(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3768738号
(P3768738)
(45)発行日　平成18年4月19日(2006.4.19)　(24)登録日　平成18年2月10日(2006.2.10)

(51)Int.Cl.　　F I
G06T 1/00 (2006.01)　　G06T 1/00 ２００Ｄ
G06F 17/30 (2006.01)　　G06F 17/30 １７０Ｂ

請求項の数 4　(全 12 頁)

(21)出願番号　特願平11-199753
(22)出願日　平成11年7月14日(1999.7.14)
(65)公開番号　特開2001-28041(P2001-28041A)
(43)公開日　平成13年1月30日(2001.1.30)
審査請求日　平成14年10月29日(2002.10.29)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100087848
弁理士　小笠原　吉義
(74)代理人　100074848
弁理士　森田　寛
(74)代理人　100087147
弁理士　長谷川　文廣
(72)発明者　勝山　裕
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

最終頁に続く

(54)【発明の名称】電子ファイリングシステム，表紙識別処理装置およびそれらのプログラム記録媒体

(57)【特許請求の範囲】

【請求項１】

文書を電子化して登録し保存する電子ファイリングシステムにおいて，

複数文書の各ページを連続して読み込み可能な入力装置により各ページの文書画像を入力する文書画像入力処理手段と，

前記文書画像の領域を識別し，その領域情報をもとに文書画像が表紙であるか否かを識別する表紙識別処理手段と，

前記表紙識別処理手段による識別結果にもとづいて，入力された文書画像群を表紙の文書画像で区切り，文書を登録する文書登録処理手段とを備え，

前記表紙識別処理手段として，少なくとも，

文書画像中の文字列，図または表などの領域を識別する手段と，

前記領域を識別した結果にもとづいて，それらの領域のレイアウト情報を解析し，表紙らしさの評価得点を計算する手段と，

前記表紙らしさの得点をもとに，その文書画像を表紙と判定する手段とを備える

ことを特徴とする電子ファイリングシステム。

【請求項２】

文書画像が表紙であるか否かを識別する表紙識別処理装置であって，

文書画像中の文字列，図または表などの領域を識別する手段と，

前記領域を識別した結果にもとづいて，それらの領域のレイアウト情報を解析し，表紙らしさの評価得点を計算する手段と，

前記表紙らしさの得点をもとに，その文書画像を表紙と判定する手段とを備える
ことを特徴とする表紙識別処理装置。
【請求項３】
文書を電子化して登録し保存する電子ファイリングシステムをコンピュータによって実現するため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であって，
複数文書の各ページを連続して読み込み可能な入力装置により各ページの文書画像を入力する処理と，
前記文書画像の領域を識別し，その領域情報をもとに文書画像が表紙であるか否かを識別する処理と，
表紙であるか否かの識別結果にもとづいて，入力された文書画像群を表紙の文書画像で区切り，文書を登録する処理とを，コンピュータに実行させるとともに，
前記表紙であるか否かを識別する処理では，少なくとも，
文書画像中の文字列，図または表などの領域を識別する処理と，
前記領域を識別した結果にもとづいて，それらの領域のレイアウト情報を解析し，表紙らしさの評価得点を計算する処理と，
前記表紙らしさの得点をもとに，その文書画像を表紙と判定する処理とを，
コンピュータに実行させるプログラムを記録した
ことを特徴とする電子ファイリングシステムのプログラム記録媒体。
【請求項４】
文書画像が表紙であるか否かを識別する表紙識別処理装置を，コンピュータによって実現するため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であって，
文書画像中の文字列，図または表などの領域を識別する処理と，
前記領域を識別した結果にもとづいて，それらの領域のレイアウト情報を解析し，表紙らしさの評価得点を計算する処理と，
前記表紙らしさの得点をもとに，その文書画像を表紙と判定する処理とを，
コンピュータに実行させるプログラムを記録した
ことを特徴とする表紙識別処理プログラム記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は，複数ページを連続して入力可能な自動ドキュメントフィーダ（ＡＤＦ）機能を持つスキャナ等から複数文書を連続して読み込み，文書単位に登録し保存するための電子ファイリングシステム，表紙識別処理装置およびそれらのプログラム記録媒体であって，特に複数の文書を自動的に文書単位に区切り，自動登録することを可能にした技術に関する。
【０００２】
【従来の技術】
近年，オフィスの効率化のために，紙文書を電子化して光ディスクなどに保存する「電子ファイリングシステム」が広く用いられるようになってきている。
【０００３】
従来の電子ファイリングシステムでは，多量の文書のページを連続して読み込むことができるＡＤＦ機能を有するスキャナ等が利用されている。しかし，複数文書をＡＤＦ付スキャナなどで入力する際に，１文書ごとの区切りを意識して操作する必要があった。例えば，
１）通常の電子ファイリングシステムでは，ユーザは，複数文書を登録する場合，スキャナのＡＤＦのスタッカに１文書ずつまとめて文書を載せ，１文書単位で入力操作を行い，１文書を登録した後に，同様な操作によって次の１文書の登録を繰り返していた。
【０００４】
２）また，複数文書登録機能を持つ電子ファイリングシステムでは，連続入力した複数ページを文書単位に区切って登録するために，区切り認識用の専用紙を用いる。このため，

