(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4230478号
(P4230478)
(45)発行日　平成21年2月25日(2009.2.25)　(24)登録日　平成20年12月12日(2008.12.12)

(51)Int.Cl.　F I
H04N 1/387 (2006.01)　H04N 1/387

請求項の数 9　(全 12 頁)

(21)出願番号　特願2005-150406 (P2005-150406)
(22)出願日　平成17年5月24日 (2005.5.24)
(65)公開番号　特開2006-332823 (P2006-332823A)
(43)公開日　平成18年12月7日 (2006.12.7)
審査請求日　平成17年5月24日 (2005.5.24)

(73)特許権者　503360115
独立行政法人科学技術振興機構
埼玉県川口市本町４丁目１番８号
(74)代理人　100103171
弁理士　雨貝　正彦
(72)発明者　寅市　和男
茨城県つくば市吾妻３－１－１ダイアパレスつくば学園都市１２１４
(72)発明者　諸岡　泰男
茨城県日立市塙山町２－２－９

審査官　渡辺　努

最終頁に続く

(54)【発明の名称】文書処理装置、方法およびプログラム

(57)【特許請求の範囲】
【請求項１】
テキスト文書領域と非テキスト文書領域とが混在するビットマップ形式の文書ファイルの中から、前記テキスト文書領域および前記非テキスト領域のいずれかに含まれる輪郭線を有する１つ以上の部分画像を抽出する部分画像抽出手段と、
前記部分画像抽出手段によって抽出された前記部分画像のレイアウト情報を生成するレイアウト情報生成手段と、
前記部分画像抽出手段によって抽出された前記部分画像に対して関数化近似処理を行って特徴量を抽出する関数化処理手段と、
前記レイアウト情報生成手段によって生成された前記レイアウト情報とともに、前記関数化処理手段によって抽出された特徴量を格納する文書情報格納手段と、
を備えることを特徴とする文書処理装置。
【請求項２】
請求項１において、
前記関数化処理手段は、前記部分画像の輪郭形状、濃度分布、色変化などを一あるいは複数の関数で近似する処理を行うことにより前記特徴量の抽出を行うことを特徴とする文書処理装置。
【請求項３】
請求項１または２において、
紙媒体に印刷された画像を光学的に読み取って前記文書ファイルを作成する文書ファイ

ル取込手段をさらに備えることを特徴とする文書処理装置。
【請求項４】
テキスト文書領域と非テキスト文書領域とが混在するビットマップ形式の文書ファイルの中から、前記テキスト文書領域および前記非テキスト領域のいずれかに含まれる輪郭線を有する１つ以上の部分画像を抽出する部分画像抽出ステップと、
前記部分画像抽出ステップによって抽出された前記部分画像のレイアウト情報を生成するレイアウト情報生成ステップと、
前記部分画像抽出ステップによって抽出された前記部分画像に対して関数化近似処理を行って特徴量を抽出する関数化処理ステップと、
前記レイアウト情報生成ステップによって生成された前記レイアウト情報とともに、前記関数化処理ステップによって抽出された特徴量を格納する文書情報格納ステップと
、
を有することを特徴とする文書処理方法。
【請求項５】
請求項４において、
紙媒体に印刷された画像を光学的に読み取って前記文書ファイルを作成する文書ファイル取込ステップをさらに有することを特徴とする文書処理方法。
【請求項６】
コンピュータを、
テキスト文書領域と非テキスト文書領域とが混在するビットマップ形式の文書ファイルの中から、前記テキスト文書領域および前記非テキスト領域のいずれかに含まれる輪郭線を有する１つ以上の部分画像を抽出する部分画像抽出手段と、
前記部分画像抽出手段によって抽出された前記部分画像のレイアウト情報を生成するレイアウト情報生成手段と、
前記部分画像抽出手段によって抽出された前記部分画像に対して関数化近似処理を行って特徴量を抽出する関数化処理手段と、
前記レイアウト情報生成手段によって生成された前記レイアウト情報とともに、前記関数化処理手段によって抽出された特徴量を格納する文書情報格納手段と、
して機能させる文書処理プログラム。
【請求項７】
請求項６において、
コンピュータを、さらに、紙媒体に印刷された画像を光学的に読み取って前記文書ファイルを作成する文書ファイル取込手段として機能させる文書処理プログラム。
【請求項８】
請求項１～３のいずれかに記載された前記文書情報格納手段から前記レイアウト情報と前記特徴量を読み出し、前記特徴量に基づいて前記部分画像を復元し、前記レイアウト情報に基づいてこの復元された部分画像の合成を行うことを特徴とする文書表示装置。
【請求項９】
請求項１～３のいずれかに記載された前記文書情報格納手段に格納された前記特徴量と、検索対象画像に対応する前記特徴量とに基づいて、前記検索対象画像に類似する前記部分画像の有無を判定することを特徴とする文書検索装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、文字と各種画像とが混在した文書をコンピュータで扱われる形式に変換する文書処理装置、方法およびプログラムに関する。
【背景技術】
【０００２】
従来から、ＸＭＬ、ＳＧＭＬ、ＨＴＭＬなどの論理構造言語によって文書処理を行う手法が知られている（例えば、特許文献１～４参照。）。これらの論理構造言語を用いるこ

