(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2006-106972
(P2006-106972A)
(43)公開日　平成18年4月20日(2006.4.20)

(51)Int.Cl.		F I		テーマコード（参考）
G06K 9/00	(2006.01)	G06K 9/00	P	5B064

審査請求 未請求　請求項の数 8　ＯＬ　（全 16 頁）

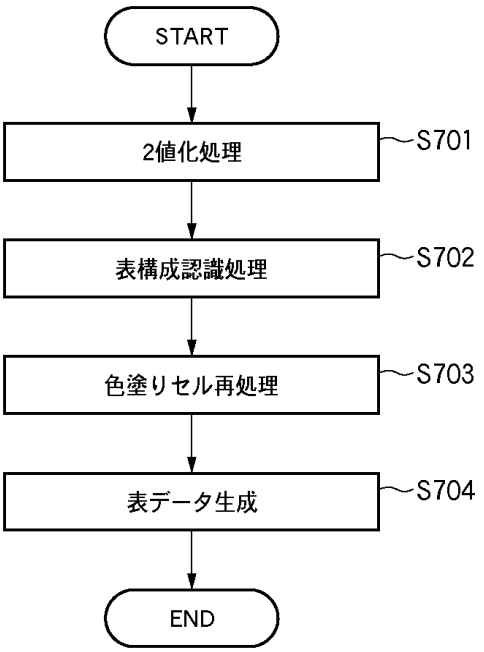
(21)出願番号　特願2004-290386 (P2004-290386)
(22)出願日　平成16年10月1日 (2004.10.1)

(71)出願人　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　100076428
弁理士　大塚　康徳
(74)代理人　100112508
弁理士　高柳　司郎
(74)代理人　100115071
弁理士　大塚　康弘
(74)代理人　100116894
弁理士　木村　秀二
(72)発明者　鵜沢　充
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

Ｆターム(参考)　5B064 AA01　AB09　BA01　CA03

(54)【発明の名称】画像処理方法及び画像処理装置

(57)【要約】
【課題】　本来文字情報が埋め込まれているセルが２値化処理により塗りつぶされることを防止し、表中の全ての文字情報を抽出する。
【解決手段】　原稿上の表領域における表枠の２値データを抽出し、表枠の２値データに基づいて表枠の表構成を認識し、その表枠の中より文字情報を抽出し、表枠を再処理する際に、表枠の表構成の認識結果に応じて文字情報を抽出する領域を選択し、該選択位置における文字情報を再抽出し、その表枠を再処理する。
【選択図】　図７



【特許請求の範囲】
【請求項１】
原稿上の表領域における表枠の２値データを抽出する工程と、
前記表枠の２値データに基づいて表枠の表構成を認識する工程と、
前記表枠の中より文字情報を抽出し、表枠を再処理する工程とを有し、
前記表枠の表構成の認識結果に応じて文字情報を抽出する領域を選択し、該選択位置における文字情報を再抽出し、前記表枠を再処理することを特徴とする画像処理方法。
【請求項２】
前記表構成の認識結果は、矩形図形の集合として表を表現し、前記文字情報を抽出する領域は該矩形図形単位で選択されることを特徴とする請求項１記載の画像処理方法。
【請求項３】
前記再処理は、少なくとも２値化処理か像域分離処理の何れかであることを特徴とする請求項１記載の画像処理方法。
【請求項４】
前記再処理された表枠を表現するベクトルデータを生成する工程を有することを特徴とする請求項１記載の画像処理方法。
【請求項５】
原稿上の表領域における表枠の２値データを抽出する抽出手段と、
前記表枠の２値データに基づいて表枠の表構成を認識する認識手段と、
前記表枠の中より文字情報を抽出し、表枠を再処理する再処理手段とを有し、
前記表枠の表構成の認識結果に応じて文字情報を抽出する領域を選択し、該選択位置における文字情報を再抽出し、前記表枠を再処理することを特徴とする画像処理装置。
【請求項６】
前記再処理手段は、前記表枠でないと想定される領域について、テキスト情報が入っていないと判定した場合は再処理しないことを特徴とする請求項５記載の画像処理装置。
【請求項７】
請求項１記載の画像処理方法をコンピュータに実行させるためのプログラム。
【請求項８】
請求項７記載のプログラムを記録したコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、スキャナなどの入力装置より読み込まれた紙文書を編集可能な電子データへ変換する技術に関し、特に紙文書中の表枠オブジェクトを解析し、罫線に置き換える技術に関する
【背景技術】
【０００２】
近年、情報の電子化が進み、文書を紙ではなく電子化して保存、あるいは送信するシステムが急速に普及している。特に、フルカラーの文書を保存、送信に適した電子データとしては、紙原稿を文字、表、図等のオブジェクトへ像域分離し、各オブジェクトに適した形態でデータ化したベクトルデータが適しており、データ量を削減し、更にオブジェクトを再利用することが可能となる。
【０００３】
ここで、文書中の表オブジェクトについては、像域分離処理により表中のテキスト情報を抽出する一方、表枠を２値画像データとして抽出することが可能である。この２値画像データとして抽出された表枠は、表枠の解析処理により、罫線などで表現されるベクトルデータに変換される（例えば、特許文献１参照）。そして、ベクトルデータに変換された表枠はベクトルデータとしてデータサイズが小さくなるだけでなく、表枠としての再利用性も高い。
【特許文献１】特開平５－１２４８９号公報