ユーザは，入力操作の前に，予め文書と文書の間にこの区切り用紙を挿入しておかなければならなかった。

【０００５】

【発明が解決しようとする課題】

このように，従来の技術では，いずれの方法によっても，ユーザ自らが，文書のまとまり（文書単位）を意識して入力操作を行わなければならなかった。特に，登録対象の文書が大量にある場合などは，このような文書単位にまとめてから入力操作を行うことは，ユーザにとって大変な労力の負担となっていた。

【０００６】

本発明は，文書の先頭ページは通常表紙であることに着目し，スキャナのスタッカに積まれた文書の１枚１枚の画像から，それが表紙であるか否かを自動的に識別する手段，および表紙とみなした場合に，そこを文書の区切りと判断し，複数文書を１文書ずつ分離して自動登録する手段を提供し，ユーザは文書単位を意識することなく複数文書を連続して入力操作でき，ユーザの労力を軽減し，また，入力装置のＡＤＦ機能も有効に活用することができるようにすることを目的とする。

【０００７】

【課題を解決するための手段】

本発明の電子ファイリングシステムは，文書を電子化して登録し保存するにあたって，複数文書の各ページを連続して読み込み可能な入力装置により各ページの文書画像を入力する文書画像入力処理手段と，入力した文書画像の領域を識別し，その領域情報をもとに文書画像が表紙であるか否かを識別する表紙識別処理手段と，これによる識別結果にもとづいて，入力された文書画像群を表紙の文書画像で区切り，文書を登録する文書登録処理手段とを備える。

【０００９】

特に本発明は，文書画像が表紙であるか否かを自動識別するため，文書画像中の文字列，図または表などの領域を識別する手段と，領域を識別した結果にもとづいて，それらの領域のレイアウト情報を解析し，表紙らしさの評価得点を計算する手段と，計算した表紙らしさの得点をもとに，その文書画像を表紙と判定する手段とを備えることを特徴とする。

【００１０】

以上の各処理手段をコンピュータによって実現するためのプログラムは，コンピュータが読み取り可能な可搬媒体メモリ，半導体メモリ，ハードディスクなどの適当な記録媒体に格納することができる。

【００１１】

動作の概要は，以下のとおりである。まず，入力した文書画像の領域を識別し，文字列，図，表等の文書構成要素に関する情報（領域矩形の位置，サイズ，領域の属性など）を抽出し，抽出した情報を用いて，文書構成要素中の文字列のタイトルらしさの得点を計算する。または文書構成要素のレイアウト（配置）情報から，その文書画像の表紙らしさの評価得点を計算する。これらの計算結果から，入力した文書画像が表紙であるか否かを識別し，表紙であると識別した場合には，そこで文書が切れたと判断し，そこを１文書の区切りとして文書登録を行う。

【００１２】

【発明の実施の形態】

図１は，本発明に係る電子ファイリングシステムのソフトウェアを含めた構成例，図２は，そのハードウェア構成例を示す。この電子ファイリングシステムは，図２に示すＣＰＵ２０およびメモリ２１からなる処理装置１と，ＡＤＦ付きのスキャナ等の複数ページ文書入力装置２と，磁気ディスク装置等の文書ファイル４を格納する文書格納装置３からなり，処理装置１は，複数ページ文書入力装置２から文書画像を入力する文書画像入力処理部１１，入力された文書画像が表紙か否かを識別する表紙識別処理部１２，表紙ページを区切りとして文書単位で入力された文書を文書格納装置３に登録・保存する文書登録処理部

