストデータの行を含むテキスト領域４、５及び６、ハーフトーンピクチャ領域７、フレーム領域８、ならびに表９を含む。米国特許出願第０７／８７３、０１２号「文字識別の方法及び装置」ならびに米国特許出願第０８／３３８、７８１号「頁分析装置」に記載されたブロック選択技術、によると、文書頁１の各領域が、その中に位置するイメージデータのタイプに従って同定及び指定され、次いでイメージデータがその各々のタイプに基づいて区分される。ブロック選択処理の結果、各イメージ領域の矩形ブロックが図３に示すように作成され、対応する階層木が形成される。即ち、図３に示すように、文書１に対応するイメージデータ１１の場合は、ブロック選択により、題名ブロック１２、テキストブロック１４、１５及び１６などの多様なテキストタイプ領域ならびに線ブロック１３、ハーフトーンイメージブロック１７、フレーム領域１８及び表領域１９等の多様な非テキストブロックが定められる。

【０００５】

図３に示すように、非テキストブロックのあるものはその中にテキストブロックを含みうるが、矩形ブロックのいずれも他のブロックに重複していない。例えば、フレーム領域１８は、非テキスト線画領域１８ａ及びテキスト領域１８ｂを含み、表領域１９は、総じて１９ａで指定されるテキストタイプ表記載事項を含む。

【０００６】

【発明が解決しようとする課題】

上に述べたように、イメージデータ１１内のイメージ領域のブロック矩形指定に基づき、各情報領域のイメージデータをその後の適切な処理のために抽出することができる。さらに、そして同じく前述したように、矩形ブロックのいずれも他と重複していないので、抽出は直接的である。

【０００７】

ブロック矩形領域が互いに他と重複するときは困難が生じる。そのような重複は、例えば、文書の全体の外見に深く影響する編集書式の単純な変更とともに生じ得る。かくして、図４に示すように、フレーム領域８を文書の２つのコラムにまたがる８ａで示される位置に移動する編集書式の単純な変更が行われている。従来のブロック選択及び特徴抽出技術では、図５に示すように、ブロック１８ａがブロック１４及びブロック１５に重なるブロック矩形フォーマットが得られる結果となる。そのように定められた矩形ブロックに基づいてイメージデータを抽出するときになると、これらのブロックの重複は困難を引き起こす。即ち、領域４のテキストイメージデータにのみ対応すると想定されているブロック１４についてイメージデータを抽出する時になると、非テキストフレーム領域１８ａの不要な部分も抽出されることが分かる。同様に、テキストタイプイメージデータのみを含むと推定されている領域１５のイメージデータを抽出するときになると、非テキストフレーム領域１８ａの不要なもう一つの部分が抽出されることが分かる。かくして、特徴抽出またはブロック選択技術により設定されるブロック矩形領域における重複は、現在までのところ困難を生じている。

【０００８】

本発明は、上述の問題点に鑑み、特徴抽出またはブロック選択技術により設定されるブロック矩形領域における重複の問題を克服できるイメージ処理方法及び装置を提供することを目的とする。

【０００９】

【課題を解決するための手段】

本発明のイメージ処理方法は、上記目的を達成するため、入力されたイメージ内に含まれる連結成分に基づいて特定される領域の外接矩形を得る外接矩形取得工程と、前記入力されたイメージ内に含まれる連結成分に基づいて特定される領域の輪郭情報を得る輪郭情報取得工程と、前記外接矩形取得工程で得た複数の外接矩形のうち、重複する外接矩形を

判別する重複判別工程と、前記重複判別工程で重複すると判別された外接矩形が、当該外接矩形に対応する所望の連結成分のイメージデータ抽出に影響するかどうか判別する影響判別工程と、前記重複判別工程で重複すると判別された外接矩形に関して、前記輪郭情報に基づいて複数の矩形に分解する分解工程と、前記分解工程で複数の矩形に分解された場合には、当該分解された複数の矩形に基づいてイメージデータを抽出し、前記分解工程で複数の矩形に分解されなかった場合には、当該分解されなかった外接矩形に基づいてイメージデータを抽出する抽出工程と、前記抽出工程で抽出されたイメージデータを処理する処理工程とを有し、前記分解工程では、前記影響判別工程で影響すると判別された場合、前記重複判別工程で重複すると判別された外接矩形を前記輪郭情報に基づいて複数の矩形に分解し、前記影響判別工程で影響しないと判別された場合、前記重複判別工程で重複すると判別された外接矩形の分解を行わないことを特徴とする。

また、本発明のイメージ処理装置は、入力されたイメージ内に含まれる連結成分に基づいて特定される領域の外接矩形を得る外接矩形取得手段と、前記入力されたイメージ内に含まれる連結成分に基づいて特定される領域の輪郭情報を得る輪郭情報取得手段と、前記外接矩形取得手段で得た複数の外接矩形のうち、重複する外接矩形を判別する重複判別手段と、前記重複判別手段で重複すると判別された外接矩形が、当該外接矩形に対応する所望の連結成分のイメージデータ抽出に影響するかどうか判別する影響判別手段と、前記重複判別手段で重複すると判別された外接矩形に関して、前記輪郭情報に基づいて複数の矩形に分解する分解手段と、前記分解手段で複数の矩形に分解された場合には、当該分解された複数の矩形に基づいてイメージデータを抽出し、前記分解手段で複数の矩形に分解されなかった場合には、当該分解されなかった外接矩形に基づいてイメージデータを抽出する抽出手段と、前記抽出手段で抽出されたイメージデータを処理する処理手段とを有し、前記分解手段は、前記影響判別手段で影響すると判別された場合、前記重複判別手段で重複すると判別された外接矩形を前記輪郭情報に基づいて複数の矩形に分解し、前記影響判別手段で影響しないと判別された場合、前記重複判別手段で重複すると判別された外接矩形の分解を行わないことを特徴とする。

本発明は、上述の困難に対してブロック選択及び特徴抽出処理の結果得られる重複矩形を同定し、該重複矩形をより小さい非重複矩形に分解することにより対処する。好ましくは、重複する矩形が全て分解されるのではなく、むしろ該重複矩形がまず分解が必要かどうかを判別するために分析されて、第１の矩形についてイメージデータを抽出するときに第２の矩形中の不要なイメージデータを抽出するのを回避する。

【００１０】

好ましい形態においては、重複矩形の分解は、各矩形内のイメージデータの輪郭ペアに基づく。即ち、多くの水平の行にわたるイメージデータの場合、隙間のない輪郭が定められ、そこから各イメージ行について輪郭ペアを得ることが可能であるが、該輪郭ペアはその線についてイメージデータの最左端及び最右端を定める。輪郭ペアは外接矩形内のイメージデータの３つの互いに排他的な異なる種類のイメージデータの各々について別様に得られる。即ち、輪郭ペアは、（１）テキストデータ、（２）内部対象物を全く含まない非テキストデータ及び（３）内部対象物を含む非テキストデータについて、別様に得られる。

【００１１】

イメージデータの重複する矩形ブロックは、その重複がしかるべきイメージ抽出に干渉するかどうかを判別するために分析される。もし重複がしかるべきイメージ抽出に干渉しない場合には、分解は行われず、輪郭ペアは用いられない。他方、矩形ブロックにおける重複がしかるべきイメージ抽出に干渉すると判別された場合は、輪郭ペアを用いて重複矩形を分解する。即ち、分解は、関係のある重複矩形の種類に応じて別様に行われる。即ち、分解は、（１）他のタイプ（即ちテキストまたは非テキスト）ブロックに重複するテキストブロック、（２）内部コンテンツを含まず、他の種類のブロックに重複する非テキストブロック、及び（３）内部コンテンツを含み他の種類のブロックに重複する非テキストブロックについて、別様に行われる。

【００１２】

かくして、本発明の１つの形態においては、テキスト領域と非テキスト領域の両方を含むイメージデータを処理する処理方法が提供される。本方法は、ブロック選択を行って、イメージデータ内のテキストタイプ領域の各ブロックの周囲及びイメージデータ内の非テキスト領域の各ブロックの周囲に外接する矩形を得、各テキスト及び非テキストブロックについて輪郭ペアを得るステップを含む。その次に、外接する矩形が重複するかどうが判別される。輪郭ペアに基づいて重複矩形を分解し、非重複矩形の場合、外接矩形に基づいて、重複矩形の場合、分解された矩形に基づいて、イメージデータが抽出される。次に抽出されたイメージデータが処理される。

【００１３】

この簡単な概要は、本発明の本質が速やかに理解されるように提供したものである。本発明のより完全な理解は、添付の図面に基づく好ましい形態の下記の説明を参照することにより得ることができる。

【００１４】

【発明の実施の形態】

図６は、本発明に係る重複矩形分析及び分解を含む本発明の代表的な実施の形態の外観を示す図である。図６には、プログラム化された汎用コンピュータが示されているが、本発明は、他のイメージ処理装置に組み込み可能な、専用、ＲＯＭベースまたは据えつけ型の装置などの他の装置において具体化可能であることは理解されねばならない。

【００１５】

図６には、マイクロソフトＷｉｎｄｏｗｓ ＯＳ等のウインドウ型オペレーションシステムを有するＩＢＭ ＰＣまたはＰＣ互換機等のコンピュータ機器４１０が示してある。コンピュータ機器４１０には、モノクロまたはカラーディスプレイモニタ４１２が設けられ、ユーザに対しイメージを表示する。コンピュータ機器４１０には更に、文書イメージファイル等のデータファイルならびにブロック選択及び重複矩形分析分解プログラム等のアプリケーションプログラムファイルを記憶する固定ディスクドライブ４１１が設けられている。さらにコンピュータ機器４１０には、テキストデータを入力し、ディスプレイ４１２の表示画面に表示される対象の操作を可能にするキーボード４１３と、ディスプレイ４１２に表示される対象を指示したり操作したりするための、マウス等のポインティング・デバイス４１４とが設けられる。

【００１６】

複数頁を有する文書はスキャナ４１６によって入力される。スキャナ４１６は文書の各頁または他のイメージをスキャンして、これらの頁のビットマップイメージデータをコンピュータ機器４１０に供給する。該イメージデータは、圧縮または非圧縮フォーマットでディスク４１１に記憶される。

【００１７】

コンピュータ機器４１０が処理した文書を出力するために従来のカラープリンタ４１８が設けられる。

【００１８】

更に、ローカル・エリア・ネットワークとインターフェースするためのネットワーク・インターフェース４２４、及びファクシミリ／データモデムを介してファクシミリ・メッセージ及び他のデータファイルを送受信するためのファクシミリ／モデム・インターフェース４２６が設けられる。そのようなインターフェースは、ドキュメントイメージデータを入力するためのスキャナ４１６に加えて、またはその代わりに用いることができる。