【発明の開示】
【発明が解決しようとする課題】
【０００４】
表枠及び表中のテキストを含めて、文書中のオブジェクトを２値化して抽出する際に、画像領域で２値化、もしくはオブジェクト毎に２値化するために、それぞれ２値化の閾値は画像領域、もしくはオブジェクト領域により設定される。
【０００５】
例えば、表領域について２値化を行った場合、図２５に示す（ａ）のように、色の濃いセルがあると、図２５に示す（ｂ）のように、色の濃いセルは塗りつぶされた画像として生成されるという問題が生じる。また、２値化した際に、セル内にノイズが多く発生するという問題が生じる。
【０００６】
尚、このようなセルについては、正常に文字情報を抽出することができない。元原稿がカラー原稿である場合は、各セル色の相関関係によって２値化によりセルの塗りつぶしが発生する場合が多く、表枠の正確な抽出は困難であり、非常に深刻な問題となっている。
【０００７】
本発明は、上述の課題を解決するためになされたもので、本来文字情報が埋め込まれているセルが２値化処理により塗りつぶされることを防止し、表中の全ての文字情報を抽出することを目的とする。
【課題を解決するための手段】
【０００８】
本発明の画像処理方法は、原稿上の表領域における表枠の２値データを抽出する工程と、前記表枠の２値データに基づいて表枠の表構成を認識する工程と、前記表枠の中より文字情報を抽出し、表枠を再処理する工程とを有し、前記表枠の表構成の認識結果に応じて文字情報を抽出する領域を選択し、該選択位置における文字情報を再抽出し、前記表枠を再処理することを特徴とする。
【０００９】
また、本発明の画像処理装置は、原稿上の表領域における表枠の２値データを抽出する抽出手段と、前記表枠の２値データに基づいて表枠の表構成を認識する認識手段と、前記表枠の中より文字情報を抽出し、表枠を再処理する再処理手段とを有し、前記表枠の表構成の認識結果に応じて文字情報を抽出する領域を選択し、該選択位置における文字情報を再抽出し、前記表枠を再処理することを特徴とする。
【発明の効果】
【００１０】
本発明によれば、カラー原稿より表枠中のノイズを除去し、表枠を正確に抽出し、セル内のノイズ、塗りつぶしにより文字情報が抽出されないことを防止できる。また、正確な表枠ベクトル情報を抽出できる。
【発明を実施するための最良の形態】
【００１１】
以下、図面を参照しながら発明を実施するための最良の形態について詳細に説明する。
【実施例１】
【００１２】
図１は、実施例１における文書処理装置の外観を示す図である。図１において、１０１はコンピュータ装置であり、後述するフローチャートを参照して説明する処理を実現するためのプログラムを含む、文書の電子化処理プログラムを実行する。また、コンピュータ装置１０１は、ユーザに状況や画像を表示するためのディスプレイ装置１０２と、ユーザの操作を受け付けるキーボードやマウス等のポインティングデバイスを含んで構成される入力装置１０３とを付随する。このディスプレイ装置１０２としては、ＣＲＴやＬＣＤ等が用いられる。１０４はスキャナ装置であり、文書画像を光学的に読み取って電子化し、得られた画像データをコンピュータ装置１０１に送る。尚、スキャナ装置１０４としては

、カラースキャナを用いるものとする。

【００１３】

図２は、実施例１における文書処理装置の構成の一例を示すブロック図である。図２において、２０１はＣＰＵであり、後述するＲＯＭ又はＲＡＭに格納された制御プログラムを実行することにより、後述する電子化処理を含む各種機能を実現する。２０２はＲＯＭであり、ＣＰＵ２０１によって実行される各種制御プログラムや制御データが格納されている。２０３はＲＡＭであり、ＣＰＵ２０１によって実行される各種制御プログラムを格納したり、ＣＰＵ２０１が各種処理を実行するのに必要な作業領域が定義されている。

【００１４】

２０４は外部記憶装置であり、詳細は後述する実施例１における処理をＣＰＵ１０１によって実現するための制御プログラムや、スキャナ装置１０４で読み取って得られた文書画像データ等を格納する。そして、２０５はコンピュータバスであり、上述した各構成を接続するものである。

【００１５】

図３は、文書処理装置における文書の電子化処理の概要を示す図である。ここで、電子化処理の流れは、まず入力部３０１において、電子化の対象であるカラー文書をスキャナ装置１０４によって読み込み、画像データとして外部記憶装置２０４に格納する。次に、２値化処理３０２において、後段の像域分離処理、アウトライン生成処理のために、外部記憶装置２０４に格納された文書の画像データに対して２値化処理を施す。そして、像域分離処理３０３では、２値化処理３０２で得られた２値画像から、文字、図、表、枠、線などの要素を抽出し、各領域に分割する。

【００１６】

次に、ベクトル化処理３０４において、領域分割された画像データに対して、文字部は文字認識部３０５で文字認識を行い、アウトライン作成部３０６でアウトラインベクトルデータへ変換する。また、表、枠の要素については表処理部３０８でアウトライン化し、アウトラインを罫線化する。尚、アウトライン作成部３０６で変換された画像データは、各オブジェクトの輪郭線が滑らかな曲線により表現される高画質で、解像度に依存しない、かつ編集容易なベクトルデータへ変換される。

【００１７】

一方、その他の図、写真画、背景については、例えば背景については、圧縮部３０９でＪＰＥＧ圧縮など各々に適した形態で保持、圧縮する。

【００１８】

次に、電子文書作成処理３１０は、分割された要素毎の属性に基づいて文字認識データや表構造データを用い、それぞれ変換された画像データに基づき電子化文書を作成する。そして、出力部３１１は生成された電子化文書を外部記憶装置２０４に格納する。

【００１９】

尚、出力部３１１の出力形態は外部記憶装置２０４への格納に限られるものではなく、ディスプレイ装置１０２へ表示したり、不図示のネットワークインターフェースを介してネットワーク上の他の装置へ出力したり、不図示のプリンタへ送出したりすることも可能である。

【００２０】

ここで、図１及び図２に示す文書処理装置において実行される文書の電子化処理（図３参照）における各処理の詳細について、以下順に説明する。

【００２１】

［２値化処理］

２値化処理３０２では、入力された文書画像データより輝度情報を抽出し、その輝度値のヒストグラムを作成する。ヒストグラム上より複数の閾値を設定し、各々の閾値で２値化された２値画像上の黒画素の連結等を解析することで最適な閾値を導出し、その閾値による２値画像を得る。

【００２２】

［像域分離処理］

像域分離処理３０３とは、図４に示す左側の読み取られた１ページのイメージデータをオブジェクト毎の塊（ブロック）として認識し、各々の塊を文字／図画／写真／線／表等の属性に判定し、図４に示す右側のように、異なる属性（TEXT／PICTURE／PHOTE／LINE／TABLE）を持つ領域に分割する処理である。

【００２３】

像域分離処理３０３では、２値化処理３０２で得られた２値画像より、黒画素の輪郭線追跡を行って黒画素輪郭で囲まれる画素の塊を抽出する。また、面積の大きい黒画素の塊については、内部にある白画素に対しても輪郭線追跡を行い、白画素の塊を抽出し、更に一定面積以上の白画素の塊の内部からは再帰的に黒画素の塊を抽出する。

【００２４】

このようにして得られた黒画素の塊を、大きさ及び形状で分類し、異なる属性を持つ領域へ分類していく。例えば、縦横比が１に近く、大きさが一定の範囲のものを文字相当の画素塊とし、更に近接する文字が整列良くグループ化可能な部分を文字領域、扁平な画素塊を線領域、一定の大きさ以上で、かつ四角系の白画素塊を整列よく内包する黒画素塊の占める範囲を表領域、不定形の画素塊が散在している領域を写真領域、それ以外の任意形状の画素塊を図画領域、などとする。