１３とを備える。文書画像入力処理部１１，表紙識別処理部１２および文書登録処理部１３は，ソフトウェアプログラムによって構成される。

【００１３】

表紙識別処理部１２は，さらに文書画像入力処理部１１によって入力した文書画像中の文字列，図または表などの領域を識別する領域識別処理部１２１と，領域を識別した結果にもとづいて，領域に含まれる文字列のタイトルらしさや，領域のレイアウト情報を解析し，入力した文書画像の表紙らしさの評価得点を計算する評価得点計算処理部１２２と，評価得点を計算した結果にもとづいてその文書画像が表紙であるか否かを判定する表紙判定処理部１２３とを備える。

【００１４】

図３は，図１に示す電子ファイリングシステム全体の処理の流れを示すフローチャートである。

【００１５】

複数ページ文書入力装置２のスタッカに載置された複数の紙文書から１枚（１ページ）を読み込み，その文書画像（画像データ）をメモリに格納する（ステップＳ３１）。読み込んだページがスタッカ上の文書の最初のページであるかどうかを調べ（ステップＳ３２），読み込んだページが最初のページである場合には，それは必ず文書の表紙であるから，その表紙画像から文書名を獲得し，または日時等の情報から文書名を生成し，新規の文書として登録する（ステップＳ３５）。

【００１６】

さらにスタッカ上に文書がある場合には（ステップＳ３７），次の１ページを読み込み，その文書画像をメモリに格納する（ステップＳ３１）。読み込んだページがスタッカ上の文書の２ページ以降の場合には，最初のページではないので（ステップＳ３２），表紙識別処理部１２によって文書画像が表紙であるかどうかの識別処理を行う（ステップＳ３３）。文書画像が表紙であると識別された場合には（ステップＳ３４），表紙画像から文書名を獲得し，または日時等の情報から文書名を生成して，新規の文書として文書格納装置３に保存・登録する（ステップＳ３５）。一方，文書画像が表紙でないと識別された場合には，その文書画像を，最も最近に登録した文書に追加して登録・保存する（ステップＳ３６）。スタッカ上の文書が全てなくなるまで，ステップＳ３１～Ｓ３６の処理を繰り返す（ステップＳ３７）。

【００１７】

次に，文書画像が表紙であるか否かの識別方法について具体的に説明する。ここでは，タイトル抽出結果を使用する方法と文書画像の全体的なレイアウト情報を使用する方法とがある。これらの二つの方法を併用してもよい。

【００１８】

最初にタイトル抽出結果を使用する方法について説明する。一般的な社内文書を対象とした例で説明する。通常，社内文書は，一般の文書に比べて文書の表紙にタイトルが分かりやすく記載されている場合が多い。分かりやすいタイトルとは，例えば，

・他の文字より大きなサイズの文字で書かれている，
・他の文字列と離れて，見やすい位置に配置されている，
・下線や枠線が施されている，

というものである。これらの特性を利用して，読み込んだ文書画像にタイトルがあるか否かを判定し，タイトル有無の判定結果を用いて，その文書画像が表紙であるか否かの判定を行う。ここでは，特開平９－１３４４０６号「文書画像からのタイトル抽出方法」に開示する技術の一部である「タイトルらしさの計算技術」を用いる。

【００１９】

図４に，タイトルを用いた表紙識別処理の処理フローチャートを示す。まず，文書画像入力処理部１１により入力した文書画像から領域を識別して，文字列，図，表を抽出し（ステップＳ１１），それらの外接矩形およびその属性タグ（文字列／図／表）を検出する（ステップＳ１２）。