とにより、コンピュータに適した形式で文書を処理することが可能になる。
【特許文献１】特開平１１－２５００４１号公報（第６－２１頁、図１－１４）
【特許文献２】特開２００３－３０８３１１号公報（第３－６頁、図１－６）
【特許文献３】特開２００３－３１６７６６号公報（第５－１４頁、図１－２３）
【特許文献４】特開２００４－１７８０１０号公報（第７－１８頁、図１－１７）
【発明の開示】
【発明が解決しようとする課題】
【０００３】
ところで、上述した特許文献１～４に開示された各種の文書処理では、主にテキストに対してレイアウト解析処理や文書論理構造解析処理がなされており、図形や画像等の非テキストに対しては原画像の状態で取り扱われている。例えば、「ＢＭＰ」や「ＧＩＦ」等の拡張子が付された画像データが文書の一部に含まれている場合には、これらの画像データがそのままの状態で文書データの一部としてコンピュータに格納される。したがって、図形や画像等の非テキスト情報に対して表示等の処理を行う場合には、データの種類に合わせた処理が必要になり、文書処理が複雑になるという問題があった。例えば、市場に出回っている多くの種類の非テキスト情報を対象にした文書処理を行う場合には、これら全ての種類の非テキスト情報を扱うことが可能な表示処理プログラムをあらかじめコンピュータにインストールしておく必要があり、処理の負担も大きくなる。また、テキスト部分と非テキスト部分とでは扱うデータ形式が全く異なるため、さらに文書処理が複雑になる。さらに、このような文書の内容を表示する場合に、テキスト部分を様々なフォントで表示しようとすると、これらのフォントデータを表示装置においてあらかじめインストールしておく必要があるため、その作業に要する手間や表示装置における処理の負担も大きくなる。
【０００４】
本発明は、このような点に鑑みて創作されたものであり、その目的は、文書処理の負担を軽減可能な文書ファイルを作成することができる文書処理装置、方法およびプログラムを提供することにある。
【課題を解決するための手段】
【０００５】
上述した課題を解決するために、本発明の文書処理装置は、テキスト文書領域と非テキスト文書領域とが混在する文書ファイルの中から、テキスト文書領域および非テキスト領域のいずれかに含まれる１つ以上の部分画像を抽出する部分画像抽出手段と、部分画像抽出手段によって抽出された部分画像のレイアウト情報を生成するレイアウト情報生成手段と、部分画像抽出手段によって抽出された部分画像に対して関数化近似処理を行って特徴量を抽出する関数化処理手段と、レイアウト情報生成手段によって生成されたレイアウト情報とともに、関数化処理手段によって抽出された特徴量を格納する文書情報格納手段とを備えている。なお、テキスト文書領域に含まれる部分画像とは、通常の画像のみでなく、文字も一つの画像として含まれる。
【０００６】
また、本発明の文書処理方法は、テキスト文書領域と非テキスト文書領域とが混在する文書ファイルの中から、テキスト文書領域および非テキスト領域のいずれかに含まれる１つ以上の部分画像を抽出する部分画像抽出ステップと、部分画像抽出ステップによって抽出された部分画像のレイアウト情報を生成するレイアウト情報生成ステップと、部分画像抽出ステップによって抽出された部分画像に対して関数化近似処理を行って特徴量を抽出する関数化処理ステップと、レイアウト情報生成ステップによって生成されたレイアウト情報とともに、関数化処理ステップによって抽出された特徴量を格納する文書情報格納ステップとを有している。
【０００７】
また、本発明の文書処理プログラムは、コンピュータを、テキスト文書領域と非テキスト文書領域とが混在する文書ファイルの中から、テキスト文書領域および非テキスト領域

のいずれかに含まれる１つ以上の部分画像を抽出する部分画像抽出手段と、部分画像抽出手段によって抽出された部分画像のレイアウト情報を生成するレイアウト情報生成手段と、部分画像抽出手段によって抽出された部分画像に対して関数化近似処理を行って特徴量を抽出する関数化処理手段と、レイアウト情報生成手段によって生成されたレイアウト情報とともに、関数化処理手段によって抽出された特徴量を格納する文書情報格納手段として機能させる。