【００１９】

オペレータの指示に従って且つウインドウ型オペレーションシステムの制御のもとで、デスクトップ・パブリッシング・プログラム、ドローイング・アプリケーション・プログラム、ブロック選択アプリケーション等の記憶されたアプリケーション・プログラムが選択的に起動され、データを処理したり操作したりする。また、オペレータの指示に従って且つこれらの記憶されたアプリケーション・プログラムに基づいて、イメージをモニタ４１２に表示したり、モニタ４１２に表示されているイメージをプリンタ４１８で印刷するよ

うにコマンドが発せられる。

【００２０】

最も典型的には、本発明を具体化するブロック選択プログラムや重複矩形分析・分解プログラムを含むコンピュータディスク４１１に記憶されたアプリケーション・プログラムは、フロッピィディスク、ＣＤ－ＲＯＭ、またはコンピュータ掲示板などのコンピュータ読み出し可能媒体から各アプリケーションをダウンロードしてディスク４１１に記憶したものである。

【００２１】

図７は、コンピュータ機器４１０の内部構成を示す詳細ブロック図である。図７に示すように、コンピュータ機器４１０は、コンピュータバス５２１に接続された、プログラム式マイクロプロセッサ等から成る中央演算処理装置（ＣＰＵ）５２０を有する。また、スキャナ・インターフェース５２２、プリンタ・インターフェース５２３、ネットワーク・インターフェース５２４、ファクシミリ・モデム・インターフェース５２６、ディスプレイ・インターフェース５２７、キーボード・インターフェース５２８、及びマウス・インターフェース５２９が、コンピュータバス５２１に接続されている。

【００２２】

ＣＰＵ５２０がアクセスできるように、ランダム・アクセス・メモリ（ＲＡＭ）等の主メモリ５３０がコンピュータバス５２１に接続されている。特に、ディスク４１１に記憶されたアプリケーション・プログラムに関連する命令等、記憶されたアプリケーション・プログラムの命令列を実行するときに、ＣＰＵ５２０は、これらの命令列を、ディスク４１１（またはネットワーク・インターフェース５２４を介してまたはフロッピィ・ディスク・ドライブ（図示せず）を介してアクセスされる媒体等の他のコンピュータ読み出し可能な媒体）から主メモリ５３０にロードして、主メモリ５３０からこれらの記憶されたアプリケーション・プログラムの命令列を読み出して実行する。

【００２３】

図１は、コンピュータで実現される方法を示すフローチャートであって、文書頁を表す入力イメージが入力され、該入力イメージに対してブロック選択が実行されて文書イメージ内のテキスト及び非テキストブロックの周囲の外接矩形を定め、輪郭ペアを各ブロックについて得、外接矩形を分析してそれらが重複しているか否かを判別し、重複していた場合は、該重複がイメージ抽出に影響するかどうかを判別し、次に、それらがイメージ抽出に影響すると判別された場合は、重複矩形を分解する。即ち、図１に示すように、分解は、重複矩形が、テキストブロック、内部コンテンツを含まない非テキストブロック、または内部コンテンツを含む非テキストブロックに外接するかどうかに応じて別様に行われる。図１に示された処理ステップは、上述のようにコンピュータ読み出し可能な媒体に記憶され、それらの処理ステップを主メモリ５３０にロードしてそこから実行するＣＰＵ５２０によって実行される。

【００２４】

即ち、ステップＳ５０１で、文書頁を表すイメージデータが入力される。前述したように、イメージデータはスキャナ４１６を介して入力してよく、またはディスク４１１上や、ネットワーク・インターフェース４２４もしくはファクシミリ／モデム・インターフェース４２６を介して記憶されたイメージデータファイルから得てもよい。

【００２５】

ステップ５０２において、入力イメージに対してブロック選択が実行されて、テキスト及び非テキストブロックの周囲の外接矩形を得、また各ブロックについて属性情報と位置情報を記憶する階層木を得る。好適なブロック選択技術は、前述の米国出願第０７／８７３，０１２号及び第０８／３８８，７８１号に記載されており、その内容は、完全に記述したのと同じようにここに引用により加入する。

【００２６】

ステップＳ５０３においては、ステップＳ５０２のブロック選択中に既に輪郭ペアが算出されていない程度だけ、輪郭ペアが、ステップＳ５０２でブロック選択により定められた

各ブロック矩形領域について得られる。即ち、各ブロックは、通常多数の水平の線にわたる。ブロックの中に入るそのような行の各々について、該ブロック中の最左端及び最右端の領域を定める輪郭ペアがステップＳ５０３で得られる。輪郭ペアは、矩形ブロック内のデータの種類に応じて別様に得られる。即ち、輪郭ペアは、該ブロックが（１）テキストブロック、（２）内部コンテンツを含まない非テキストブロックまたは（３）内部コンテンツを含む非テキストブロックであるかどうかに応じて、別様に定められる。ステップＳ５０３にしたがって輪郭ペアを得る方法については、図８及び図９を参照して下記において詳述する。

【００２７】

ステップＳ５０４では、ステップＳ５０２で得た外接矩形が重複するかどうかを判別する。外接矩形が重複するかどうかの判別は、図１３を参照して、下記に詳述する。もし重複する外接矩形がなければ、本発明によりさらに処理を行う必要はなく、フローは、ステップＳ５０９までスキップする。

【００２８】

他方、ステップＳ５０４で外接矩形が重複すると判別されたときは、ステップＳ５０５で、重複がイメージ抽出に影響するかどうかを判別する。一般的に、ほとんどの重複は、イメージ抽出に影響するが、中にはイメージ抽出に影響しないものも存在しうる。例えば、重複するけれども、第１の外接矩形が第２の外接矩形のイメージデータを全く含まないかまたはその逆であるという点で、互いに排他的である２つの外接矩形がブロック選択により定められることはあり得る。そのような外接矩形は、重複しているが、イメージ抽出には影響を与えない。重複がイメージ抽出に影響するかどうかを判別する方法についての詳細は、図１４（Ａ）～（Ｃ）を参照して詳述する。

【００２９】

ステップＳ５０５で重複がイメージ抽出に影響を与えないと判別された場合は、本発明によりさらに処理を行う必要はなく、フローは、ステップＳ５０９までスキップして進む。他方、重複がイメージ抽出に影響する場合は、フローは、ステップＳ５０６、Ｓ５０７及びＳ５０８に進み、各ステップで重複ブロックをそれぞれ分解してイメージ抽出に影響しないようにする。

【００３０】

ステップＳ５０６、Ｓ５０７及びＳ５０８では、重複するブロックの種類に応じて重複ブロックが別様に分解される。かくして、ステップＳ５０６では、他の種類のブロック、即ちテキストまたは非テキストブロックに重複するテキストブロックが分解される。ステップＳ５０７では、内部コンテンツを含まない重複非テキストブロックが分解される。そしてステップＳ５０８では、内部コンテンツを含む非テキストブロックが分解される。階層木は、ステップＳ５０６、Ｓ５０７またはＳ５０８においてどのような分解が行われたかに基づいて更新される。

【００３１】

重複テキストブロック及び非テキストブロックをステップＳ５０６またはＳ５０７またはＳ５０８で分解した後で、フローはステップＳ５０９に進み、イメージデータがイメージデータの境界を定める外接矩形に応じて抽出される。かくして、例えば、重複矩形が存在しない状況では、イメージデータは、ステップＳ５０２でブロック選択により定められた外接矩形に応じて抽出される。他方、重複矩形が存在し、ステップＳ５０６、Ｓ５０７またはＳ５０８で分解された場合は、各ブロックのイメージ抽出がそれらのステップで決定された分解ブロックに応じて実行される。

【００３２】

フローは、次に抽出イメージデータが適切に処理されるステップＳ５１０に進む。例えば、上述のように、テキストタイプイメージデータを抽出する時は、適切な処理は、テキストデータ内の文字イメージの同一性を判別するための光学式文字認識であってよい。同様に、非テキスト表領域用の適切な処理には、該表内に含まれるテキスト用のＯＣＲ処理が含まれてよい。さらに別の例として、非テキストハーフトーンイメージデータの適切な処

理には、より小さな記憶領域にハーフトーンピクチャを保存することを可能にする簡単なイメージ圧縮が含まれてよい。

【００３３】

図１の処理は、必要に応じて、ブロック選択及びイメージ抽出が望まれる各文書頁のイメージデータについて繰り返し行われる。

【００３４】

図８は、図１の上記ステップＳ５０３で簡単に説明したように、イメージ情報の各矩形ブロックについて輪郭ペアを得る方法を示すフローチャートであり、図９は、「輪郭ペア」が何を意味するかを説明する図である。

【００３５】

簡単に図９を参照すると、ブロック選択ステップＳ５０２で定められたイメージデータの任意のブロック４０は、垂直方向に多数の走査線を含み、該走査線の各々はブロックを水平に横切って延びている。任意に形作られた連結成分４１は、例えば、文字のイメージまたはその他のイメージであってよいが、やはり多数の走査線を横切って垂直方向に延びている。「輪郭ペア」は、対象物４１を含む各走査線について定められる。走査線上の各輪郭ペア（対象物が違う位置で走査線を横切る場合は各走査線には２個以上の輪郭ペアが存在しうる）は、正確に２つのポイント：対象物が走査線上で始まる第１（または左側）の点と対象物が走査線上で終わる第２（または右側）の点とを含む。かくして、例えば、図９を参照すると、走査線ｉには２つの輪郭ペアが含まれ、第１のものは（ａ，ｂ）からなり、第２のものは（ｃ，ｄ）からなる。輪郭ペア（ａ，ｂ）には、対象物４１が走査線ｉ上で始まる第１（左側）の点ａと対象物４１が輪郭線ｉ上で終了する第２（右側）の点ｂが含まれる。同様に、輪郭ペア（ｃ，ｄ）には、対象物４１が走査線ｉ上で始まる第１（左側）の点ｃと対象物が輪郭線ｉ上で終了する第２（右側）の点ｄが含まれる。走査線「ｊ］の場合は、３つの輪郭ペア、即ち、（ｅ，ｆ）、（ｇ，ｈ）及び（ｋ，ｌ）があり、走査線「ｋ」の場合は、ただ１つの輪郭ペア（ｍ，ｎ）がある。輪郭ペアの内側にあり輪郭ペアを含むイメージは、対象物４１の連結成分に属するイメージである。したがって、対象物のイメージは、各走査線に沿う輪郭ペアにのみもとづいて抽出することができる。