【００２０】
次に，文字列の矩形について，タイトルらしさの評価ポイントを算出する（ステップＳ１３）。具体的には，予め，文字列の大きさ（文書画像中の平均文字高さ／幅からみてどのくらい大きいか），文字列の修飾属性（下線付き，枠線付き），文字列の画像中での位置（縦方向および横方向），文字列間の相対的位置関係（周囲の文字列，図，表からどれくらい離れているか）などの種々の指標に具体的な得点を配分しておき，抽出した文字列の矩形のすべてについて，これらの指標が当てはまるか否かを独立に評価する。指標が当てはまる場合には，その矩形にタイトルらしさの得点を加算していく。文字列の矩形は，高ポイントほどタイトルらしいということになる。
【００２１】
次に，最もタイトルらしさの評価の高い文字列の矩形の評価ポイントが所定のしきい値以上であるかどうかを調べ（ステップＳ１４），その矩形の評価ポイントが所定のしきい値以上である場合には，その文字列の矩形をタイトルと判定し，その文書画像を表紙と識別する（ステップＳ１５）。
【００２２】
次に，表紙識別に，文書画像の全体的なレイアウト情報を使用する方法について説明する。表紙識別処理部１２は，例えば文書画像の各構成要素のレイアウト情報を用いて文書画像が表紙であるか否かを識別することもできる。この方法では，文書画像中の文字，図，表などの文書構成要素の絶対位置情報，または各文書構成要素間の相対的関係情報などから，その文書画像の表紙らしさの得点を求める。
【００２３】
具体的には，予め，文字，図，表等の文書構成要素の領域矩形の位置およびサイズ情報から，一つの領域矩形とその周囲にある領域矩形の間の相対的位置関係，サイズ関係および複数の矩形のそれぞれの絶対位置等で定まる表紙らしさの得点を定めておく。これを用いて，対象となる文書画像から求めた文字，図，表等のすべての文書構成要素について，その矩形の位置情報および形状の情報から，表紙らしさの得点を計算する。計算したすべての文書構成要素から，その文書画像の最終的な表紙らしさの得点を求め，その値が予め定めたしきい値以上の場合には，その文書画像を表紙として識別する。
【００２４】
図５に，文書構成要素のレイアウト情報を用いた表紙識別処理の処理フローチャートを示す。文書画像中の文字，図，表等の文書構成要素の領域を識別してその矩形領域を切り出し（ステップＳ２１），文書構成要素のレイアウト情報を抽出する（ステップＳ２２）。レイアウト情報は，例えば，以下のようなものである。
【００２５】
・領域矩形のサイズ（幅，高さ）
・領域矩形の位置（左上点の座標）
・領域の属性（文字／図形／表）
・内部の文字サイズ（文字領域の場合）
・最も近い上の領域情報，
・最も近い上の領域までの距離，
・最も近い下の領域情報，
・最も近い下の領域までの距離，
・最も近い左の領域情報，
・最も近い左の領域までの距離，
・最も近い右の領域情報，
・最も近い右の領域までの距離。
【００２６】
例えば，文書画像中の文書構成要素の配置が，図６に示すようなものである場合に，「図形ａ」を注目する領域矩形とすると，「図形ａ」の周囲の領域矩形への関係を表すデータ構造は，以下のようになる。「図形ａ」の左上角の座標を（ｘ１，ｙ１），右下角の座標