【００２５】

図５は、像域分離処理３０３で分離された各ブロックに対するブロック情報と入力ファイル情報を示す図である。図５に示すように、ブロック情報は、各ブロックの属性、座標（Ｘ，Ｙ）、幅（Ｗ）、高さ（Ｈ）、ＯＣＲ情報を含み、属性１は文字、属性２は図画、属性３は表、属性４は線、属性５は写真である。そして、入力ファイル情報は、ブロック総数Ｎ（図５に示す例では、ブロック１～ブロック６までの６である）を有する。

【００２６】

尚、各ブロックに対して、より鮮明な２値画像を得ようとした場合は、ここでブロック毎に上述した２値化処理を行っても良い。

【００２７】

［文字認識部］

文字認識部３０５では、文字単位で切り出された画像に対して、パターンマッチングの一手法を用いて認識を行い、対応する文字コードを得る。この認識処理は、文字画像から得られる特徴を数十次元の数値列に変換した観測特徴ベクトルと、予め字種毎に求められている辞書特徴ベクトルとを比較し、最も距離の近い字種を認識結果とする処理である。この特徴ベクトルの抽出には種々の公知手法があり、例えば文字をメッシュ状に分割し、各メッシュ内の文字線を方向別に線素としてカウントしたメッシュ数次元ベクトルを特徴とする方法がある。

【００２８】

像域分離処理３０３で抽出された文字領域に対して文字認識を行う場合、まず該当領域に対して横書き、縦書きの判定を行い、それぞれ対応する方向に行を切り出し、その後、文字を切り出して文字画像を得る。この横書き、縦書きの判定は、該当領域内で画素値に対する水平／垂直の射影を取り、水平射影の分散が大きい場合には横書き領域と判定し、垂直射影の分散が大きい場合には縦書き領域と判定すれば良い。また、文字列及び文字への分解は、横書きならば水平方向の射影を利用して行を切り出し、更に切り出された行に対する垂直方向の射影から、文字を切り出すことで行う。縦書きの文字領域に対しては、水平と垂直を逆にすれば良い。尚、この時、文字のサイズが検出できる。

【００２９】

［アウトライン生成部］

アウトライン作成部３０６では、像域分離処理３０３で図画或いは線、表領域とされた領域を対象に、抽出された画素塊の輪郭をベクトルデータに変換する。具体的には、輪郭をなす画素の点列を角と看做される点で区切って、各区間を部分的な直線或いは曲線で近似する。角とは曲率が極大となる点であり、曲率が極大となる点は、図６に示すように、

任意点Ｐｉに対して左右ｋ個の離れた点Ｐｉ－ｋ，Ｐｉ＋ｋの間に弦を引いたとき、この弦とＰｉの距離が極大となる点として求められる。更に、点Ｐｉ－ｋ，Ｐｉ＋ｋ間の弦の長さ／弧の長さをＲとし、Ｒの値が閾値以下である点を角とみなすことができる。角によって分割された後の各区間は、直線は点列に対する最小二乗法など、曲線は３次スプライン関数などを用いてベクトル化することができる。

【００３０】

また、対象が内輪郭を持つ場合、像域分離処理で抽出した白画素輪郭の点列を用いて、同様に部分的直線或いは曲線で近似する。

【００３１】

［表処理部］

表処理部３０８では、表中のセル及びその構成を認識し、表枠を罫線により表現する等、セル毎に編集可能な電子データへ変換する。尚、表部は、像域分離処理３０３により、表枠として表枠中の文字部と分離して抽出されている。

【００３２】

図７は、表処理部３０８における表処理を示すフローチャートである。まず、像域分離処理３０３で分離された表部を２値化し、表枠の２値データを得る（ステップＳ７０１）。ここでは、像域分離処理３０３より入力された表枠の２値データをそのまま出力しても良い。また、像域分離処理３０３で得られた表領域に対して、再度２値化処理、表枠抽出処理を行い、正確に表枠を抽出しても良い。

【００３３】

次に、ステップＳ７０１で得られた表枠について表構成を認識する（ステップＳ７０２）。図８は、表枠を認識する処理を示すフローチャートである。まず、表枠について上述したアウトライン処理によりアウトライン化し、滑らかな直線及び曲線で表枠を表現する（ステップＳ８０１）。次に、表枠のアウトラインよりセルを表現しているセルアウトラインを抽出する（ステップＳ８０２）。ここで、元々アウトラインは、外輪郭と内輪郭に分類されているが、まず外輪郭のうち、表全体の外枠を構成しているアウトラインを抽出する。尚、表の内部に表が存在するような場合もあるので、外枠は複数抽出される場合もある。

【００３４】

次に、外枠の内側に存在するセルを構成しているアウトラインを抽出する。尚、ここでの処理はアウトラインよりそのアウトラインがセルを構成するサイズであるか否かを判定し、更にアウトラインを図形認識処理によりアウトラインが矩形図形、もしくは三角図形を構成しているか否かを判定する。尚、矩形図形、三角図形、もしくは矩形図形の集合と判定されたアウトラインをセルアウトラインとする。図９に示す（ａ），（ｂ），（ｃ）は、矩形図形、三角図形、矩形図形の集合と判定されたアウトラインの一例である。

【００３５】

次に、ステップＳ７０２で外枠、もしくはセルを構成していると判定されたセルアウトラインについてセル図形へ変換する（ステップＳ８０３）。具体的には、まず、ステップＳ８０２で外枠を構成するアウトラインと内部セルを構成するアウトラインが抽出されているが、外枠を構成するアウトラインの角度が全て90°で表現される図形であると判定された場合、90°角の間を直線で表現した図形へ変換する。次に、外枠の内側のセルを構成すると判定されたセルについて、セル図形へ変換する。ここで、セル図形とは矩形図形である。

【００３６】

例えば、図９に示すセルアウトラインをセル図形へ変換した例を図１０に示す。図９に示すセルアウトライン（ａ），（ｂ），（ｃ）はそれぞれ図１０に示す（ａ），（ｂ），（ｃ）のように変換される。ここで、図９に示す（ａ）のセルアウトラインは矩形図形の当てはめ処理によりそのまま図１０に示す（ａ）となる。図９に示す（ｂ）のような三角図形のセルアウトラインも同様に、矩形図形の当てはめ処理を行う。この三角図形に対する矩形図形の当てはめ処理では、三角を構成するセルアウトラインを囲むようにセル矩形

を当てはめる。

【００３７】

尚、当てはめられた矩形図形は最終的にその位置関係よりセル図形同士統合され一つのセル図形として抽出する。例えば、図１０に示す（ｂ）のセル図形は、図１１に示すセル図形に統合され、一つのセル図形として表現される。統合されないセル図形も当然あり、それらの図形についてはそのまま三角アウトラインに当てはめられたセル図形をセル図形とする。