【００３６】

図９からよく理解されるように、イメージ中の各対象物が輪郭ペアで表わされれば、イメージのどの部分が次のイメージ処理のために抽出されなければならないかについてもはや混乱はないであろう。他方、かなりのメモリスペースがイメージ内の各輪郭ペアを記憶するのに必要とされ、特にイメージが何千もの輪郭ペアが必要とされるであろうテキストから構成されている時にはそうである。さらに、輪郭ペアの使用は、ブロック矩形フォーマットでより自然に提示されるイメージの非直覚的な表現になるために、ユーザにとっては不便である。勿論、ブロック矩形フォーマットは、各矩形に対して単に左上の角と右下の角のみが設定に必要であるので、ずっと少ないメモリ記憶装置条件しか要求しない。重複矩形が発生するときに生じる上述の欠点は、下記においてより詳細に説明されるように、部分的には、輪郭ペアのしかるべき使用法により対処される。

【００３７】

図８を参照すると、３つの異なる種類のブロック：テキストブロック、内部成分を含まない非テキストブロック及び内部成分を含むテキストブロックの各々について、輪郭ペアを導出するフローチャートが示されている。よく理解されるように、イメージ内で出会ういかなるブロックもこれらの互いに排他的な３つのカテゴリの一つに属する。図８からさらによく理解されるように、各異なる種類のブロックの輪郭ペアは、それぞれが他の種類のブロックについてから導出される方法とは、別様に導出される。

【００３８】

かくして、ステップＳ６０１では、ステップＳ５０２（図１）で導出された矩形ブロックがテキストブロックであるかどうか判別するために調べられる。テキストブロックである場合は、次に輪郭ペアがステップＳ６０２、Ｓ６０３及びＳ６０４により得られるが、これらのステップにおいては、該ブロックの連結成分の頂部、底部、左側及び右側の縁がま

ず得られ、それらの４つの縁が組み合わされて、ブロック内の全ての連結成分の隙間のない輪郭にされ、輪郭ペアがその隙間のない輪郭から作成される。この処理は、図１０（Ａ）～（Ｆ）に示されている。図１０（Ａ）を参照すると、テキスト連結成分を含む矩形ブロックが描かれている。１２のテキスト連結成分が示されているが、この数字は、通常生じるよりもずっと少ないものであり、簡潔さのためにのみ示してある。図１０（Ｂ）～（Ｅ）では、図８のステップＳ６０２により、全ての連結成分の頂部の縁、底部の縁、左側の縁及び右側の縁がそれぞれ得られる。次に、ステップＳ６０３により図１０（Ｆ）に示されるように、頂部、底部、左側及び右側の縁が組み合わされて全ての連結成分が隙間のない輪郭にされる。４つの縁を組み合わせることにより、隙間のない輪郭が、テキストブロック内のテキストの全てを隙間なく囲む閉じたループに形成されることはよく理解されるであろう。最後に、輪郭ペアが図１０（Ｆ）の隙間のない輪郭の輪郭ペアを得ることにより、得られる（ステップＳ６０４）。

【００３９】

ステップＳ６０２ないしＳ６０４においては、テキストブロックを囲む隙間のない輪郭を形成するため連結成分の矩形境界を用いると、各連結成分基づいて隙間のない輪郭を算出する場合に多すぎる時間を費やすことなく良好な結果が得られるであろうと感じられる。しかし、この人為的に作られた輪郭は、他の対象物の輪郭と重複する結果となった場合に、該重複に関わる連結成分の各々の輪郭ペアを用いて重複する部分を修正することができる。

【００４０】

図８に戻って、ステップＳ６０１で矩形ブロックがテキストブロックでないと判別された場合は、フローは、ステップＳ６０５に進み、矩形ブロックが内部成分を含まない非テキストブロックであるかどうか判別される。矩形ブロックが内部成分を含まない非テキストブロックである場合は、フローはステップＳ６０６及びＳ６０７に切り替わり、そこで、輪郭ペアがブロック内の各連結成分について得られ、そのようにして得られた輪郭ペアが連結される。この処理は、図１１に示されるが、該図においては、３つの任意な連結成分４６、４７及び４８が示されている。輪郭ペアは、該連結成分の各々について導出され、輪郭ペアを結び組み合わせて、全体に４９で示されるように、全非テキストブロック囲む単一の輪郭を形成すべきかどうか考慮される。もし４９で示されるような連結が望まれる場合は、連結のための空の道筋、即ち、重複を引き起こさない道筋が見出されねばならない。その後で、連結された輪郭線が輪郭ペアとして出力される。他方、連結が望まれていない場合は、各分離した連結成分の輪郭ペアが出力される。

【００４１】

図８に戻り、ステップＳ６０５で、矩形ブロックが内部成分を含まない非テキストブロックであると判別されなかった場合は、カテゴリは互いに排他的であるので、該矩形ブロックは必然的に内部成分を有する非テキストブロックでなければならない。したがって、フローは、ステップＳ６０９及びＳ６１０（ステップＳ６０８は単に完全さのために示してあり、実際には実行を要しない）に進む。ステップＳ６０９とＳ６１０においては、非テキスト対象物の外側の輪郭の輪郭ペアのみならず内側白輪郭の輪郭ペアも得られ、ブロックの内部成分を抽出するのを補助する。

【００４２】

即ち、図１２（Ａ）を参照して、フレーム対象物、表対象物または線画対象物等の内部成分を有する非テキスト成分は、対象物の最も外側の輪郭の内側に含まれる白輪郭を有していてよい。内側白輪郭は、内部対象物を抽出する補助のために用いられる。かくして、図１２（Ａ）に示されるように、そしてステップＳ６０９で説明されたように、テキスト成分などの内部成分５２を含む任意の非テキストブロック５０について、まず輪郭ペアが、該非テキスト対象物の最も外側の輪郭を定める連結成分について得られる。したがって、図１２（Ａ）に示してある状況では、輪郭ペアが連結成分５１について得られる。その後、ステップＳ６１０にしたがって最も外側の輪郭の内側白輪郭について輪郭ペアが得られる。そのような内側白輪郭が５３で示されている。（内側白輪郭を得る方法についての詳

細な説明は米国特許出願第０７／８７３，０１２号でなされており、その内容は上述のように引用によりここに加入される。）最も外側の輪郭を定める連結成分と内側白輪郭の両方についての輪郭ペアが、次に、必要な輪郭ペア情報として出力される。

【００４３】

米国特許出願第０７／８７３，０１２号は、図１２（Ｃ）に示す４方向パターンでの内側白輪郭の導出を記述しているが、図１２（Ｄ）に示される８方向パターンで内側白輪郭を導出することも可能である。８方向導出は、非テキストの最も外側の輪郭が単に垂直に配された縁だけではなく斜めの縁も有するような図１２（Ｂ）に示されるような状況の場合に有利である。勿論、該８方向パターンは、図１２（Ａ）に示されるような状況パターンでも使用可能である。８方向導出は、５３ａで示されるように斜めの縁が存在する場合でも、内側白輪郭を良好に定めることを可能にする。しかし、８方向検索パターンが用いられると、各ステップにおいて白輪郭が外側の黒い境界で完全に包まれているかどうかを判別するために白輪郭トレースを検査しなければならない。

【００４４】

要約すると、図８及び図９は、ステップＳ５０３で簡単に触れた処理の詳細な処理を示すが、それによりブロック選択により同定された各矩形ブロックについて輪郭ペアが得られ、またそれにより、輪郭ペアは、該ブロックがテキストブロックであるか、内部成分を含まない非テキストブロックであるか、内部成分を含む非テキストブロックであるかに応じて別様に得られる。

【００４５】

図１３及び図１４（Ａ）～（Ｃ）は、重複する外接矩形があるかどうかを判別し、重複矩形が存在する場合は、重複がイメージ抽出に影響するかどうかを判別する図１の処理ステップＳ５０４及びＳ５０５を説明する図である。

【００４６】

即ち、ステップＳ５０５では、ブロック選択により同定されたブロックに重複するものがあるかどうかが判別される。そのような重複は、各ブロックを定める２つの座標（即ち、左上の角と右下の角）を他のブロックの対応する座標と比較することにより判別可能である。かくして、図１３は、４つのブロック、即ちテキストブロック５４、テキストブロック５５、テキストブロック５６、及び非テキストブロック５７を含む任意の文書のイメージ５３を示す。各ブロックについての左上の角及び右下の角と他のブロックについての対応する座標との比較により、文書５３には重複するブロックがないことが示される。したがって、イメージ５３中のブロックについては分解は必要でなく、文書５３の処理は、直接イメージデータの抽出（ステップＳ５０９）に進むことが可能である。

【００４７】

図１４（Ａ）～（Ｃ）は、矩形ブロックにおける重複がイメージ抽出に影響するかどうかを説明する図である。かくして、図１４（Ａ）においては、任意の文書のイメージ６０は３つのブロック、即ち、ピクチャブロック６１、ピクチャブロック６２、及びテキストブロック６３を含む。テキストブロック６３の左上の角及び右下の角の座標（即ち、座標（Ｘ１，Ｙ１）及び（Ｘ２，Ｙ２））の比較により、テキストブロック６３が非テキストブロック６１及び非テキストブロック６２に重複していると判別される。しかし、図１４（Ａ）の状況下では、テキストブロック６３がブロック６１及び６２に重複しているといっても、テキストブロックがそれらに重複している領域には、ブロック６１及び６２のイメージデータは存在しない。かくして、重複にも関わらず、ブロック６３のイメージ抽出は影響を受けず、テキストブロック６３の分解は必要とされない。他方、非テキストブロック６１及び６２のイメージ抽出は、両方とも、重複により影響される。即ち、ブロック６１の矩形座標に基づいてイメージデータを抽出する場合は、必要なピクチャデータが抽出されるのみならずブロック６３から不要なテキストデータ部分も抽出されるであろう。したがって、非テキストブロック６１及び６２の両方について、ブロック分解がステップＳ５０６、Ｓ５０７またはＳ５０８にしたがって適切な処理として、必要となる。（ここで、非テキストブロック６１及び６２は内部成分を含まないと仮定すると、その場合は、分