を（ｘ２，ｙ２）とする。
【００２７】
・サイズ：幅＝ｘ２－ｘ１＋１，高さ＝ｙ２－ｙ１＋１，
・位置：（ｘ１，ｙ１），
・属性：図形，
・内部の文字サイズ：０，
・最近の上の領域：文字ｂ，
・最近の上の領域への距離：ｄ１
・最近の下の領域：文字ｃ，
・最近の下の領域への距離：ｄ２
・最近の左の領域：表ａ，
・最近の左の領域への距離：ｄ３
・最近の右の領域：図形ｂ，
・最近の右の領域への距離：ｄ４
次に，これらの情報を表紙識別関数に入力して，その文書画像が表紙であるかどうかの評価得点を計算する（ステップＳ２３）。例えば，予め定めた以下のようなルールを適用する。
【００２８】
・文字の領域矩形のサイズ：しきい値より大きい→＋１０，
・文字の領域の内部の文字サイズ：しきい値より大きい→＋２０，
・文字の領域矩形の位置：中央付近→＋３０，
・文字の領域矩形の位置：上から順に，＋１５，＋１４，＋１３，…，
・文字の領域の上下左右の関係：左右の領域との距離がしきい値以上，上下の領域との距離がしきい値以上→＋３０，
・左右に同じ程度の上辺または下辺の文字の領域があり，両方ともしきい値より小さい文字→－３０，
このようなルールに従って表紙らしさの得点を計算し，その計算の結果，文書画像の得点が所定のしきい値以上のときは（ステップＳ２４），その文書画像を表紙と識別する（ステップＳ２５）。
【００２９】
以上が，文書構成要素のレイアウト情報を使用して表紙識別を行う際の表紙識別関数を実現する方法の一つであり，矩形間の相対的関係から予め定めた表紙らしさの得点を求める方法である。さらに，表紙識別関数を実現する方法として，遺伝的アルゴリズム（ＧＡ）を用いて得点を定める方法，ニューラルネットにより表紙を識別する方法，判別分析により表紙を識別する方法などを用いることもできる。これらを順番に説明する。
【００３０】
〔遺伝的アルゴリズムを用いて得点を求める方法〕
予め，文字列，図，表の文書構成要素の位置と形状の情報から，一つの領域矩形とその周囲にある領域矩形の間の相対的位置とサイズ関係，対象となる複数の領域矩形のそれぞれの絶対位置などで定まる表紙らしさの得点を定めておく方法として，学習用の表紙画像を多量に用意しておき，遺伝的アルゴリズムを使用して，得点配分を表紙識別に最も適合するように学習する方法を用いる。
【００３１】
具体的には，注目する一つの領域矩形，その周辺の領域矩形の間の関係，領域矩形と上下左右の画像端との関係，領域矩形中心の位置座標およびその画像端との関係などの特徴を記述し，それらの特徴に対してランダムな得点を割り振った一つの規則を作成する。例えば，以下のような構造を持つ一つの遺伝子を作り，各項目にランダムに得点を配分しておく。
【００３２】
・文字領域のサイズ：しきい値より大きい，

・文字領域の内部の文字サイズ：しきい値th１より大きい，
・文字領域の内部の文字サイズ：しきい値th２より大きい，
・文字領域の位置：中央付近，
・文字領域の位置：上の方，
・文字領域の上下左右の関係：左右距離がしきい値以上，上下距離がしきい値以上，
・左右に同じ程度の上辺または下辺の文字領域があり，両方ともしきい値より小さい文字
，
・図形領域のサイズ：しきい値より大きい，
・図形領域の位置：中央付近，
・図形領域の位置：上の方，
・図形領域の上下左右の関係：左右距離がしきい値以上，上下距離がしきい値以上，
・表領域のサイズ：しきい値より大きい
・表領域の位置：中央付近，
・表領域の位置：上の方，
・表領域の上下左右の関係：左右距離がしきい値以上，上下距離がしきい値以上。

【００３３】

同様な遺伝子を多数独立に生成し，以下のような処理を行う。
１）多数の学習用表紙画像に適用して評価する。
２）しきい値より高い得点を示す複数の遺伝子を残し，他のものは捨てる。
３）選択された遺伝子を交差させ，または遺伝子に突然変異を加えて，元の数と同数の遺伝子を生成する。
４）１）の処理を繰り返し，これを全ての評価画像に対して高得点を与えるような遺伝子を一つ見つけるまで繰り返す。

【００３４】

処理を何回か繰り返すと，規則はだいたい高い表紙らしさの得点を与えるものだけになってくる。処理を適当な回数繰り返した後に，全ての評価画像に対して高得点を得る遺伝子にある得点配分（最も高い得点を与える規則）を抽出し，それを表紙らしさの識別用規則として用いる。

【００３５】

表紙であるか否かの識別は，上記遺伝的アルゴリズムにより抽出した規則を文書画像に適用して出力された得点と，所定のしきい値との比較によって行い，得点がしきい値以上の場合には，表紙と識別する。

【００３６】

〔ニューラルネットにより表紙を識別する方法〕
また，別の表紙らしさの得点計算処理（ステップＳ２３）として，ニューラルネットによる表紙識別方法を用いてもよい。この方法では，まず学習用の表紙画像を多量に用意しておき，それらの文字列，図，表の文書構成要素の領域矩形の位置と形状の情報を表紙識別用のニューラルネット（ＮＮ）に入力し，誤差逆伝搬法などでニューラルネットの重みを学習させ，表紙識別ニューラルネットを構成しておく。図７に，このニューラルネットの例を示す。