【００３８】

また、図９に示す（ｃ）のような矩形図形の集合として抽出されるセルアウトラインは、図１０に示す（ｃ）のように各矩形図形へ分離する。この矩形図形への分離処理では、アウトラインの中の直角をなすであろう角を見つけ出し、その角点の構成から矩形図形へ分解する。尚、更に抽出されたセル全てを用いて表の水平方向及び垂直方向を求め、セル図形全てを求められた水平方向及び垂直方向の成分からなるセル図形へ変換する。このように外枠図形とセル図形は抽出されるが、各セル図形はその構成されるセルアウトラインにより属性情報がつけられる。図１２は、セル図形と、それを構成するセルアウトライン及び各属性情報の一例を示す図である。

【００３９】

次に、ステップＳ８０３で変換したセル図形をマッピングするためのマッピング領域を作成する（ステップＳ８０４）。ここで、マッピング領域とはセル図形をマッピングするための領域であり、外枠内部の領域がそのままマッピング領域となる。また、外枠の交点を抽出しておく。交点とは、表の罫線と罫線が交差する点のことであり、外枠においては、外枠の角点がそのまま交点となる。図１３にマッピング領域と交点の一例を示す。

【００４０】

次に、セル図形マッピング領域にセル図形をマッピングし（ステップＳ８０５）、セル図形の構成を認識する。具体的には、ステップＳ８０４で作成されたマッピング領域内にセル図形をマッピングしていき、マッピングされるセル図形より表の罫線と罫線が交差する交点を抽出していくことで、表構成を認識する。即ち、交点の隣接関係を調べていくことで、表構成を認識する。

【００４１】

ここで、ステップＳ８０５におけるセル図形をマッピングするセル図形マッピング処理の詳細について説明する。

【００４２】

図１４は、セル図形マッピング処理を示すフローチャートである。まず、上述した外枠の交点より注目点を抽出する（ステップＳ１４０１）。ここで、注目点とは、右側と下側に隣接して繋がっている交点を持つ交点で、それら３点を含む矩形領域がマッピング領域であり、かつ該矩形領域に対し、まだ何もマッピングされていない交点である。図１５にマッピング領域と交点とマッピングセルと注目点の関係を示した例を示す。

【００４３】

次に、抽出した注目点に対し、該注目点を左上の交点とするセル図形が存在するか否かを判定する（ステップＳ１４０２）。具体的には、未だマッピングされていない全てのセル図形の左上の角点と注目点との距離を調べ、セル図形の左上の角点と注目点との距離が一定値以内であり、最も注目点に近いセル図形を注目点と左上の角点が一致するセル図形と判定する。ここで、注目点と左上の角点が一致するセル図形が存在すれば、該セル図形をマッピング領域上にマッピングする（ステップＳ１４０３）。

【００４４】

また、セル図形全ての左上の角点と注目点との距離が一定値以内にあるセル図形が存在しなければ、色塗りセルを作成し、マッピングする（ステップＳ１４０４）。色塗りセルは、矩形図形である。まず、注目点とその隣接する右側と下側の交点より少し広げた矩形領域（以後矩形領域Ａと呼ぶ）内に、まだマッピングされていないセル図形の角点がないか判定し、角点が存在すれば、その角点を通る水平方向及び垂直方向の直線によって領域

を区切る。この区切り作業を矩形領域Ａ内に存在する角点全てに対して行い、水平線及び垂直線によって区切られた領域の最も左上にある区切られた矩形図形を色塗りセルとし、マッピングする。

【００４５】

図１６は、矩形領域Ａ内に色塗りセルを作成する例を示す図である。図１６に示す例では、矩形領域Ａ内に２つのセル図形の角点が存在するので、それらの角点を通る水平方向及び垂直方向の直線によって矩形領域Ａを区切り、区切られた領域の最も左上にある区切られた矩形図形を塗りつぶしセルとしてマッピングする。

【００４６】

次に、ステップＳ１４０３、Ｓ１４０４でマッピングされたセル図形よりセル図形上の交点を作成する（ステップＳ１４０５）。交点はこのマッピング図形の角点がそのまま交点となるが、もしマッピング図形の角点が、既に存在する交点との距離がある閾値以内であれば、その角点により作成される交点は既に存在すると判断できるため、その角点より新たな交点は作成しない。ここで、マッピング図形の左上の交点は注目点と一致と判断されているため、左上の角点より新たな交点は作成されない。また、マッピング図形の右上の角点より作成される交点は注目点より水平線上にあるとして作成し、左下の角点は注目点より垂直線上にあるとして作成する。

【００４７】

次に、現在抽出されている交点の中で注目点が存在するか否かを判定する（ステップＳ１４０６）。注目点とは、上述したように、右側と下側に隣接し繋がっている交点を持つ交点で、そられ３点を含む矩形領域がマッピング領域であり、かつ該矩形領域に対しまだ何もマッピングされていない交点である。交点が囲む領域内にセル図形及び塗りつぶしセルがマッピングされていない領域が存在すれば注目点は存在する。尚、注目点が存在しない場合には表構成認識処理を終了とする。また、まだ注目点が存在する場合は、ステップＳ１４０１に戻り、再度注目点を抽出して一連の処理を繰り返す。

【００４８】

以上の繰り返し処理により、交点の隣接関係が作成され、罫線を表現することが可能となる。図１７に作成された交点の隣接関係と、マッピングされたセル図形及び色塗りセル、またセル図形の場合はその属性情報を記述した例を示す。尚、図１７に示す属性情報により得られる表罫線は図１８に示すようになる。

【００４９】

ここで、図１４に示したセル図形マッピング処理が終了すると、図８に示すステップＳ８０６へ進み、罫線の太さ及び位置関係を調節し、交点の正確な位置を抽出する。罫線の太さは、ステップＳ８０５でマッピングされたセル図形のうち、隣接するセル図形の距離から求められる。また、罫線の位置は隣接するセル図形の中間となるように調節する。

【００５０】

尚、交点の位置はステップＳ８０６で得られる罫線同士の交わる交点として正確な位置を求める。

【００５１】

［色塗りセル再処理］

図７に示すステップＳ７０３では、表構成認識処理（ステップＳ７０２）で抽出された色塗りセルについて再処理を行う。

【００５２】

図１９は、色塗りセル再処理を示すフローチャートである。まず、セルの判別を行う（ステップＳ１９０１）。セル判別では、ステップＳ７０２で作成した色塗りセルの区切りについて原稿上でもセルの区切りが存在するか否かを判別し、場合によっては色塗りセルを統合する。具体的には、原画像について、隣り合う色塗りセルの区切り線に対応する近隣画素のエッジ成分を調べ、原画像において区切り線上の罫線があるか否かを判別する。もし、原画像上で罫線が無いと判別された場合は、２つのセルを統合し一つのセルとする。以上の統合処理を隣接する色塗りセル全てに対して行い、セルを抽出する。