解処理はステップＳ５０７にしたがって行われるであろう）。
【００４８】
図１４（Ｂ）は、テキストブロック６５及び非テキストブロック６６を含む任意の文書６４について、重複がブロック６５と６６間に存在する状況を示す。この状況では、図１４（Ａ）の状況とは違って、重複のために、ブロック６５または６６のいずれについてもイメージ情報を抽出するのは、同じく他のブロックについての不要なイメージデータを得ることなしには不可能である。したがって、ブロック６５及び６６の両方がステップＳ５０６ないしＳ５０８にしたがって分解に付される。
【００４９】
図１４（Ａ）及び（Ｂ）に示される重複は、編集スタイルによるものであり、グラフィックが単一のページ上でテキストと混合され、インデントされ位置決めされたために生じたものである。しかし、重複は他の原因によるものであり得、したがって、編集スタイルによる重複に厳格に限定されるべきものではない。例えば、重複は、イメージデータ内のスキュー（斜行）によっても発生可能であり、スキューは、意図的なもの（やはり編集スタイルによる）またはある角度で文書を走査したことによる非意図的なもののいずれかであり得る。この状況は、任意の文書６７が第１のテキストブロック６８及び第２のテキストブロック６９を含む図１１Ｃに示されている。両方のテキストブロックが斜めになっており、そのスキューのため、ブロック６８と６９の間に重複が形成される。その重複のために、他のブロックからの不要な情報も抽出することなしには一つのブロックからのイメージ情報を抽出できない。したがって、ブロック６８及び６９の両方の分解が必要である。
【００５０】
［重複するテキストブロックの分解］
他のテキストブロックまたは他のイメージタイプのブロックに重複するテキストをしかるべく抽出するために、抽出されるべきテキストを含むテキストブロックがより小さい非重複の矩形に分解される。ステップＳ５０６について上で論じた重複するテキストブロックを分解する工程は、図１５（Ａ）～（Ｉ）、１６（Ａ）及び（Ｂ）、図１７～図１９、ならびに図２０（Ａ）及び（Ｂ）を参照してより詳細に下記で論じる。
【００５１】
一度重複テキストブロック領域が存在すると判別され、該重複テキストブロック領域がイメージ抽出に影響すると判別された場合は、重複状態がどのように存在しているかが、どの領域が直ぐに分解されるべきか（非重複領域）そしてどの領域がさらに分解を要するか（重複領域）を判別するために吟味される。図１５（Ａ）～（Ｉ）に示されるように、２つのブロックに重複が発生する仕方には９通りの例がある。（これらの９つの例の鏡像及び回転が存在してもよい）。例えば、２つの矩形が何らかの重複を有する場合、図１５（Ｄ）に示す重複状態等のように、その水平及び垂直の縁が完全にまたは一部重複しているかも知れず、一つの矩形のいずれかの縁が、他の矩形の縁内に完全に入っているかも知れない。
【００５２】
重複領域に関わる領域は、図１５（Ａ）～（Ｃ）及び１５（Ｅ）～（Ｉ）の各重複状態に示されるように、少なくとも２つのそして多くとも４つの矩形に直ぐに分解可能である。勿論、図１５（Ｄ）に示すように、２つのブロックが完全に互いに重複する場合がある。分解の第１のステップは、重複していない領域を分解して１以上の非重複矩形に分解することにより刈り込んで取り去ることである。図１５（Ａ）～（Ｃ）及び１５（Ｅ）～（Ｉ）に示すように、非重複領域は非ボールド体の輪郭で示されている。例えば、図１５（Ａ）に示されるように、テキストブロック１２０は、重複しない第１及び第２の領域に分解可能であり、テキスト領域１２１は、テキストブロック１２０と重複しない第１及び第２の領域に分解可能である。その結果、直ぐに分解可能な４つの非重複領域が得られる。しかし、重複領域１２２の場合は、更なる分解が必要となる。この工程は下記においてより詳細に論じられるであろう。
【００５３】

まず、明瞭さのために、分解されるべきテキストブロックを「テキストブロック」と呼び、対象物または該テキストブロックの矩形領域が重複する非テキスト領域を「重複対象物」と呼ぶ。この場合に重複対象物は、テキストブロック、非テキストブロック、または既に分解された矩形でありうる。重複対象物が分解された矩形である場合は、最初の分解処理が、テキストブロックを分解するのに十分であるべきである。重複ブロックがテキストまたは非テキストブロックであれば、分解の第２のステップが、テキストまたは非テキストブロックの「輪郭ペア」を用いて必要となるであろう。

【００５４】

さて、図１６（Ａ）を参照すると、非テキストブロック１３１に重複するテキストブロック１３０の例が示されている。図１５（Ａ）で注意した重複状態のように、イメージ抽出されるべきテキストが重複領域１３２内に存在し、一度非重複ブロックが分解されるとその結果図１６（Ｂ）に示す重複領域１３２が得られる。重複領域１３２が次に、より小さい非重複矩形を作るために更なる分解に付され、テキストブロックが重複領域からすぐに抽出することが可能となる。

【００５５】

かくして、重複領域１３２をさらに分解する工程を図１７～図１９に示すフローチャートを参照して論じる。ステップＳ１４００においては、領域１３２内のテキストブロックの全ての連結成分が集められる。一度、それらが集められてしまうと、各成分が図２０（Ａ）に示すような重複領域に入るように刈り込まれる。例えば、図２０（Ａ）に示すように、ブロック１５１が刈り込まれ、重複領域内のブロックの部分のみが残る。即ち、ブロック１５１について図２０（Ａ）に示されたものは、ブロック全体の一部にすぎない。同じことが、ブロック１５２及び非テキストイメージ１５３の残余の部分についても当てはまる。

【００５６】

ステップＳ１４０１では、重複領域１３２内にある重複非テキスト対象物の全ての輪郭ペアが集められ、上述したようにして非重複部分が重複部分から刈り込まれる。ステップＳ１４０２では、連結成分が重複非テキストイメージ内に入らない矩形の組にグループ化される。この点については、ステップＳ１４０２での連結成分を非重複矩形の組にグループ化する処理は二つの異なる方法、即ち方法Ａ及び方法Ｂによって行うことが可能である。方法Ａにおいては、領域１３２内のテキストブロックの刈り込まれた連結成分がステップＳ１４０４で用いられ、方法Ｂにおいては、重複テキストブロックの刈り込まれた連結成分から導出される輪郭ペアが用いられる。方法Ａまたは方法Ｂのいずれかより、テキストブロックが非テキストイメージに重複しない矩形にさらに分解される。

【００５７】

かくして、方法Ａの分解を用いて、ステップＳ１４０５において、テキストブロックが水平なテキストブロックであるかどうか判別される。ステップＳ１４０５において、水平なテキストブロックであると判別されたときは、フローは、ステップＳ１４０７に進み、そこで、水平方向に沿う全ての連結成分が一緒にグループ化される。一度水平方向の全ての連結成分がグループ化されると、フローは、ステップＳ１４０８に進み、そこで、垂直方向に沿う全ての連結成分がグループ化される。

【００５８】

ステップＳ１４０９において、重複非テキスト対象物の輪郭ペアのいずれとも重複しない重複領域内の連結成分が残存しているかどうかが判別される。もし輪郭ペアと重複しない更なる連結成分が存在しないならば、ステップＳ１４１２において、グループ化処理が終了する。グループ化された水平方向連結成分及びグループ化された垂直方向連結成分が次いで、組み合われて４つの非重複矩形１５４、１５５、１５６及び１５７を形成する。これらの矩形の座標は、非重複領域内のテキストと同様にして重複領域からテキストを抽出するのに利用される。しかし、もっと多くの連結成分が存在する場合は、フローはステップＳ１４０７に戻る。

【００５９】

ステップＳ１４０５で、テキストブロックが水平方向のテキストブロックではない場合は、フローは、ステップＳ１４１４に進み、そこで連結成分がまず垂直方向に沿って連結される。一度、それらが垂直方向に沿って連結されると、フローは、ステップＳ１４１５に進み、その時点で、全ての連結成分が水平方向に沿って一緒にグループ化される。ステップＳ１４０９におけると同様に、ステップＳ１４１６において、重複非テキストイメージの輪郭ペアのいずれとも重複しない連結成分が残存しているかどうか判別される。もし重複対象物の輪郭ペアのいずれとも重複しない成分が残存していないと判別された場合は、グループ化処理がステップＳ１４１２で終了する。グループ化された水平方向連結成分及びグループ化された垂直方向連結成分は、次に、組み合わされて非重複矩形を形成する。これらの矩形の座標は、記憶され、非重複領域のテキストと同様に重複領域からテキストを抽出するのに使用される。重複しない連結成分が存在する場合は、フローはステップＳ１４１４に戻る。

【００６０】

図１８の連結成分のグループ化の結果、今や更なる処理のために抽出可能となったブロック化された領域が得られる。更なる分解により今度は重複領域内のテキストデータが直ぐにしかるべき後処理のために抽出可能となる。

【００６１】

一方、方法Ｂが用いられた（ステップＳ１４０６）場合は、ステップＳ１４２０で、輪郭ペアが集めれらた連結成分から導出されるか、即ちテキストブロックの輪郭ペアが図２０（Ｂ）に示すように重複領域に入るように刈り込まれる。なお、該グループ化は、テキストブロックの連結成分ではなくて、輪郭ペアにより生じるため、図２０（Ａ）に示されているのとは少し異なっている。一度輪郭ペアが刈り込まれた連結成分から導出されると、フローは、ステップＳ１４２１に進む。ステップＳ１４２１では、全ての垂直方向の連結された輪郭ペアが一緒にグループ化される。ステップＳ１４２２では、一度連結された垂直方向の輪郭ペアの全てが垂直方向に沿ってグループ化されると、次に水平方向の輪郭ペアが一緒にグループ化される。ステップＳ１４２３では、垂直方向及び水平方向グループ輪郭ペアから作られた矩形が重複対象物の輪郭ペアのいずれとも重複しないような垂直方向又は水平方向の連結輪郭ペアが残存するかが判別される。もし、ステップＳ１４２３において、残存する連結輪郭ペアが存在しないと判別されると、グループ化処理はステップＳ１４２４で終了する。垂直方向及び水平方向両方のグループ化された連結輪郭ペアは重複対象物の輪郭ペアのいずれともまた前に分解された矩形の非グループ化輪郭ペアのいずれとも重複しない矩形に分解される。

【００６２】

上述の処理の結果として、重複テキストブロック内のテキストは、さらにテキストブロックの重複領域をより小さな非重複矩形に分解することにより抽出することができる。

【００６３】

［内部コンテンツを含まない非テキストブロックの分解］

上記ステップＳ５０７で論じた内部コンテンツを含まない重複非テキストブロックを分解する処理を図２１（Ａ）～（Ｃ）及び図２２を参照してさらに詳細に論じる。

【００６４】

非テキストブロックの矩形領域がもう一つの非テキストブロックと重複する場合は、非テキストブロックの一方をより小さな非重複矩形に分解することが可能である。より小さな非重複矩形の組によって提供される情報に基づいて該ブロック内の重複非テキストイメージを直ぐに抽出することが可能である。

【００６５】

明瞭さのために、分解されるべき非テキストブロックを「非テキストブロック」と呼び、該非テキストブロックにより重複されている領域を「重複対象物」と呼ぶ。

【００６６】

さて、図２１（Ａ）を参照すると、非テキストイメージブロック１６０及び１６１が領域１６２で重複している。上記において重複テキストブロックについて論じたように、非重

複領域は直ぐに最大４つの非重複矩形に分解される。非テキストブロック１６０及び１６１の「輪郭ペア」に基づいて、図２２のフローチャートに示される処理を用いて、重複領域１６２、非テキストブロック１６０及び１６１内のより小さな非重複矩形の組を作ることが可能である。