【００３７】

入力した文書画像から求められた文字列，図，表の文書構成要素の領域矩形の位置と形状の情報を，表紙識別ニューラルネットにおける入力層の各ノード（図中，丸で示す）から入力し，表紙か否かの識別を行う。具体的には，上記の遺伝的アルゴリズムを用いた処理で説明した遺伝子の項目と同様の数値を並べてベクトルを生成し，３層のバックプロパゲーション（誤差逆伝搬）型のニューラルネットを構成する。入力層はベクトルの次元数のノードを持ち，中間層はベクトルの次元数＊２のような適当な数のノードを持ち，出力層は，表紙であるか表紙でないかをそれぞれ出力する二つのノードを持つ。

【００３８】

このニューラルネットを用いる方法では，以下のような処理を行う。

１）学習用の表紙画像からベクトルを生成する。
２）出力に表紙ベクトル信号を与えて，ネットの重みを誤差逆伝搬法で変更する。
３）上記１），２）の処理を評価増がなくなるまで繰り返す。
４）できあがった表紙識別ニューラルネットを用いて，表紙識別を行う。
【００３９】
図７に示すニューラルネットは，ハードウェアによって実現することも，またソフトウェア・プログラムによって実現することも可能である。
【００４０】
〔判別分析により表紙を識別する方法〕
さらに，別の表紙らしさの得点計算処理（ステップＳ２３）として，判別分析による表紙識別方法を用いてもよい。この方法では，学習用の表紙画像を多量に用意しておき，それらの画像の文字列，図，表の文書構成要素の領域矩形の位置と形状の情報から，線形判別式またはベイズ識別のような表紙識別用の判別関数を構成する。
【００４１】
入力した文書画像から求めた文字列，図，表の領域矩形の位置と形状の情報を，この表紙識別用判別関数に入力し，表紙か否かの識別を行う。
【００４２】
この判別分析により表紙を識別する方法では，以下のような処理を行う。
１）複数の表紙画像（集合ｓ１）から求めた構成要素の特徴をベクトルに表現する。
２）複数の表紙以外の画像（集合ｓ２）からも，同様にベクトルを求める。
３）各集合で，平均μ１，μ２を求める。
４）共分散行列Σを求める。
５）ある文書画像が入力された場合に，その文書画像の特徴ベクトルをｘとする。次の線形判別式ｚ
$z = x' \Sigma^{-1} (\mu_1 - \mu_2) - (1/2) (\mu_1 + \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$
が正であれば表紙画像と判別し，それ以外であれば表紙でないと判別する。
【００４３】
以上のように，本実施の形態によれば，電子ファイリングシステムにおいて，自動的に表紙文書を識別し，表紙ページで登録文書を区分して，別文書として登録することが可能になる。
【００４４】
また，文書画像処理において，文書画像を領域識別し，文字列，図，表の文書構成要素を抽出し，これらの情報を使用して表紙であるか否かを識別することができる。この文書画像処理においては，文書画像を領域識別し，文字列，図，表の文書構成要素を抽出した後，例えばタイトル抽出処理を行い，タイトルがある文書画像を表紙画像であると判断する。この場合，抽出した領域中の文字列についてタイトルらしさの得点を計算し，各領域において最も高得点のタイトル得点と予め設定したしきい値とを比較し，タイトル得点がしきい値以上ならば，タイトルであると判断する。
【００４５】
また，上記の文書画像処理において，予め学習用の表紙画像から文字列，図，表などの文書構成要素を求め，一つの矩形（領域）とその周囲矩形の間の相対的位置関係とサイズ関係と対象となる複数の矩形のそれぞれの絶対位置と矩形の属性（タグ）で決まる表紙らしさの得点を定めておき，入力された文書画像から求めた文字列，図，表の位置および形状情報から，表紙らしさの得点を計算する。
【００４６】
ここで，予め学習用の表紙画像からなる文書構成要素を求め，一つの矩形（領域）とその周囲矩形の間の相対的位置関係とサイズ関係と対象となる複数の矩形のそれぞれの絶対位置と矩形の属性（タグ）で決まる表紙らしさの得点を定める際に，遺伝的アルゴリズム（ＧＡ）を用いて得点配分を表し識別に最も適合するように学習する。
【００４７】