【００５３】
次に、ステップＳ１９０１で抽出されたセルについて、一つのセルを一枚の画像とみなし、セル毎に２値化処理を行う（ステップＳ１９０２）。セル毎に２値化処理を行うことで、塗りつぶされていた色塗りセルより、セル内のテキスト、模様等の２値オブジェクトを抽出することが可能である。そして、セル毎に像域分離処理を行う（ステップＳ１９０３）。
【００５４】
尚、像域分離処理は、特にテキストデータを抽出するものである。抽出されたテキストデータが抽出された場合はＯＣＲをかけてテキストコードを抽出し、更にアウトライン化してフォント化し、文字部のデータへ追加する。
【００５５】
［表データ生成処理］
図７に示すステップＳ７０４では、ステップＳ７０２、Ｓ７０３で作成されたベクトル情報を使用し、表データを作成する。例えば、図１８に示す表は図２４に示すようになる。図２４において、２４０１～２４０３の部位からは、テキスト情報が抽出されている。また、もしテキスト情報がないセルについては、そのまま色塗りセルのままで良いとする。
【００５６】
以上の処理により、表枠は塗りつぶされるようなセルが作成されることを回避し、表中よりテキスト情報を正確に抽出し、表データが作成される。
【００５７】
尚、上述の処理では、色塗りセルについて、セル判別を行い、得られる各セルについて２値化、像域分離処理を行ったが、より単純に抽出された複数の色塗りセル全ての領域を一枚の原稿と見立て、それに対して２値化、像域分離処理を行ってもよく、色塗りセルの領域より文字、線等の情報を得ることが可能である。
【００５８】
［アプリデータへの変換処理］
以上の通り、１頁分のイメージデータを像域分離処理３０３し、ベクトル化処理３０４した結果は図２０に示すような中間データ形式のファイルとして変換される。このようなデータ形式は、ドキュメント・アナリシス・アウトプット・フォーマット（ＤＡＯＦ）と呼ばれる。
【００５９】
図２０は、ＤＡＯＦのデータ構造を示す図である。図２０において、２００１はHeaderであり、処理対象の文書画像データに関する情報が保持される。２００２はレイアウト記述データ部であり、文書画像データ中の文字（TEXT）、タイトル（TITLE）、キャプション（CAPTION）、線画（LINEART）、自然画（PICTURE）、枠（FRAME）、表（TABLE）等の属性毎に認識された各ブロックの属性情報とその矩形アドレス情報を保持する。２００３は文字認識記述データ部であり、TEXT、TITLE、CAPTION等のTEXTブロックを文字認識して得られる文字認識結果を保持する。２００４は表記述データ部であり、TABLEブロックの構造の詳細を格納する。２００５は画像記述データ部であり、PICTUREやLINEART等のブロックのイメージデータを文書画像データから切り出して保持する。
【００６０】
このようなＤＡＯＦは中間データとしてのみならず、それ自体がファイル化されて保存される場合もあるが、このファイルの状態では、所謂一般の文書作成アプリケーションで個々のオブジェクトを再利用することはできない。
【００６１】
そこで、このＤＡＯＦからアプリケーションデータに変換する電子文書作成処理３０９について説明する。
【００６２】
図２１は、電子文書作成処理の全体の概略を示すフローチャートである。まずステップ

Ｓ２１０１において、ＤＡＯＦデータの入力を行う。次に、ステップＳ２１０２において、アプリデータの元となる文書構造ツリー生成を行う。そして、ステップＳ２１０３で、文書構造ツリーに基づいてＤＡＯＦ内の実データを流し込み、実際のアプリデータを生成する。

【００６３】

図２２は、文書構造ツリー生成処理の詳細を示すフローチャートである。また、図２３は文書構造ツリーを説明するための図である。尚、全体制御の基本ルールとして、処理の流れはミクロブロック（単一ブロック）からマクロブロック（ブロックの集合体）へ移行する。尚、以下の説明で、「ブロック」はミクロブロック及びマクロブロック全体を指すものとする。

【００６４】

まず、ステップＳ２２０１では、ブロック単位に縦方向の関連性に基づいて再グループ化する。スタート直後はミクロブロック単位での判定となる。ここで、関連性とは、距離が近い、ブロック幅（横方向の場合は高さ）がほぼ同一であることなどで定義することができる。また、距離、幅、高さなどの情報はＤＡＯＦを参照し、抽出する。

【００６５】

図２３は、ページの構成とその文書構造のツリーを示す図である。図２３に示す（ａ）は実際のページ構成、図２３に示す（ｂ）はその文書構造ツリーである。

【００６６】

ステップＳ２２０１での結果、図２３に示すＴ３、Ｔ４、Ｔ５が１つのグループＶ１として生成され、Ｔ６、Ｔ７が１つのグループＶ２として生成され、図２３に示す（ｂ）のように、グループＶ１とグループＶ２が同じ階層のグループとして生成される。そして、ステップＳ２２０２において、縦方向のセパレータの有無をチェックする。セパレータは、例えば物理的にはＤＡＯＦ中でライン属性を持つオブジェクトである。また、論理的な意味としては、アプリ中で明示的にブロックを分割する要素である。ここでセパレータを検出した場合は、同じ階層で再分割する。

【００６７】

次に、ステップＳ２２０３において、分割がこれ以上存在し得ないか否かをグループ長を利用して判定する。ここで、縦方向のグループ長がページ高さとなっている場合、文書構造ツリー生成を終了する。また、図２３に示す例の場合、セパレータもなく、グループ高さはページ高さではないのでステップＳ２２０４へ進み、ブロック単位で横方向の関連性に基づいて再グループ化する。ここもスタート直後の第一回目はミクロブロック単位で判定を行うことになる。尚、関連性、及びその判定情報の定義は、縦方向の場合と同じである。

【００６８】

図２３に示す例の場合、Ｔ１、Ｔ２でＨ１が、Ｖ１、Ｖ２でＨ２がＶ１、Ｖ２の１つ上の同じ階層のグループとして生成される。そして、ステップＳ２２０５において、横方向セパレータの有無をチェックする。図２３に示す例では、Ｓ１があるので、これをツリーに登録し、Ｈ１、Ｓ１、Ｈ２という階層を生成する。

【００６９】

次に、ステップＳ２２０６において、分割がこれ以上存在し得ないか否かをグループ長を利用して判定する。ここで、横方向のグループ長がページ幅となっている場合、文書構造ツリー生成を終了する。また、そうでない場合はステップＳ２２０１に戻り、再びもう一段上の階層で、縦方向の関連性チェックから繰り返す。図２３に示す例の場合、分割幅がページ幅になっているので、ここで終了し、最後にページ全体を表す最上位階層のＶ０が文書構造ツリーに付加される。