【００６７】

かくして、ステップＳ１７００において、非テキストブロック１６０及び重複対象物１６１の輪郭ペアを用いて、非重複領域を刈り込んで領域１６２内に入るイメージのみが残存するようにする。即ち、図１５（Ａ）を参照して前述したように、矩形１２０及び１２１が重複領域１２２から刈り込まれる。同様に、ブロック１２３及び１２４が刈り込まれ、重複領域のみが残される。

【００６８】

図２１（Ｃ）に示されるように、重複非テキストブロックが刈り込まれ、分解される。ステップＳ１７０１においては、重複領域１６２において輪郭ペア情報が存在するか否かが判別される。輪郭ペア情報が重複領域に存在しないとステップＳ１７０２で判別されると、さらに分解を行う必要はなく、処理が終了する。しかし、重複領域１６２に輪郭ペア情報が存在するとフローはステップＳ１７０３に進む。

【００６９】

ステップＳ１７０３においては、矩形重複領域（重複領域の境界領域）に接触する全ての輪郭ペアが非テキストブロック及び重複対象物の両方について集められる。それらの集められた輪郭ペアは図２１（Ｂ）に示す重複領域内に入るように刈り込まれる。

【００７０】

次に、ステップＳ１７０４においては、非テキストブロック１６０の刈り込まれた輪郭ペアが集められる。ステップＳ１７０５では、非テキストブロック１６０及び重複ブロック１６１の重複領域１６２が、連結された垂直方向の輪郭ペアをグループ化し、次いで、垂直方向に沿うグループ化が終了した後で水平方向の輪郭ペアがグループ化されることにより、分解される。一度垂直方向の輪郭ペアがグループ化され、水平方向の輪郭ペアがグループ化されると、重複対象物１６１の輪郭ペアのいずれとも重複しないように矩形が形成される。非テキストブロック１６０の重複領域１６２を分解した結果を図２１（Ｃ）に示す。

【００７１】

図２１（Ｃ）に示すように、非テキストブロック１６０の重複領域１６２は２つのより小さい非重複矩形１６３及び１６４に分解されている。

【００７２】

［内部コンテンツを含む非テキストブロックの分解］

ステップＳ５０８で説明したように、テキストなどの内部コンテンツを含む非テキストブロックを分解する処理について図２３（Ａ）～（Ｄ）及び図２４ならびに図２５（Ａ）及び（Ｂ）を参照してより詳細に論じる。

【００７３】

初めは、フレーム、表、ピクチャ等のブロックには、その矩形領域が、該フレーム、表またはピクチャ内の白輪郭によって包まれる異なる種類のイメージデータを包んでいるものがあってもよい。例えば、フローチャートの場合は、その中にテキスト（アクション）を含む非テキストイメージ（工程ボックス）があり、したがって、イメージは、イメージまたはテキストが、他を抽出しないでしかるべく抽出され得るように、テキストとは別に定められねばならない。

【００７４】

コンテンツを有するまたは有しない非テキストイメージをしかるべく抽出するために、コンテンツを有する非テキストブロックがコンテンツと重複しない最小数のより小さな外接矩形に分解されねばならない。それらのより小さな外接矩形から、内容イメージというよりむしろブロック化されたイメージがしかるべく抽出可能になる。

【００７５】

さて図２３（Ａ）を参照すると、テキストコンテンツを含む非テキストイメージの例が示されている。図２３（Ａ）は、非テキストイメージ１８０～１８２を含むフローチャートの例である。非テキスト領域１８０、１８１及び１８２の各々の内側に、それぞれ３つの白輪郭領域１８６、１８７及び１８８があり、各白輪郭内に、テキストブロック１８３、１８４及び１８５がある。

【００７６】

図２３（Ｂ）に示される非テキストイメージのみを得るために、非テキストイメージ１８０、１８１及び１８２をより小さな外接矩形によって分解できるように、非テキストイメージ、白輪郭及びテキストブロックの輪郭ペアが集められる。

【００７７】

かくして、図２３（Ｂ）に示すイメージをしかるべく得るために、下記のステップが図２４のフローチャートに示されたように実行される。

【００７８】

図２３（Ｂ）に示されるイメージの分解を開始するために、非テキストイメージの各行の輪郭ペアが、該輪郭ペアの一つを白輪郭の輪郭ペアで置き換えることにより修正される。例えば、図２３（Ｃ）に示すように、行ｉの輪郭ペアは、白輪郭の対応する行で各行の輪郭ペアを置き換えることにより修正される。即ち、該修正に先立って、非テキストイメージ１８０及び１８１の輪郭ペアは、それぞれ（ａ，ｂ）及び（ｃ，ｄ）であり、白輪郭１８６及び１８７の輪郭ペアはそれぞれ（ｅ，ｆ）及び（ｇ，ｈ）である。一度輪郭ペアが、各行の輪郭ペアを対応する白輪郭の行に置き換えることにより修正されると、非テキストイメージの輪郭ペアは、（ａ，ｅ）、（ｆ，ｂ）、（ｃ，ｇ）及び（ｈ，ｄ）となる。

【００７９】

各行の輪郭ペアを修正した後で、ステップＳ１９０１において、修正が非テキスト対象物の各輪郭ペアについて繰り返される。一度輪郭ペアの全ての行の修正が完了すると、ステップＳ１９０３において、修正された輪郭ペアは、ピクチャそのものがしかるべく抽出可能となるより小さな外接矩形の組にグループ化される。即ち、ステップＳ１９０３では、グループ化処理が、まず垂直方向に連結された輪郭ペアのグループ化が行われ、次に、垂直方向に沿うグループ化が終了した後で、水平方向に連結された輪郭ペアの全てのグループ化が行われる。全体の外接矩形が、テキストブロック１８３、１８４及び１８５、他のグループ化されない輪郭ペアまたは他の前に形成された矩形等の内部ブロック内容物の輪郭ペアのいずれとも重複しないように、さらに修正された輪郭ペアを選択することが不可能になると、グループ化処理は終了する。一度垂直方向に連結された輪郭ペア及び水平方向に連結された輪郭ペアが全て連結されてしまうと、より小さな外接矩形が、非テキストイメージ１８０、１８１及び１８２の周囲に作られる。得られた分解イメージは図２３（Ｄ）に示される。その後、矩形の座標が更なる処理のために記憶され、分解が終了する。

【００８０】

図２３（Ｄ）に示すように、図２３（Ａ）のイメージは、最小１１個の外接矩形に分解されている。記憶されたこれらの１１個の外接矩形の座標を利用して、該１１個の矩形、１８０、１８１及び１８２の非テキストイメージがその中のコンテンツを抽出することなくしかるべく抽出可能になる。

【００８１】

図２５（Ａ）は、テキストやピクチャデータ等の、内部に含まれるブロックからイメージデータを抽出するためにどのようにフレームイメージが分解できるかを示す１例である。図２５（Ｂ）は、上記の方法を用いて、テーブルイメージを、テーブルの内容物を抽出することなく抽出できるようにするために、テーブルを分解することができる方法を示す１例である。

【００８２】

図２６（Ａ）は、不規則な形状の輪郭１９２内に含まれるテキストブロック１９１を示す。そのような状況では、テキストブロックは、非テキスト輪郭用のブロックに重複しがちである。上述の分解により、テキストブロック１９１を非重複領域１９１ａ，１９１ｂ及

び１９１ｃに分解可能であり、その全てが図２６（Ｂ）に示されるような輪郭１９２の白輪郭内にある。そのような分解により、テキストブロックのイメージデータを抽出するときに、輪郭の不要なイメージデータも間違って抽出されることが、確実になくなる。

【００８３】

以上、本発明を特定の態様に関して説明したが、本発明は上記記載に限定されることはなく、抽出されるべき全てのタイプのイメージデータに適用されることができるものと理解されるべきものである。更に、発明の精神及び範囲から逸脱することなく、当業者によって種々の変更や修正が可能である。

【図面の簡単な説明】

【図１】重複矩形を分析し分解する方法を示すフローチャートである。

【図２】文書頁の代表的な図である。

【図３】図２にブロック選択処理を行って得られた矩形ブロックの図である。

【図４】文書頁の代表的な図である。

【図５】図４にブロック選択処理を行って得られた矩形ブロックの図である。

【図６】本発明を具体化する装置の概観を示す斜視図である。

【図７】図６の装置のブロック図である。

【図８】イメージデータ内の輪郭ペアを得る方法を説明するフローチャートである。

【図９】輪郭ペアを説明する図である。

【図１０】テキストタイプイメージデータの場合に、輪郭ペアを得る方法を説明する図である。

【図１１】内部成分を含まない非テキストイメージデータの輪郭ペアを得る方法を説明する図であり（Ａ）は元のピクチャを示し、（Ｂ）は輪郭を結合した後のピクチャを示す。

【図１２】（Ａ）及び（Ｂ）は、内部成分を含む非テキストイメージデータの場合に輪郭ペアを得る方法を説明する図であり、（Ｃ）及び（Ｄ）は、それぞれ４方向及び８方向輪郭トレース間の相違を説明する図である。