【０００８】

これにより、非テキスト文書領域に含まれる図形や画像等の種類やデータの属性に関係なく関数化近似して得られた特徴量で非テキスト文書領域の内容を定義することが可能になり、テキスト文書領域と非テキスト文書領域とが混在する文書ファイルをその後に読み出して表示等の文書ファイル処理を行う場合の処理負担を軽減することができる。すなわち、文書処理の負担を軽減可能な文書ファイルを作成することができる。また、非テキスト文書領域に含まれる図形や画像等が関数化処理されるため、ビットマップ形式で画像データを保持する場合に比べてデータ量を削減することができる。さらに、テキスト文書領域と非テキスト文書領域の全体に含まれるテキスト画像を含む各種の部分画像を関数近似処理によって得られた特徴量を用いて統一的に取り扱うことができるため、数々のデータ形式に合わせた処理が不要になり、さらに処理の簡略化が可能になる。

【０００９】

また、上述した関数化処理手段は、部分画像の輪郭形状、濃度分布、色変化などを一あるいは複数の関数で近似する処理を行うことにより特徴量の抽出を行うことが望ましい。これにより、テキスト文書領域および非テキスト文書領域に含まれる部分画像の内容を関数近似することが可能になり、その後の文書ファイル処理によってこの部分画像を復元することができる。

【００１０】

また、紙媒体に印刷された画像を光学的に読み取って文書ファイルを作成する文書ファイル取込手段をさらに備えることが望ましい。あるいは、紙媒体に印刷された画像を光学的に読み取って文書ファイルを作成する文書ファイル取込ステップをさらに有することが望ましい。また、コンピュータを、さらに、紙媒体に印刷された画像を光学的に読み取って文書ファイルを作成する文書ファイル取込手段として機能させることが望ましい。これにより、紙媒体に印刷されたテキスト文書領域と非テキスト文書領域とが混在した文書を取り込んでコンピュータの処理に適した形式で格納することが可能になる。

【００１１】

また、本発明の文書表示装置は、上述した文書情報格納手段からレイアウト情報と特徴量を読み出し、特徴量に基づいて部分画像を復元し、レイアウト情報に基づいてこの復元された部分画像の合成を行っている。これにより、テキスト文書領域と非テキスト文書領域とが混在する文書ファイルの内容を表示する際に、テキスト文書と非テキスト文書の両方を同じ処理によって復元することが可能になり、表示処理手順を簡略化することができる。

【００１２】

また、本発明の文書検索装置は、上述した文書情報格納手段に格納された特徴量と、検索対象画像に対応する特徴量とに基づいて、検索対象画像に類似する部分画像の有無を判定している。これにより、テキスト文書領域と非テキスト文書領域のいずれかに含まれる部分画像に対して同じ処理手順で検索を行うことが可能になり、検索処理手順を簡略化することができる。

【発明を実施するための最良の形態】

【００１３】

以下、本発明を適用した一実施形態の文書処理装置について、図面を参照しながら詳細に説明する。図１は、一実施形態の文書処理装置の構成を示す図である。図１に示すよう

に、本実施形態の文書処理装置は、スキャナ１０、文書ファイル取込部１２、文書ファイル格納部１４、部分画像抽出部２０、関数化処理部３０、レイアウト情報生成部４０、文書ファイル格納部５０を含んで構成されている。この文書処理装置は、ＣＰＵ、ＲＯＭ、ＲＡＭ、ハードディスク装置を有するコンピュータによって、あらかじめハードディスク装置にインストールされた文書処理プログラムを実行することにより実現される。

【００１４】

文書ファイル取込部１２は、テキスト文書領域と非テキスト文書領域とが混在した文書ファイルを取り込む処理を行う。具体的には、スキャナ１０を用いて、原稿台にセットされた紙媒体に印刷された画像を光学的に読み取って、処理対象となる文書ファイルを作成する。作成された文書ファイルは、画素毎に２値（白黒）の値が対応したビットマップデータ形式の画素データによって構成されている。文書ファイル格納部１４は、文書ファイル取込部１２によって取り込まれた文書ファイルを格納する。

【００１５】

部分画像抽出部２０は、文書ファイルに含まれるテキスト文書領域と非テキスト文書領域とを抽出し、さらに、これらの各文書領域に含まれる部分画像を抽出する。例えば、テキスト文書領域内の一箇所あるいは複数箇所に含まれるテキスト（文字）に対応する部分画像や、非テキスト文書領域に含まれる図形や画像等に対応する部分画像を抽出する。なお、本実施形態では、テキスト文書領域にはＪＩＳコード等のコードデータではなく、読み取った２値画像として部分画像が含まれているため、テキスト文書領域と非テキスト文書領域のいずれに含まれる部分画像であるかを区別する必要はない。したがって、テキスト文書領域と非テキスト文書領域の抽出処理を省略して、直接部分画像の抽出処理を行うようにしてもよい。