【００７０】

文書構造ツリーが完成した後、その情報に基づいてアプリデータを生成する（ステップＳ２１０３）。図２３に示す例の場合、具体的には、以下のようになる。

【００７１】

即ち、Ｈ１は横方向に２つのブロックＴ１及びＴ２があるので、２カラムとし、Ｔ１の内部情報（ＤＡＯＦを参照、文字認識結果の文章、画像など）を出力後、カラムを変え、Ｔ２の内部情報出力、その後Ｓ１を出力する。次に、Ｈ２は横方向に２つのブロックＶ１及びＶ２があるので、２カラムとして出力、Ｖ１はＴ３、Ｔ４、Ｔ５の順にその内部情報を出力、その後カラムを変え、Ｖ２のＴ６、Ｔ７の内部情報を出力する。

【００７２】

以上の処理によりアプリデータへの変換処理を行うことができる。

【００７３】

尚、本発明は複数の機器（例えば、ホストコンピュータ，インターフェース機器，リーダ，プリンタなど）から構成されるシステムに適用しても、１つの機器からなる装置（例えば、複写機，ファクシミリ装置など）に適用しても良い。具体的には、複合機や、複写機や、ファクシミリ装置で、高品位に変倍するために、スキャンした画像データを入力し（公衆回線やネットワークから画像データを入力しても良い）、画像データから輪郭ベクトルを抽出し、抽出した輪郭ベクトルを変倍し、変倍された輪郭ベクトルから画像データを生成し、生成した画像データをプリントする際の輪郭ベクトル抽出時に適用できる。

【００７４】

また、本発明の目的は前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記録媒体を、システム或いは装置に供給し、そのシステム或いは装置のコンピュータ（ＣＰＵ若しくはＭＰＵ）が記録媒体に格納されたプログラムコードを読出し実行することによっても、達成されることは言うまでもない。

【００７５】

この場合、記録媒体から読出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記憶した記録媒体は本発明を構成することになる。

【００７６】

このプログラムコードを供給するための記録媒体としては、例えばフロッピー（登録商標）ディスク，ハードディスク，光ディスク，光磁気ディスク，ＣＤ－ＲＯＭ，ＣＤ－Ｒ，磁気テープ，不揮発性のメモリカード，ＲＯＭなどを用いることができる。

【００７７】

また、コンピュータが読出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているＯＳ（オペレーティングシステム）などが実際の処理の一部又は全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００７８】

更に、記録媒体から読出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵなどが実際の処理の一部又は全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【図面の簡単な説明】

【００７９】

【図１】実施例１における文書処理装置の外観を示す図である。

【図２】実施例１における文書処理装置の構成の一例を示すブロック図である。

【図３】文書処理装置における文書の電子化処理の概要を示す図である。

【図４】実施例１における像域分離処理を説明するための図である。

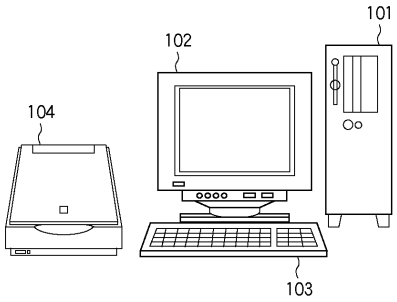
【図５】像域分離処理３０３で分離された各ブロックに対するブロック情報と入力ファイル情報を示す図である。

【図６】アウトラインベクトルデータへの変換を説明するための図である。

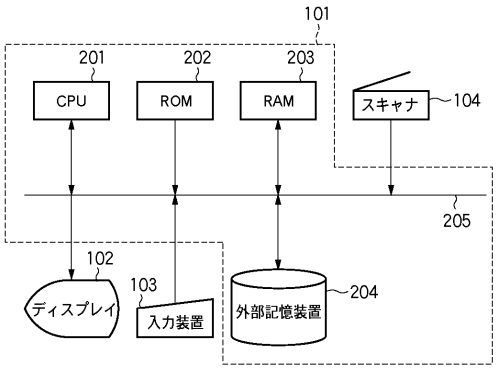
【図７】表処理部３０８における表処理を示すフローチャートである。

【図８】表枠を認識する処理を示すフローチャートである。
【図９】矩形図形、三角図形、矩形図形の集合と判定されるアウトラインの一例を示す図である。
【図１０】図９に示すセルアウトラインをセル図形へ変換した例を示す図である。
【図１１】隣接するセル図形の統合例を示す図である。
【図１２】セル図形を構成するセルアウトラインと各属性情報についての一例を示す図である。
【図１３】マッピング領域と交点の一例を示す図である。
【図１４】セル図形マッピング処理を示すフローチャートである。
【図１５】マッピング領域と交点とマッピングセルと注目点の関係を示す図である。
【図１６】矩形領域Ａ内に色塗りセルを作成する例を示す図である。
【図１７】作成された交点の隣接関係と、マッピングされたセル図形及び色塗りセル、またセル図形の場合はその属性情報を記述した例を示す図である。
【図１８】図１７に示す交点の隣接関係から得られた処理結果を示す図である。
【図１９】色塗りセル再処理を示すフローチャートである。
【図２０】ドキュメント・アナリシス・アウトプット・フォーマット（ＤＡＯＦ）のデータ構造を示す図である。
【図２１】電子文書作成処理の全体の概略を示すフローチャートである。
【図２２】文書構造ツリー生成処理の詳細を示すフローチャートである。
【図２３】ページの構成とその文書構造のツリーを示す図である。
【図２４】本実施例において作成される表枠を示す図である。
【図２５】従来の表処理における問題を説明するための図である。