【図１３】重複しない外接矩形を示す図である。

【図１４】重複輪郭がイメージ抽出に影響を与えるかどうかを判別するための分析を説明する図である。

【図１５】２つのブロックがどのようにして重複可能であるかを説明する図である。

【図１６】（Ａ）は、非テキストブロックに重複するテキストブロックの例を示し、（Ｂ）は、（Ａ）に示す重複領域の刈り込まれたものを示す図である。

【図１７】重複テキストブロックをより小さな矩形に分解する方法を示すフローチャートである。

【図１８】図１７のフローチャートの続きである。

【図１９】図１７のフローチャートの続きである。

【図２０】それぞれ重複するテキストブロック領域を分解する二つの方法を示す図である。

【図２１】二つの重複非テキストブロックを説明する図である。

【図２２】重複非テキストイメージを分解する方法を示すフローチャートである。

【図２３】内部コンテンツを含む非テキストブロックを分解する方法を説明する図である。

【図２４】内部コンテンツを含む非テキストブロックをより小さな外接矩形に分解する方法を示すフローチャートである。

【図２５】（Ａ）分解されたフレーム及び（Ｂ）分解された表の例をそれぞれ示す図である。

【図２６】（Ａ）不規則な形状の輪郭に含まれるテキストブロック及び（Ｂ）その分解後を説明する図である。

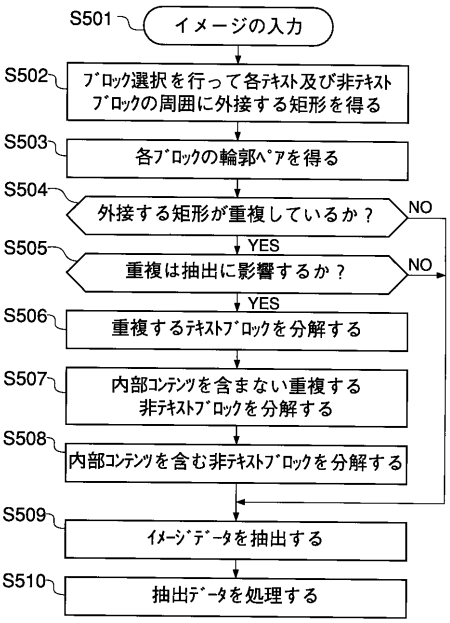
【符号の説明】

４１０　コンピュータ機器

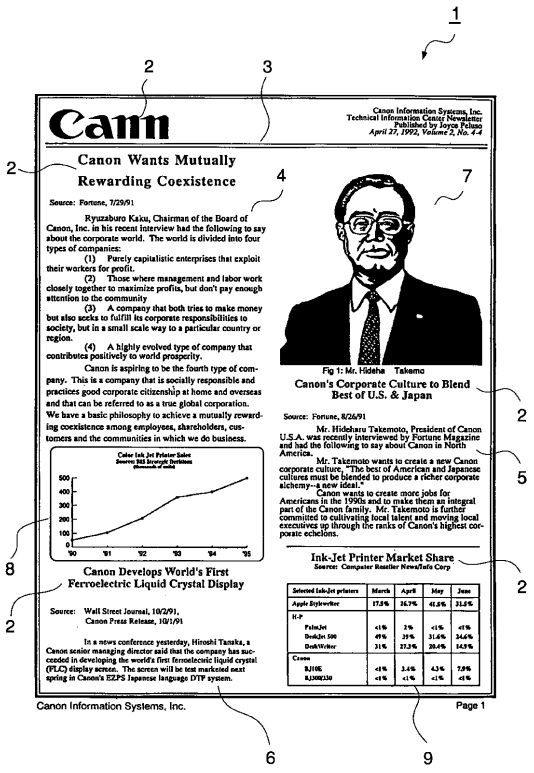
４１１　ディスク

５２０　ＣＰＵ
５２２　スキャナ・インターフェース

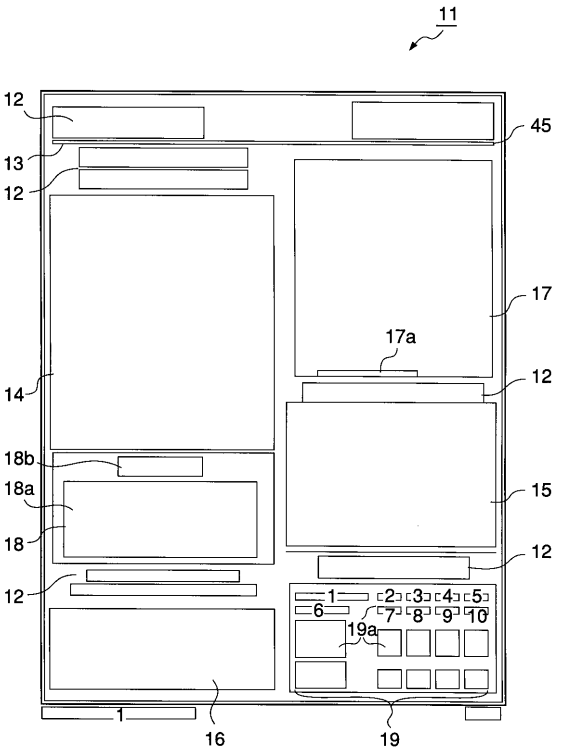
【図１】



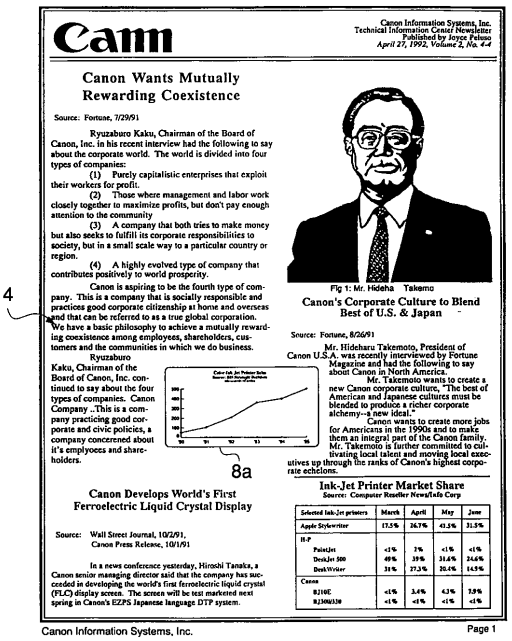
【図２】



【図３】



【図４】


Canon

Canon Information Systems, Inc.
Technical Information Center Newsletter
Published by Joyce Peluso
April 27, 1992, Volume 2, No. 4-4

Canon Wants Mutually Rewarding Coexistence

Source: Fortune, 7/29/91

Ryuzaburo Kaku, Chairman of the Board of Canon, Inc. in his recent interview had the following to say about the corporate world. The world is divided into four types of companies:

(1) Purely capitalistic enterprises that exploit their workers for profit.

(2) Those where management and labor work closely together to maximize profits, but don't pay enough attention to the community

(3) A company that both tries to make money but also seeks to fulfill its corporate responsibilities to society, but in a small scale way to a particular country or region.

(4) A highly evolved type of company that contributes positively to world prosperity.

Canon is aspiring to be the fourth type of company. This is a company that is socially responsible and practices good corporate citizenship at home and overseas and that can be referred to as a true global corporation. We have a basic philosophy to achieve a mutually rewarding coexistence among employees, shareholders, customers and the communities in which we do business.

Ryuzaburo Kaku, Chairman of the Board of Canon, Inc. continued to say about the four types of companies. Canon Company ..This is a company practicing good corporate and civic policies, a company concerned about it's employees and shareholders.

Canon Develops World's First Ferroelectric Liquid Crystal Display

Source: Wall Street Journal, 10/2/91, Canon Press Release, 10/1/91

In a news conference yesterday, Hiroshi Tanaka, a Canon senior managing director said that the company has succeeded in developing the world's first ferroelectric liquid crystal (FLC) display screen. The screen will be test marketed next spring in Canon's EZPS Japanese language DTP system.

Fig 1: Mr. Hideha Takemo

Canon's Corporate Culture to Blend Best of U.S. & Japan

Source: Fortune, 8/26/91

Mr. Hidcharu Takemoto, President of Canon U.S.A. was recently interviewed by Fortune Magazine and had the following to say about Canon in North America.

Mr. Takemoto wants to create a new Canon corporate culture, "The best of American and Japanese cultures must be blended to produce a richer corporate alchemy--a new ideal."

Canon wants to create more jobs for Americans in the 1990s and to make them an integral part of the Canon family. Mr. Takemoto is further committed to cultivating local talent and moving local executives up through the ranks of Canon's highest corporate echelons.

Ink-Jet Printer Market Share
Source: Computer Reseller NewsFacts Corp

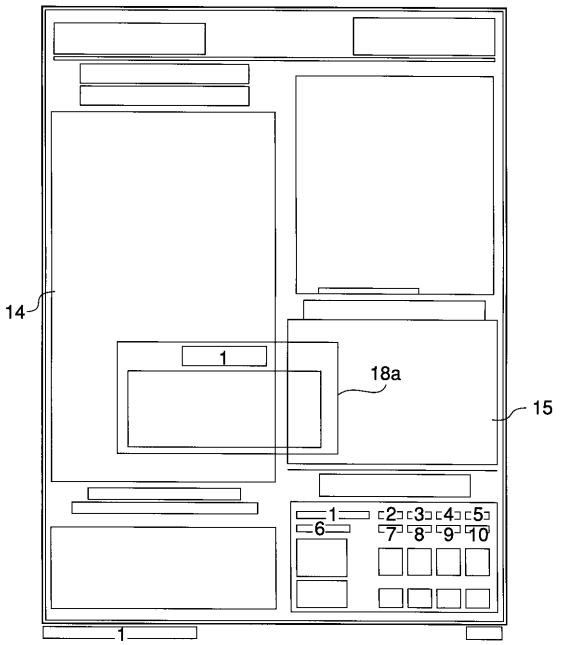
Selected Ink-Jet printers	March	April	May	June
Apple StyleWriter	17.5%	26.7%	43.3%	31.5%
H-P				
PaintJet	<1%	3%	<1%	<1%
DeskJet 500	49%	39%	31.6%	24.6%
DeskWriter	31%	27.3%	20.4%	14.5%
Canon				
BJ10E	<1%	3.4%	4.3%	7.9%
BJ300/330	<1%	<1%	<1%	<1%