【００１６】

関数化処理部３０は、部分画像抽出部２０によって抽出された部分画像の輪郭形状、濃度分布、色変化などを一あるいは複数の関数で近似する処理を行う。例えば、関数化処理の対象となる部分画像が図形や２値画像（この２値画像にはテキストに対応する部分画像も含まれる）の場合には、その輪郭形状を関数近似する処理が行われる。具体的には、この関数化処理部３０は、輪郭追跡処理部３２、接合点抽出処理部３４、関数近似処理部３６を備えている。

【００１７】

輪郭追跡処理部３２は、部分画像抽出部２０によって抽出された各部分画像に含まれる一あるいは複数の輪郭線を抽出する。具体的には、輪郭追跡処理部３２は、部分画像の画素データを用いて輪郭線を所定方向に追跡して、この輪郭線を構成する画素列（輪郭点列）を抽出する。例えば、抽出された画素列の特定は、Ｘ座標およびＹ座標のそれぞれの各座標値について別々に行われる。また、一の部分画像に複数の輪郭線が含まれている場合には、各輪郭線について輪郭点列の抽出が行われる。なお、上述した輪郭線の抽出処理は、被検索対象画像が白黒あるいは単色のみを用いた画像であるか、濃淡分布を有する中間調画像や色変化を有するカラー画像であるかによって場合を分けることが望ましい。すなわち、白黒あるいは単色のみを用いた画像の場合には、輪郭追跡処理部３２は、部分画像と背景との間の境界部を輪郭線として抽出する。また、中間調画像やカラー画像の場合には、輪郭追跡処理部３２は、部分画像と背景との間の境界部と、部分画像の内部領域に現れる同一濃淡あるいは同一色の縁部とを輪郭線として抽出する。

【００１８】

図２は、輪郭追跡処理部３２によって抽出された輪郭点列の概略を示す図である。また、図３は抽出された輪郭点列のＸ座標を媒介変数を用いて分離した変化の様子を示す図である。図４は、抽出された輪郭点列のＹ座標を媒介変数を用いて分離した変化の様子を示す図である。

【００１９】

図２では、丸印（○）が輪郭線を構成する画素を示しており、各丸印に付された数字は輪郭線を追跡していったときの画素の順番を示している。なお、実際の部分画像の輪郭線

は図２に示す例に比べて多くの画素によって構成されているが、図２では説明を簡略化するために少ない数の画素によって輪郭点列が構成されているものとする。

【００２０】

例えば、輪郭追跡処理部３２は、Ｘ座標が最も小さい位置からＹ方向に沿って走査を開始し、Ｘ座標を大きくしていって最初に検出した画素に番号「１」を付す。図２に示した例では、輪郭追跡処理部３２は、番号「１」の画素を追跡開始画素として時計回り方向に輪郭線を追跡しながら、輪郭線を構成する各画素を検出するとともにこれらの各画素に検出順に通し番号「２」、「３」、…を付す。この輪郭線に沿った画素の検出動作は、検出する画素が追跡開始画素に一巡するまで行われる。輪郭線を構成する各画素の番号を横軸に、各画素のＸ座標値を縦軸にプロットしたものが図３である。また、輪郭線を構成する各画素の番号を横軸に、各画素のＹ座標値を縦軸にプロットしたものが図４である。このように、輪郭追跡処理部３２は、輪郭線を構成する各画素に付した検出順番を示す通し番号を媒介変数として、Ｘ座標値とＹ座標値を別々に記録することにより、輪郭点列の抽出を行う。なお、上述した説明では、一例として媒介変数を用いてＸ座標値とＹ座標値を別々に記録するようにしたが、媒介変数を用いずに、Ｘ座標値とＹ座標値の組み合わせを記録するようにしてもよい。また、一般には、部分画像には、この画像と背景との間の境界部としての多くの輪郭線が含まれるが、各輪郭線毎に輪郭点列の抽出が行われる。

【００２１】

接合点抽出処理部３４は、輪郭追跡処理部３４によって抽出した輪郭点列に基づいて、輪郭線の傾向が変化する接合点を抽出する。例えば、輪郭線の角度が急に変化する角点が接合点として抽出される。接合点の抽出処理や関数近似処理は、図３に示すＸ座標についての輪郭点列と図４に示すＹ座標についての輪郭点列のそれぞれについて別々に行われる。