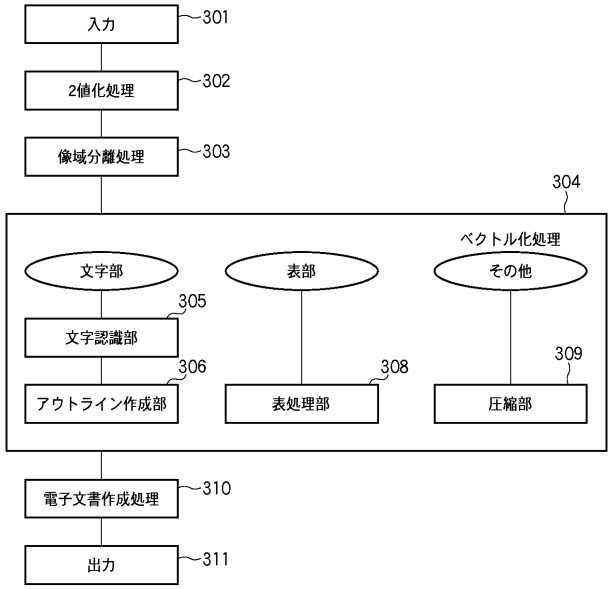
【図１】



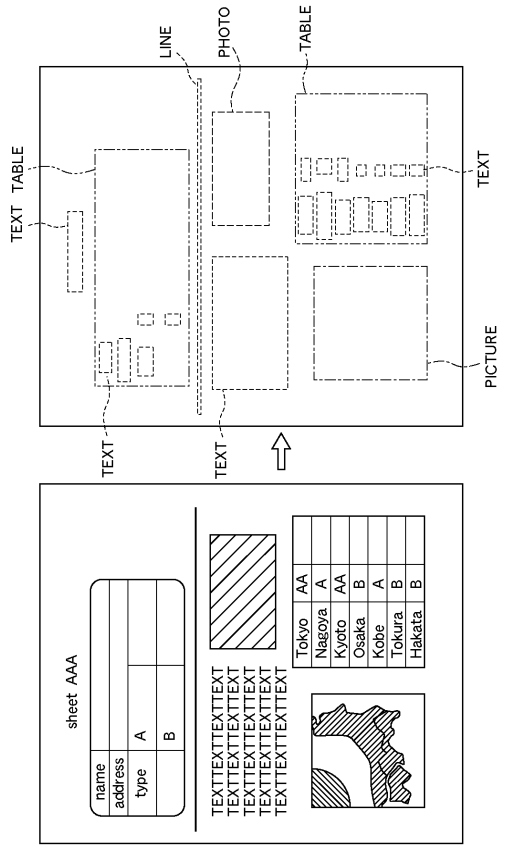
【図２】



【図３】



【図４】



【図５】

ブロック情報

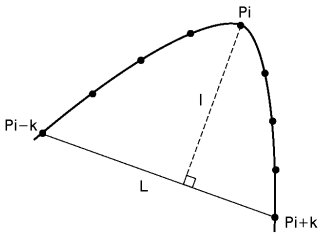
	属性	座標X	座標Y	幅W	高さH	OCR情報
ブロック1	1	X1	Y1	W1	H1	有
ブロック2	3	X2	Y2	W2	H2	有
ブロック3	2	X3	Y3	W3	H3	無
ブロック4	1	X4	Y4	W4	H4	有
ブロック5	3	X5	Y5	W5	H5	有
ブロック6	5	X6	Y6	W6	H6	無

＊ 属性 1 : text 2 : picture 3 : table 4 : line 5 : photo

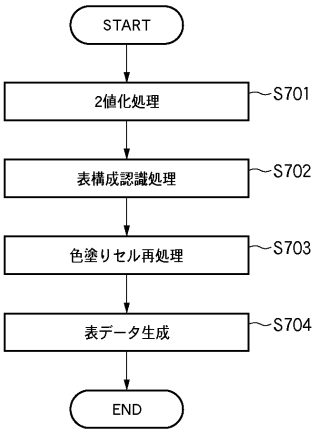
入力ファイル情報

ブロック総数	N (=6)