Canon Information Systems, Inc. Page 1

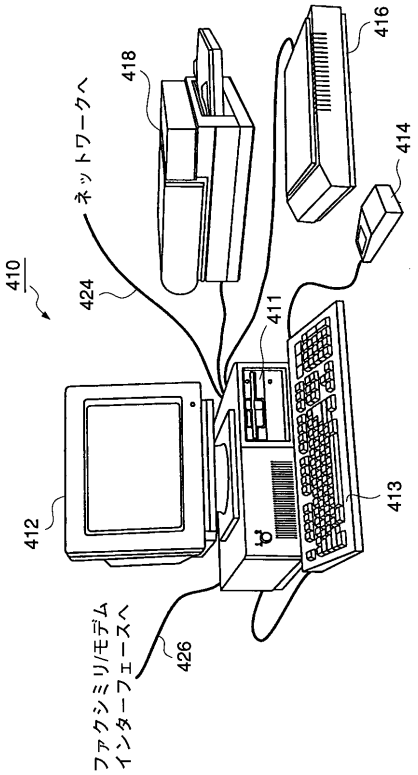
4

8a

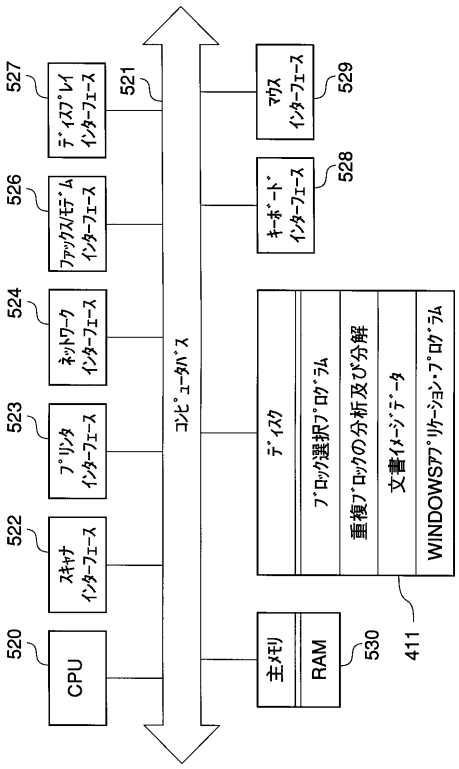
【図５】



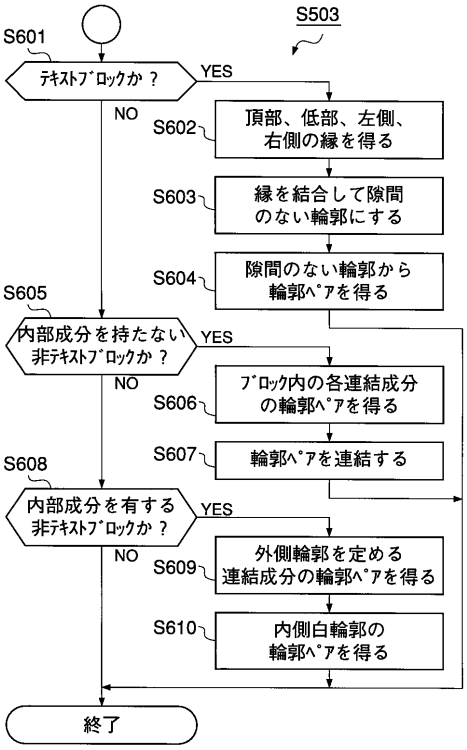
【図６】



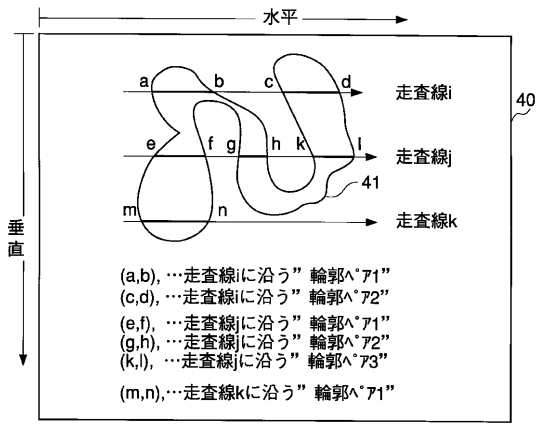
【図７】


CPU
520
スキャナ インターフェース
522
プリンタ インターフェース
523
ネットワーク インターフェース
524
ファックス/モデム インターフェース
526
ディスプレイ インターフェース
527
521
コンピュータバス
マウス インターフェース
529
キーボード インターフェース
528
ディスク
プログラム選択プログラム
重複ブロックの分析及び分解
文書イメージデータ
WINDOWSアプリケーション・プログラム
411
主メモリ
RAM
530

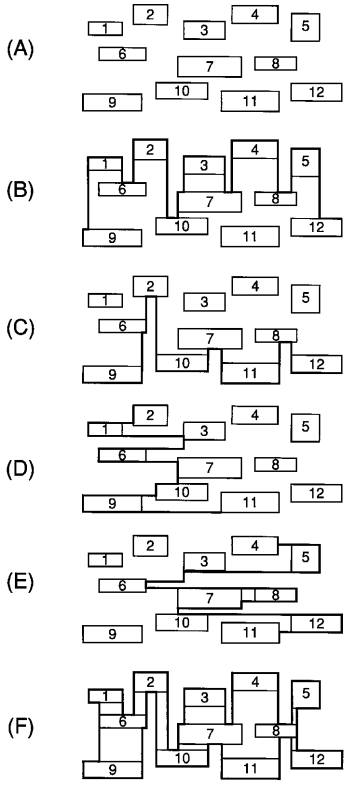
【図８】


S503
S601
テキストブロックか？
YES
NO
S602
頂部、低部、左側、右側の縁を得る
S603
縁を結合して隙間のない輪郭にする
S604
隙間のない輪郭から輪郭ペアを得る
S605
内部成分を持たない非テキストブロックか？
YES
NO
S606
ブロック内の各連結成分の輪郭ペアを得る
S607
輪郭ペアを連結する
S608
内部成分を有する非テキストブロックか？
YES
NO
S609
外側輪郭を定める連結成分の輪郭ペアを得る
S610
内側白輪郭の輪郭ペアを得る
終了

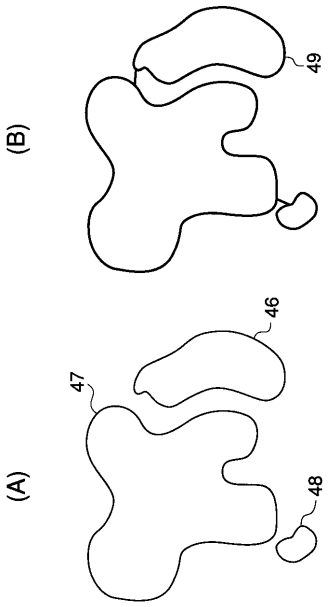
【図９】


水平
垂直
a
b
c
d
走査線i
e
f
g
h
k
l
走査線j
41
m
n
走査線k
40
(a,b), …走査線iに沿う” 輪郭ペア1”
(c,d), …走査線iに沿う” 輪郭ペア2”
(e,f), …走査線jに沿う” 輪郭ペア1”
(g,h), …走査線jに沿う” 輪郭ペア2”
(k,l), …走査線jに沿う” 輪郭ペア3”
(m,n),…走査線kに沿う” 輪郭ペア1”

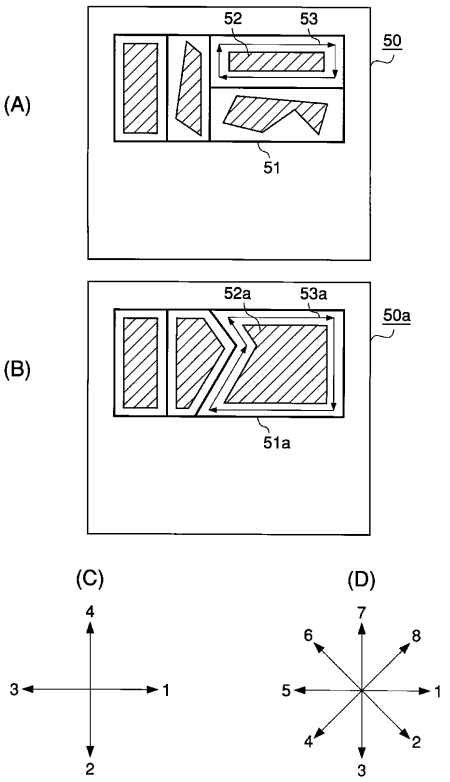
【図１０】


(A)
(B)
(C)
(D)
(E)
(F)