または，予め学習用の表紙画像から文書構成要素を求め，それらの位置と矩形情報と矩形の属性（タグ）情報を，表紙識別用のニューラルネットに入力し，誤差逆伝搬法などにより各ノードの重みを学習させ，表紙を識別できるニューラルネットを構成しておき，入力画像から求められた文字列，図，表の文書構成要素の位置と矩形情報と矩形の属性（タグ）情報を表紙識別ニューラルネットに入力して，表紙識別を実現することもできる。

【００４８】

または，予め学習用の表紙画像から文書構成要素を求め，それらの位置と矩形情報と矩形の属性（タグ）情報から，線形またはベイズ識別のような表紙識別用の判別関数を構成し，入力画像から求められた文字列，図，表の文書構成要素の位置と矩形情報と矩形の属性（タグ）情報を表紙識別用判別関数に認識して，表紙であるか否かの判別を行うことができる。

【００４９】

【発明の効果】

以上説明したように，本発明によれば，文書画像の文字列，図，表などの文書構成要素の位置およびその属性情報からタイトルの有無を判定して表紙を識別し，または文書構成要素のレイアウト情報を用いて表紙らしさの評価値を計算して表紙を識別する。これにより，電子ファイリングシステムにおいて，複数文書のページを連続して入力しても，文書画像の表紙を自動的に識別して文書ごとに登録・保存を行うことができる。これにより，ユーザは，文書単位を意識することなく入力操作を行うことができ，入力操作に関わる労力を大幅に軽減することができる。

【図面の簡単な説明】

【図１】電子ファイリングシステムの構成例を示す図である。

【図２】電子ファイリングシステムのハードウェア構成例を示す図である。

【図３】電子ファイリングシステム全体の処理の流れを示すフローチャートである。

【図４】タイトルを用いた表紙識別処理の処理フローチャートである。

【図５】文書構成要素のレイアウト情報を用いた表紙識別処理の処理フローチャートである。

【図６】文書構成要素の配置例を示す図である。

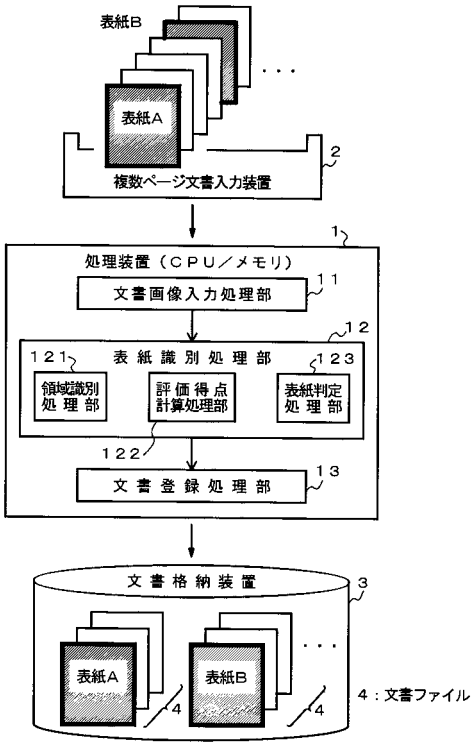
【図７】表紙識別処理に用いるニューラルネットの例を示す図である。

【符号の説明】

１　処理装置
１１　文書画像入力処理部
１２　表紙識別処理部
１２１　領域識別処理部
１２２　評価得点計算処理部
１２３　表紙判定処理部
１３　文書登録処理部
２　複数ページ文書入力装置
３　文書格納装置
４　文書ファイル