【００２２】

関数近似処理部３６は、輪郭線に沿って隣接する２つの接合点で区分される部分的な領域（区分領域）を、直線、円弧、自由曲線のいずれかの関数を用いて近似し、この近似処理に関連する特徴情報を作成する。例えば、区分領域が直線で近似可能な場合には近似関数として直線が用いられ、直線で近似不可能であって円弧で近似可能な場合には近似関数として円弧が用いられる。円弧でも近似不可能な場合には近似関数として自由曲線が用いられる。近似関数として直線を用いた場合には、用いた関数が直線であることを示す符号と、直線で近似される区分領域の形状を示すパラメータとが、この区分領域に対応する近似関数に関する特徴量として作成される。同様に、近似関数として円弧を用いた場合には、用いた関数が円弧であることを示す符号と、円弧で近似される区分領域の形状を示すパラメータとが、この区分領域に対応する近似関数に関する特徴量として作成される。近似関数として自由曲線を用いた場合には、用いた関数が自由曲線であることを示す符号と、自由曲線で近似される区分領域の形状を示すパラメータとが、この区分領域に対応する近似関数に関する特徴量として作成される。このようにして作成された部分画像内の各輪郭線に対応する特徴量が文書ファイル格納部５０に格納される。

【００２３】

なお、着目している区分領域がどの関数で近似可能であるか否かの判定は、区分領域と近似関数との間の誤差（最小二乗法で求めた誤差）が所定値以下であるか否かを調べることにより行われる。また、区分領域の形状を示すパラメータは、この区分領域の形状を特定することが可能であればよいが、例えば、特許第２６４６４７５号公報に開示されているように、以下に示すものを用いるようにしてもよい。

（１）直線の場合：直線を示すフラグ、区分領域の始点の座標

（２）円弧の場合：円弧を示すフラグ、円弧の始点の座標、接合点間の中心角の係数、接合点間に存在する輪郭点数、近似関数の係数（円弧を例えば三角関数の線形結合の式で表現した場合の各係数）

（３）自由曲線の場合：接合点間の自由曲線を示す近似関数の次元数（≧３）、接合点間に存在する輪郭点数、接合点間における輪郭点列の変動の中心、近似関数の係数。

【００２４】
図５および図６は、輪郭追跡処理部３２、接合点抽出処理部３４、関数近似処理部３６の各処理によって抽出される特徴量の概要を示す図である。図５に示すひとまとまりの特徴量が一の部分画像について抽出される。図５に示す例では、着目している部分画像には、輪郭線１、２、３、…で示される複数の輪郭線が含まれており（それぞれの輪郭長が輪郭長１、２、３…）で、その中で最も長い輪郭線の長さが「最大輪郭長」で示されている。また、各輪郭線には、Ｘ軸関数表とＹ軸関数表とが対応付けられている。
【００２５】
図６に示すように、Ｘ軸関数表には、関数総数、輪郭長、総標本点数、直線個数、直線総長、円弧個数、円弧総長、曲線個数、曲線総長の他に、各輪郭線毎の区間長、標本点数、始点標本番号（図２や図３において示した通し番号）、各関数に対応する区間長やパラメータが含まれている。関数総数は、着目している輪郭線に含まれる関数の総数であって区分領域の数に等しい。輪郭長は、着目している輪郭線の長さである。直線個数は、着目している輪郭線を構成する各区分領域の中で直線によって近似される区分領域の数である。直線総長は、着目している輪郭線を構成する各区分領域の中で直線によって近似される区分領域の長さの合計値である。円弧個数は、着目している輪郭線を構成する各区分領域の中で円弧によって近似される区分領域の数である。円弧総長は、着目している輪郭線を構成する各区分領域の中で円弧によって近似される区分領域の長さの合計値である。曲線個数は、着目している輪郭線を構成する各区分領域の中で自由曲線によって近似される区分領域の数である。曲線総長は、着目している輪郭線を構成する各区分領域の中で自由曲線によって近似される区分領域の長さの合計値である。また、図６において、「Ｘ軸関数」に対応する複数の関数は、着目している輪郭線を構成する各区分領域を近似する関数を示しており、これらの配置順が各区分領域の並びに対応している。なお、図６に示した特徴量は、後に文書ファイル内の画像検索を行うことができるように多くの項目を含ませているが、単に文書ファイルを復元して表示や印刷を行うことができればよい場合にはこれら全ての項目を抽出する必要はない。例えば、「Ｘ軸関数」、「始点座標」、「パラメータ」があれば各輪郭点列の両端座標やその間の形状が再現できるため、文書ファイル内の各部分画像の表示や印刷が可能になる。
【００２６】
レイアウト情報生成部４０は、文書ファイル内のテキスト文書領域と非テキスト文書領域のそれぞれに含まれる部分画像のレイアウト情報を作成する。このレイアウト情報は文書ファイル格納部５０に格納される。
【００２７】
上述した部分画像抽出部２０が部分画像抽出手段に、レイアウト情報生成部４０がレイアウト情報生成手段に、関数化処理部３０が関数化処理手段に、文書ファイル格納部５０が文書情報格納手段に、スキャナ１０、文書ファイル取込部１２が文書ファイル取込手段にそれぞれ対応する。また、部分画像抽出部２０による動作が部分画像抽出ステップの動作に、レイアウト情報生成部４０による動作がレイアウト情報生成ステップの動作に、関数化処理部３０による動作が関数化処理ステップの動作に、文書ファイル格納部５０によってレイアウト情報と特徴量を格納する動作が文書情報格納ステップの動作に、スキャナ１０、文書ファイル取込部１２によって文書ファイルを取り込む動作が文書ファイル取込ステップの動作に対応する。
【００２８】
本実施形態の文書処理装置はこのような構成を有しており、次にその動作を説明する。図７は、本実施形態の文書処理装置を用いて取り込んだ文書ファイルの形式をコンピュータ処理に適した形式に変換する動作手順を示す図である。
【００２９】
まず、文書ファイル取込部１２はスキャナ１０を用いて紙媒体に印刷された画像を読み取ることにより、ファイル形式の変換処理の対象となるビットマップ形式の文書ファイルの取り込みを行う（ステップ１００）。取り込まれた文書ファイルは文書ファイル格納部