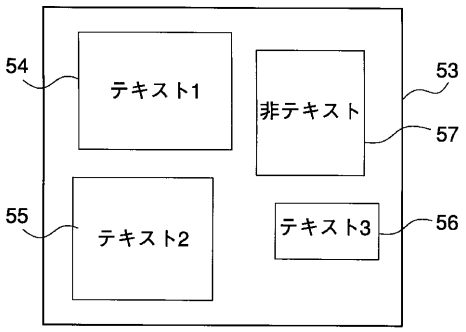
【図１１】



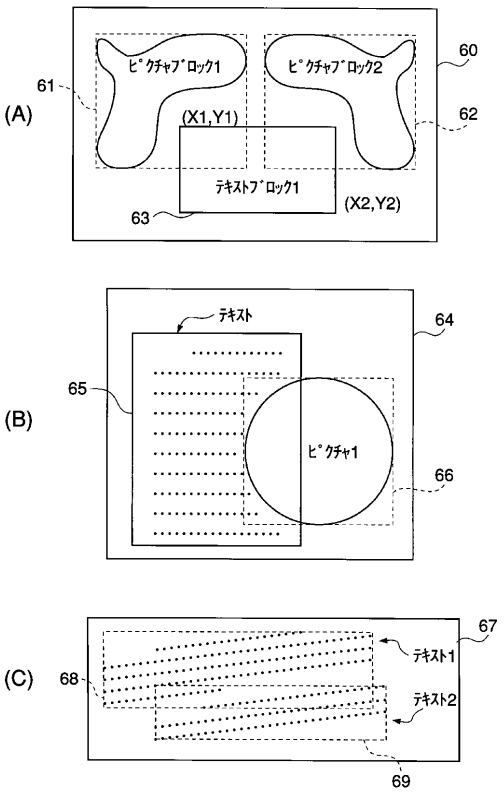
【図１２】



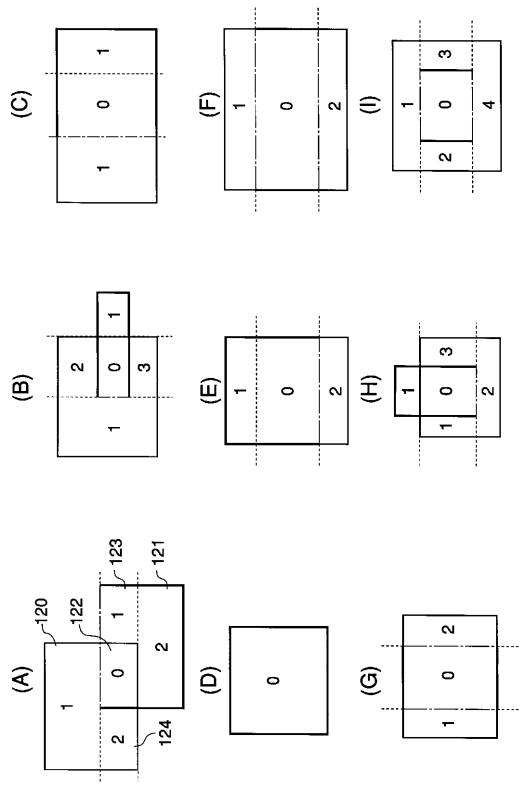
【図１３】



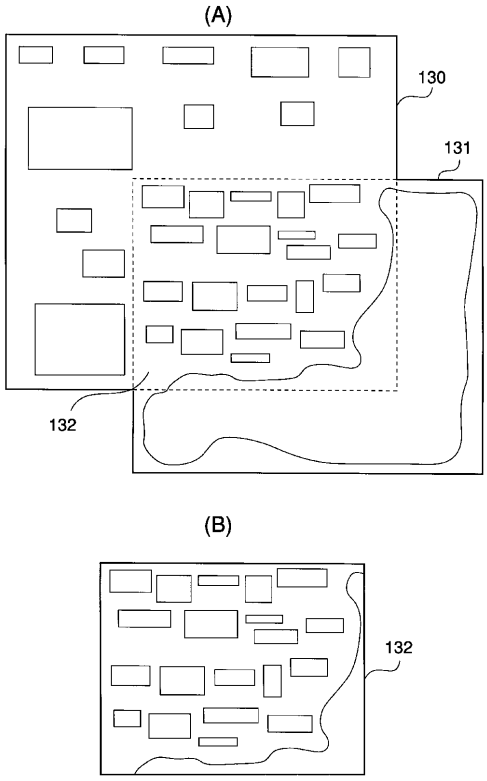
【図１４】



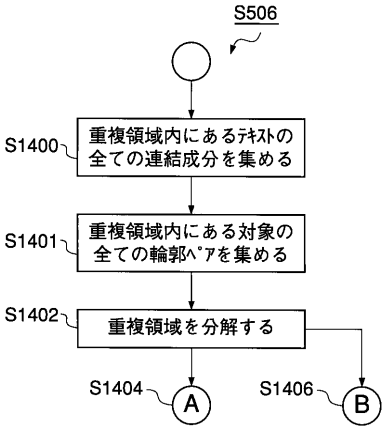
【図１５】



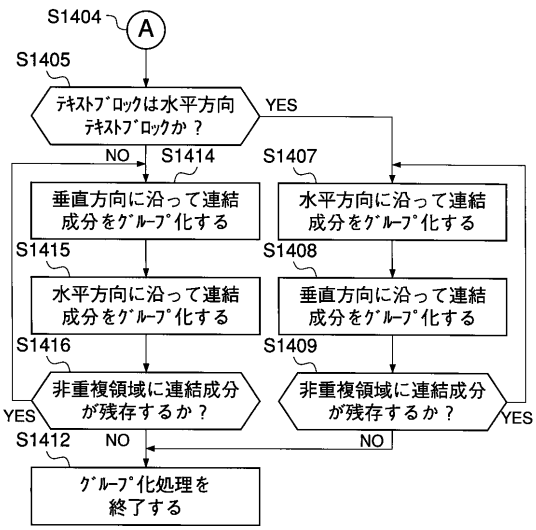
【図１６】



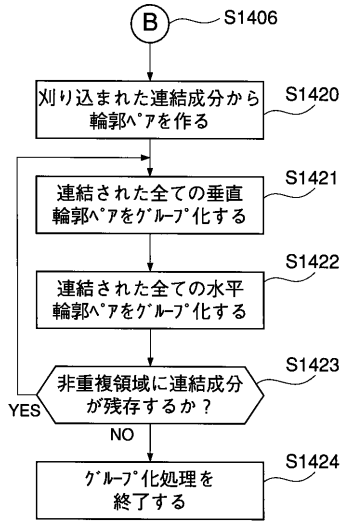
【図１７】



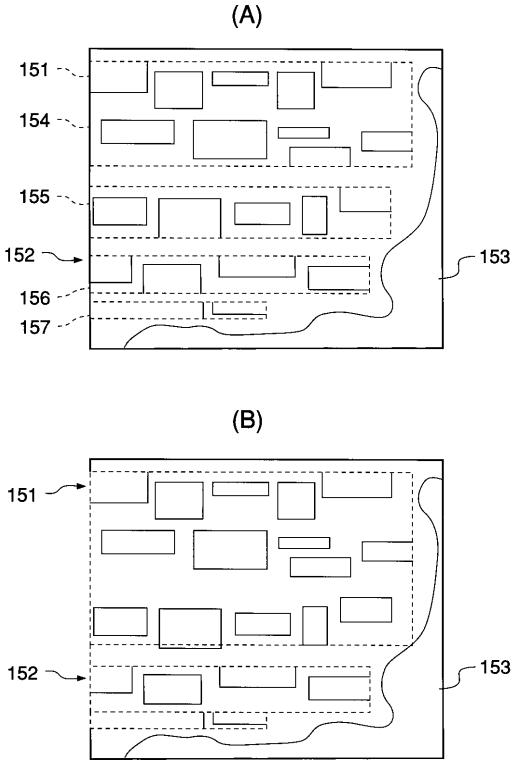
【図１８】



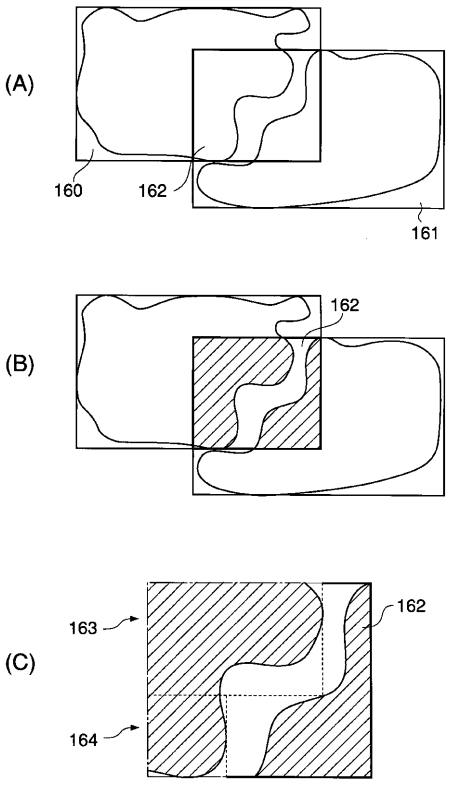
【図１９】



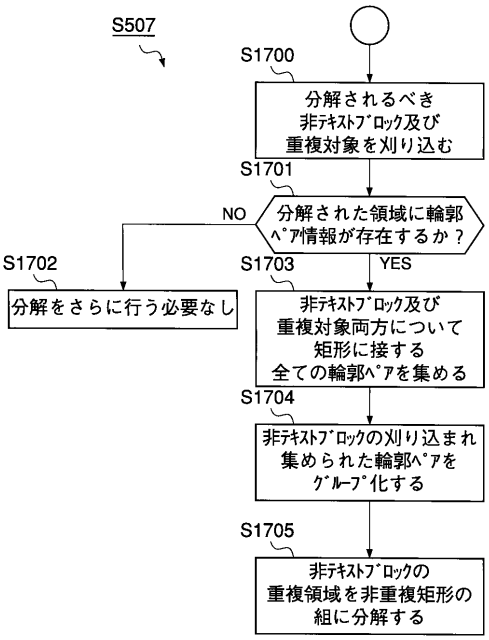
【図２０】



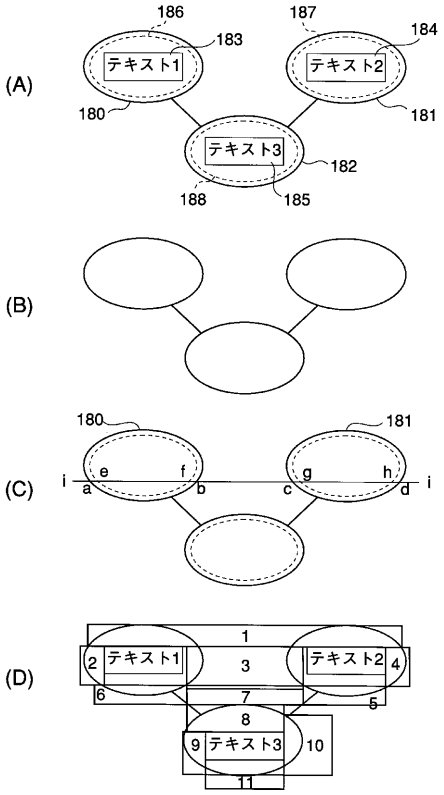
【図２１】



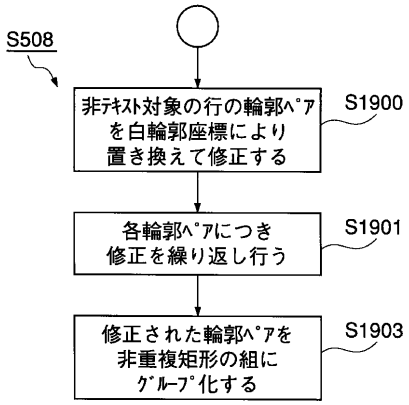
【図２２】



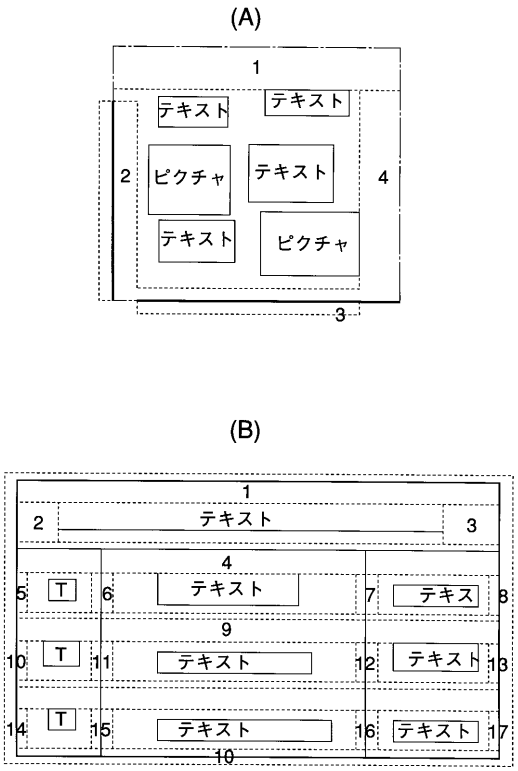
【図２３】



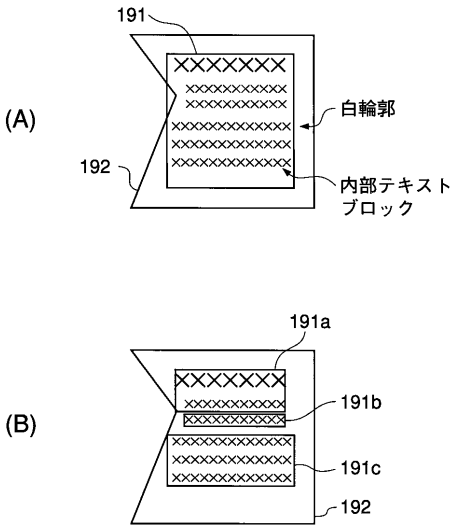
【図２４】



【図２５】



【図２６】



フロントページの続き

(72)発明者　シン　ヤン　ワン
アメリカ合衆国　カリフォルニア　９２６８０　タスティン　マクチャールズ　ドライブ　２２２１
(72)発明者　矢ヶ崎　敏明
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

審査官　飯田　清司

(56)参考文献　特開平０８－１１５３８０（ＪＰ，Ａ）
特開平０７－１０７２８１（ＪＰ，Ａ）
特開平０７－２３４９１８（ＪＰ，Ａ）
特開平０６－０６８３０１（ＪＰ，Ａ）
特開昭６２－２０３２８０（ＪＰ，Ａ）
特開昭６２－２０４３８０（ＪＰ，Ａ）
特開２００１－２５９５３２（ＪＰ，Ａ）
特開２０００－３４８１４０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06T　7/00-7/60
G06K　9/20