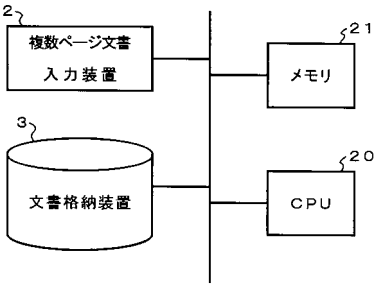
【図１】

電子ファイリングシステムの構成例



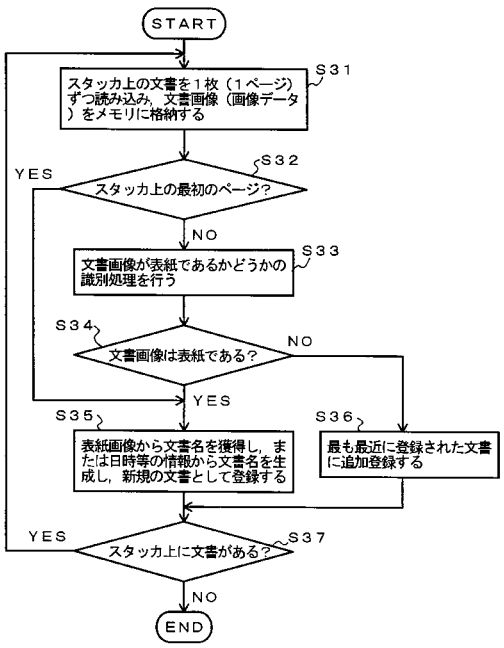
【図２】

電子ファイリングシステムのハードウェア構成例



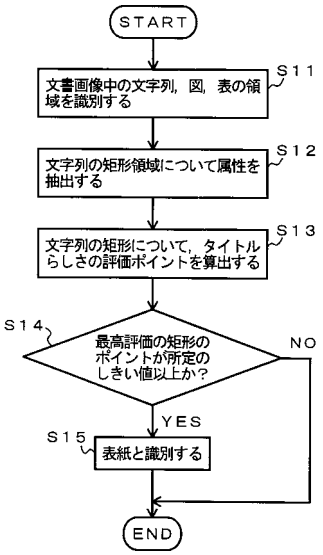
【図３】

電子ファイリングシステムの全体処理の処理フローチャート



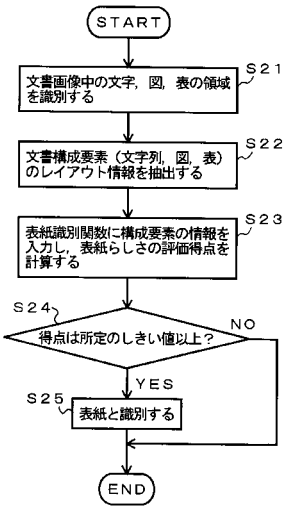
【図４】

タイトルを用いた表紙識別処理の処理フローチャート



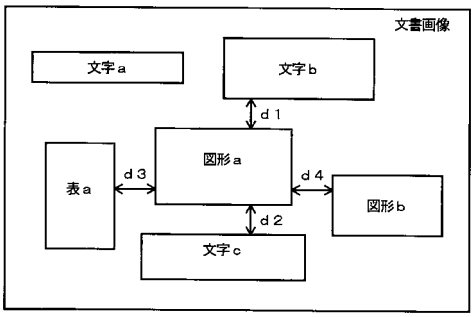
【図５】

文書構成要素のレイアウト情報を用いた表紙識別処理の処理フローチャート



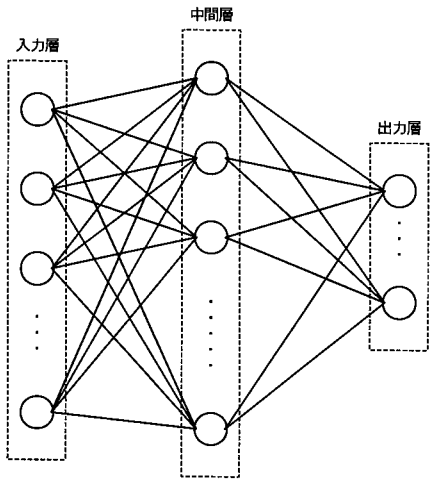
【図６】

文書構成要素の配置例



【図７】

ニューラルネットの例



フロントページの続き

(72)発明者　広田　勉
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

審査官　橋爪　正樹

(56)参考文献　特開平１０－０９７６０６（ＪＰ，Ａ）
特開平０７－０８５０８３（ＪＰ，Ａ）
特開平０９－１３４４０６（ＪＰ，Ａ）
特開平１０－０１１５３１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06T 1/00
G06K 9/00- 9/82
G06F17/30