１４に格納される。

【００３０】

次に、部分画像抽出部２０は、文書ファイル格納部１４に格納された文書ファイルを読み出して、その中に含まれるテキスト文書領域と非テキスト文書領域を抽出し、さらにこれらの文書領域に含まれる部分画像を抽出する（ステップ１０１）。

【００３１】

図８は、テキスト文書領域と非テキスト文書領域に含まれる部分画像の抽出動作の説明図である。図８に示すように、スキャナ１０の原稿台にセットされた紙媒体は、文字が含まれる２つのテキスト文書領域Ａ１、Ａ２と、イラストや地図、写真、図形が含まれる５つの非テキスト文書領域Ｂ１～Ｂ５が含まれている。部分画像抽出部２０は、これらの２つのテキスト文書領域Ａ１、Ａ２と５つの非テキスト文書領域Ｂ１～Ｂ５に含まれる部分画像を抽出する。

【００３２】

次に、レイアウト情報作成部４０は、部分画像抽出部２０による部分画像の抽出結果に基づいてこれらの部分画像の配置を示すレイアウト情報を作成し、文書ファイル格納部５０に格納する（ステップ１０２）。また、部分画像抽出部２０は抽出した一の部分画像を選択し（ステップ１０３）、関数化処理部３０はこの選択された部分画像に対する関数化処理を行って（ステップ１０４）、その処理結果としての特徴量を作成する（ステップ１０５）。作成された特徴量は文書ファイル格納部５０に格納される。

【００３３】

その後、部分画像抽出部２０は、未処理の部分画像があるか否かを判定する（ステップ１０６）。未処理の部分画像がある場合には肯定判断が行われ、次の一の部分画像を選択するステップ１０３以後の動作が繰り返される。また、全ての部分画像について処理が終了した場合にはステップ１０６の判定において否定判断が行われ、一連のファイル変換処理が終了する。

【００３４】

このように、本実施形態の文書処理装置では、非テキスト文書領域に含まれる図形や画像等の種類やデータの属性に関係なく関数化近似して得られた特徴量で非テキスト文書領域の内容を定義することが可能になり、テキスト文書領域と非テキスト文書領域とが混在する文書ファイルをその後に読み出して表示等の文書ファイル処理を行う場合の処理負担を軽減することができる。すなわち、文書処理の負担を軽減可能な文書ファイルを作成することができる。また、非テキスト文書領域に含まれる図形や画像等が関数化処理されるため、ビットマップ形式で画像データを保持する場合に比べてデータ量を削減することができる。さらに、テキスト文書領域と非テキスト文書領域の全体に含まれるテキスト画像を含む各種の部分画像を関数近似処理によって得られた特徴量を用いて統一的に取り扱うことができるため、数々のデータ形式に合わせた処理が不要になり、さらに処理の簡略化が可能になる。

【００３５】

特に、部分画像に対する関数近似処理をこの部分画像の輪郭形状、濃度分布、色変化などを一あるいは複数の関数で近似して行うことにより、図形や２値画像、濃淡画像、カラー画像のそれぞれを同じ処理手順で関数近似することが可能になり、しかも、その後の文書ファイル処理によってこの部分画像を復元することができる。

【００３６】

次に、上述した文書処理装置を用いて作成された文書ファイルを用いて各種の処理を行う場合の具体例を簡単に説明する。例えば、各種の処理として、文書ファイルを用いた表示処理と検索処理について説明する。

【００３７】

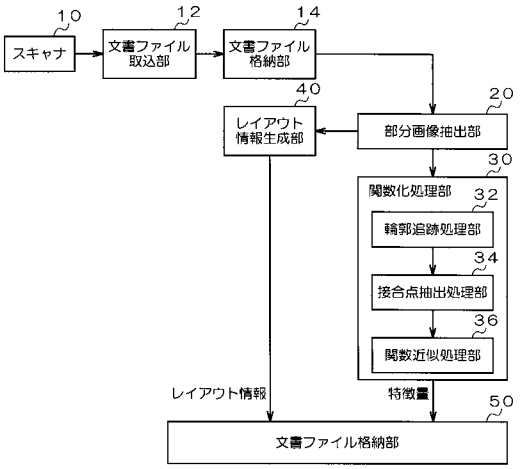
文書ファイルの表示処理（例えば、コンピュータによって構成された文書表示装置によって行われる）は、以下の手順を実行することにより行われる。