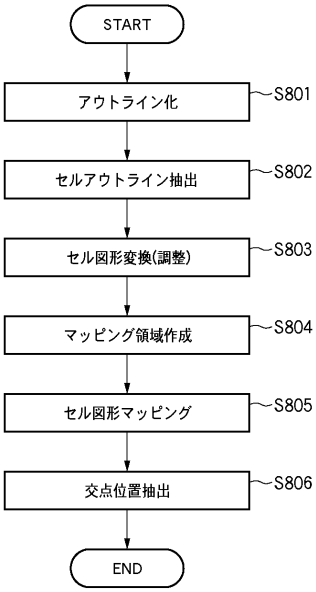
【図６】



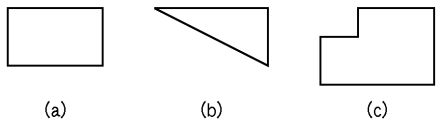
【図７】



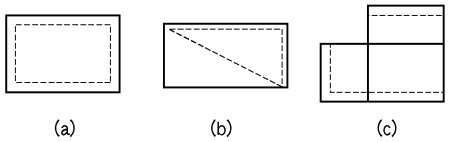
【図８】



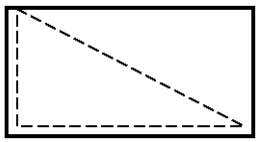
【図９】



【図１０】



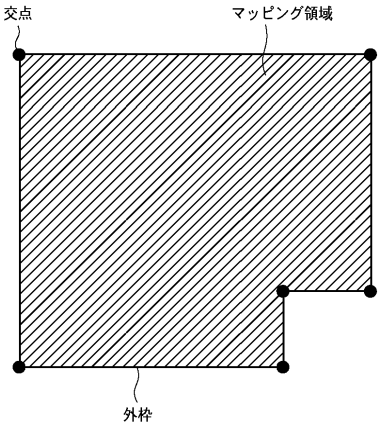
【図１１】



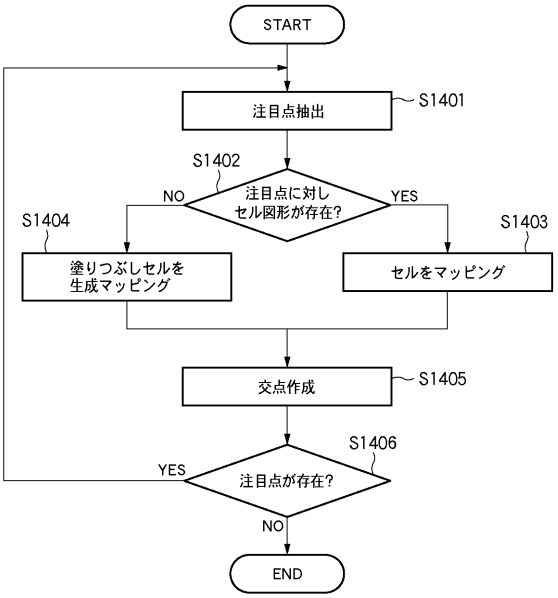
【図１２】

属性名	セルアウトライン
属性1	
属性2	
属性3	
属性4	
属性5	
属性6	
属性7	
属性8	
属性9	

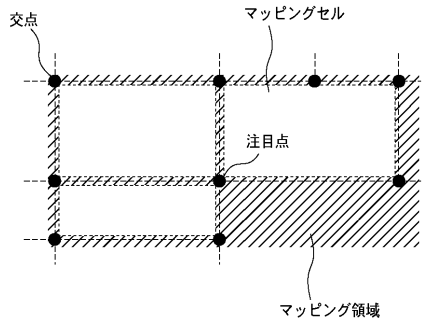
【図１３】



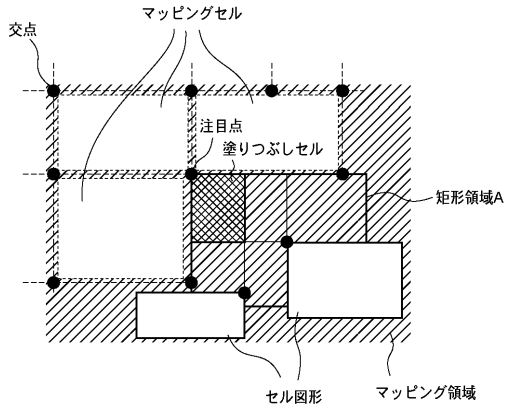
【図１４】



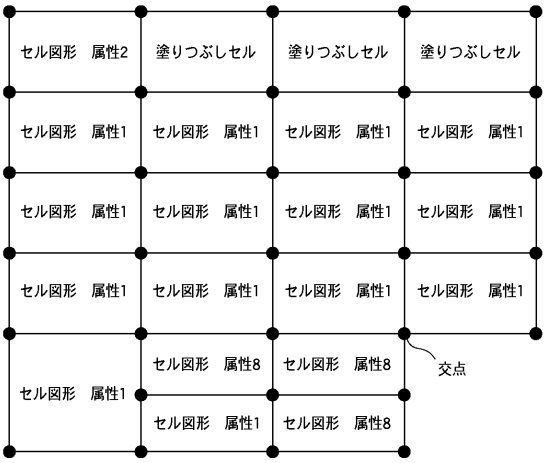
【図１５】



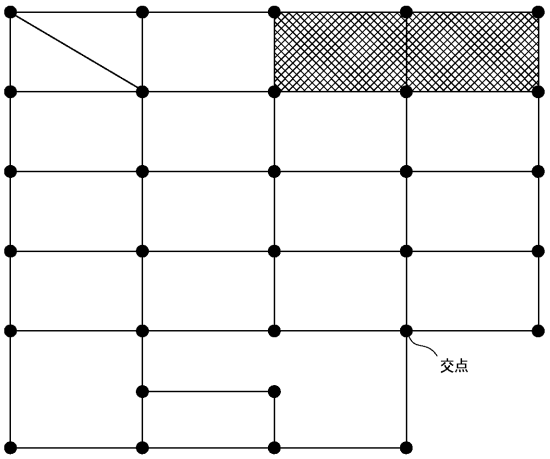
【図１６】



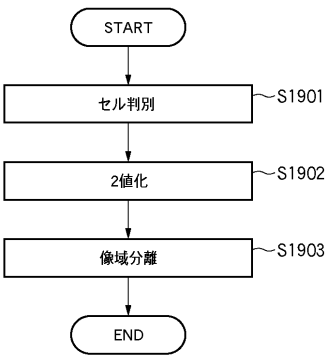
【図１７】



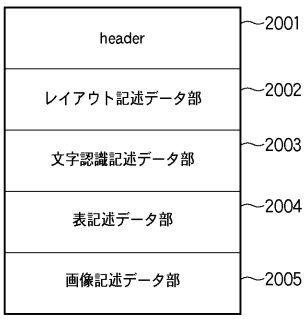
【図１８】



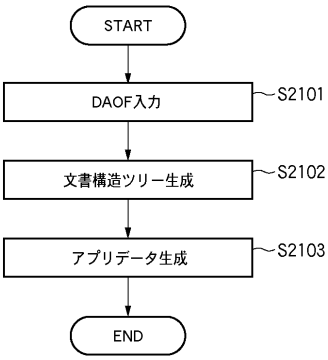
【図１９】



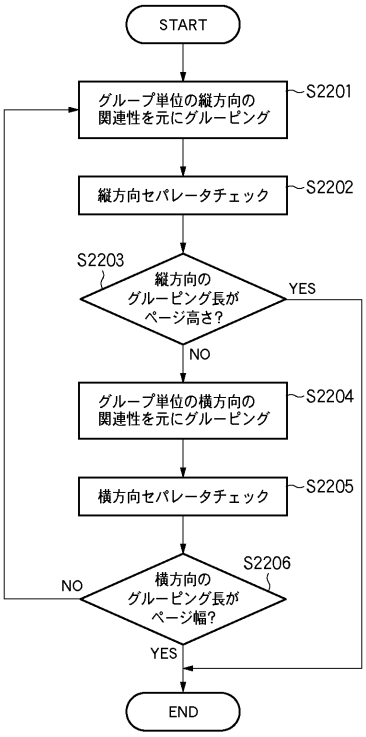
【図２０】



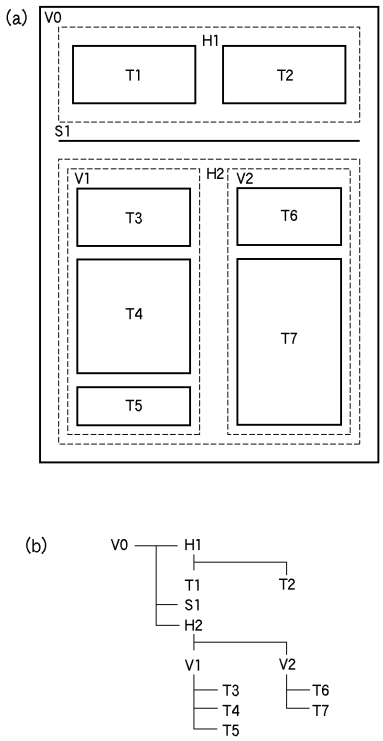
【図２１】



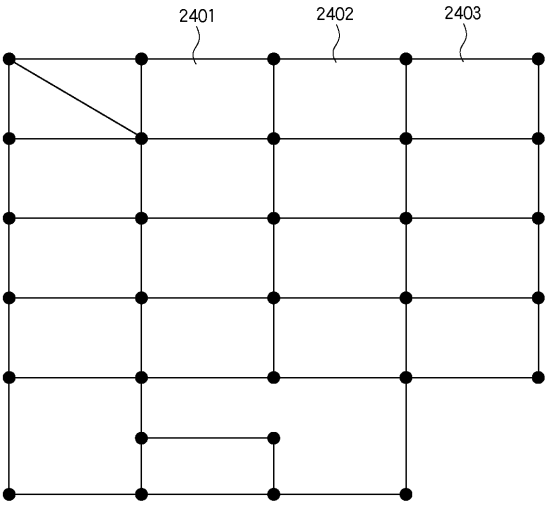
【図２２】



【図２３】



【図２４】



【図２５】