（ステップａ１）レイアウト情報を読み込む。

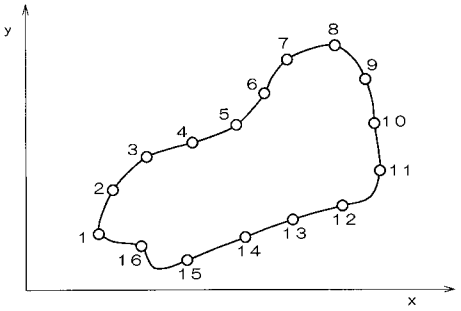
（ステップａ２）関数化処理によって得られた特徴量を読み込む。
（ステップａ３）各部分画像を復元する。
（ステップａ４）テキスト文書領域と非テキスト文書領域の両方に含まれる各部分画像をレイアウト情報に基づいて合成する。
【００３８】
これにより、テキスト文書領域と非テキスト文書領域とが混在する文書ファイルの内容を表示する際に、テキスト文書と非テキスト文書の両方を同じ処理によって復元することが可能になり、表示処理手順を簡略化することができる。
【００３９】
また、文書ファイルを用いた検索処理（例えば、コンピュータによって構成された文書検索装置によって行われる）は、以下の手順を実行することにより行われる。例えば、検索対象となる画像が指定され、非テキスト文書領域にこの検索対象画像と一致（あるいは類似）する部分画像が含まれているか否かを検索する動作が行われる。また、検索対象画像に対応する関数化近似処理がその都度あるいは前処理によって行われ、図６に示した特徴量の一部あるいは全部が検索処理前に抽出されているものとする。
（ステップｂ１）検索対象図形を指定する。この場合の検索対象図形には、テキスト画像も含まれる。
（ステップｂ２）文書ファイルに含まれる各部分画像の特徴量を読み出す。
（ステップｂ３）検索対象画像の特徴量と文書ファイルに含まれる各部分画像の特徴量とを比較して、検索対象画像に類似する部分画像の有無を判定する（あるいは各部分画像の類似度を判定する）。例えば、特徴量の中の輪郭長（輪郭線の長さ）、輪郭数（輪郭線の数）、輪郭線を構成する複数の関数の順番、輪郭線を構成する複数の関数のそれぞれに対応する区間長の並びの中の一つあるいは複数に着目して、それらの値が近いものほど類似度が高いと判定される。このような画像比較は、従来のビットマップ形式の画像データの場合には複雑な処理が必要であったが、本実施形態のように関数近似によって得られた特徴量を用いた場合には比較的簡単な処理で実施することができる。
【００４０】
これにより、テキスト文書領域と非テキスト文書領域のいずれかに含まれる部分画像に対して同じ処理手順で検索を行うことが可能になり、検索処理手順を簡略化することができる。
【００４１】
なお、本発明は上記実施形態に限定されるものではなく、本発明の要旨の範囲内で種々の変形実施が可能である。例えば、上述した実施形態では、スキャナ１０を用いて紙媒体の文書を読み込んで文書ファイルを作成したが、ワープロソフトで作成した文書ファイルや、ＨＴＭＬ形式等の文書ファイルを文書ファイル取込部１２によって直接取り込むようにしてもよい。この場合には、テキスト文書領域にテキスト画像そのものが含まれているわけではない。部分画像抽出部２０は、文書ファイルに含まれるテキスト文書領域を抽出した後、このテキスト文書領域に対応するコードデータに基づいてこのテキスト領域に含まれるテキスト画像（ビットマップデータ）を復元し、このテキスト画像を部分画像として抽出する。
【図面の簡単な説明】
【００４２】
【図１】一実施形態の文書処理装置の構成を示す図である。
【図２】輪郭追跡処理部によって抽出された輪郭点列の概略を示す図である。
【図３】抽出された輪郭点列のＸ座標を媒介変数を用いて分離した変化の様子を示す図である。
【図４】抽出された輪郭点列のＹ座標を媒介変数を用いて分離した変化の様子を示す図である。
【図５】輪郭追跡処理部、接合点抽出処理部、関数近似処理部の各処理によって抽出される特徴量の概要を示す図である。

【図６】輪郭追跡処理部、接合点抽出処理部、関数近似処理部の各処理によって抽出される特徴量の概要を示す図である。
【図７】本実施形態の文書処理装置を用いて取り込んだ文書ファイルの形式をコンピュータ処理に適した形式に変換する動作手順を示す図である。
【図８】テキスト文書領域と非テキスト文書領域に含まれる部分画像の抽出動作の説明図である。
【符号の説明】
【００４３】
１０　スキャナ
１２　文書ファイル取込部
１４、５０　文書ファイル格納部
２０　部分画像抽出部
３０　関数化処理部
３２　輪郭追跡処理部
３４　接合点抽出処理部
３６　関数近似処理部
４０　レイアウト情報生成部

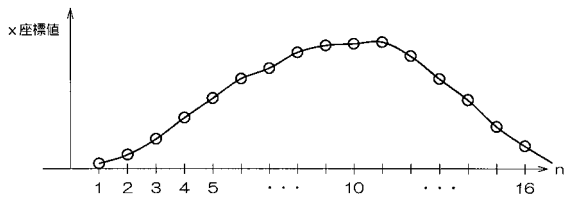
【図１】



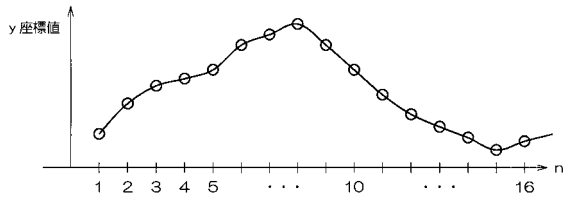
【図２】



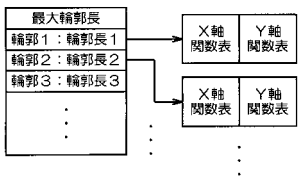
【図３】



【図４】



【図５】

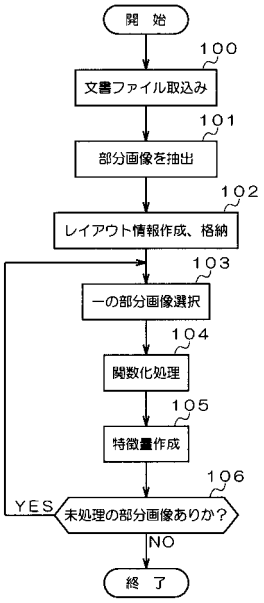


【図６】

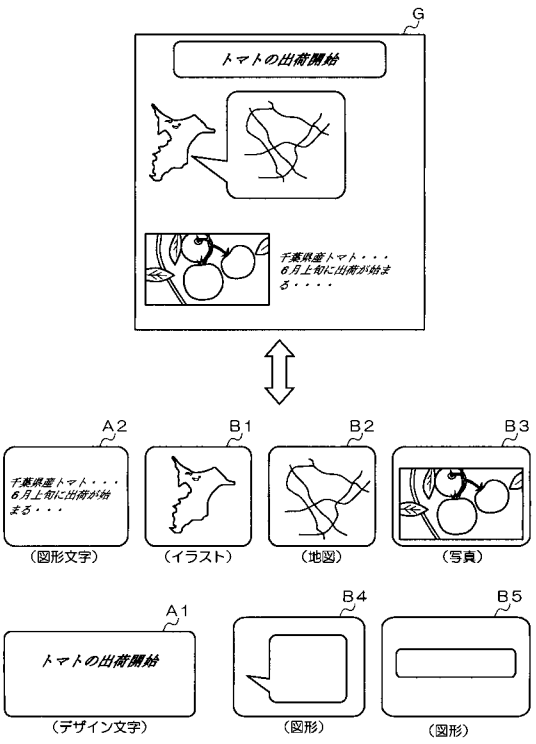
X軸関数	区間長	標本点数	始点座標	始点標本No	パラメータ
関数1	区間長L1	点数:m1	x1	k1	(a1,b1,c1)
関数2	区間長L2	点数:m2	x2	k2	(a2,b2,c2)
関数1	区間長L3	点数:m3	x3	k3	(a3,b3,c3)
関数3	区間長L4	点数:m4	x4	k4	(a4,b4,c4)
関数2	区間長L5	点数:m5	x5	k5	(a5,b5,c5)
関数3	区間長L6	点数:m6	x6	k6	(a6,b6,c6)
関数1	区間長L7	点数:m7	x7	k7	(a7,b7,c7)
・・・	・・・	・・・	・・・	・・・	・・・

関数総数	輪郭長	総標本点数
直線個数	直線総長	—
円弧個数	円弧総長	—
曲線個数	曲線総長	—

【図７】



【図８】



フロントページの続き

(56)参考文献 特開２０００－０５９６０５（ＪＰ，Ａ）
特開平０６－３４８８３７（ＪＰ，Ａ）
特開平０８－０６３４７４（ＪＰ，Ａ）
特開平０４－２４６７７２（ＪＰ，Ａ）
特開２００１－０４７７９６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｎ　　１／３８７