(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3689485号
(P3689485)
(45)発行日　平成17年8月31日(2005.8.31)　(24)登録日　平成17年6月17日(2005.6.17)

(51)Int.Cl.[7]　F I
G06K 9/20　G06K 9/20 340C
G06K 9/20 340L

請求項の数 1　(全 15 頁)

(21)出願番号　特願平8-114573
(22)出願日　平成8年5月9日(1996.5.9)
(65)公開番号　特開平9-305701
(43)公開日　平成9年11月28日(1997.11.28)
審査請求日　平成12年11月27日(2000.11.27)
審判番号　不服2003-19217(P2003-19217/J1)
審判請求日　平成15年10月2日(2003.10.2)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100075096
弁理士　作田　康夫
(72)発明者　新庄　広
東京都国分寺市東恋ケ窪１丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　嶋　好博
東京都国分寺市東恋ケ窪１丁目２８０番地
株式会社日立製作所中央研究所内

最終頁に続く

(54)【発明の名称】帳票認識方法

(57)【特許請求の範囲】
【請求項１】
帳票の表面画像を入力し文字を読み取る帳票認識方法において，
画像入力部において、帳票の画像を二値化して入力し、
枠抽出部において、前記帳票の画像から黒画素の連続状態に基づいて検出される罫線のうちの枠罫線を抽出して、枠罫線を構成しない罫線を特定し，
下線抽出部において、当該枠罫線を構成しない罫線から下線を抽出し，当該下線の延長線上に存在する黒画素を検出し，検出した黒画素の並びを下線であると判定することを特徴とする帳票認識方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は帳票、特に、不動産に関する登記情報が記載された多様な帳票に関し，特に，登記済通知書から文字データを読み取り，自動的に入力する帳票認識方法に関する。
【０００２】
【従来の技術】
帳票の種類の識別に関する従来技術の例としては，以下のものが挙げられる。第１は，全ての種類の帳票に対して同じ位置に記載された帳票の種類を表すＩＤ番号を読み取ることにより，帳票の種類を識別する方式である。第２は，帳票の種類ごとに枠の構造が異なる場合に，枠の構造を識別することにより帳票の種類を識別する方式である。この例は，特

開平７－１４１４６２号公報に記載されている。
【０００３】
【発明が解決しようとする課題】
不動産に関する登記済通知書は現在７種類ある。これらの帳票は不動産に関する課税のためのデータ入力に用いられるものであるが、この通知書には，帳票の種類を特定するＩＤ番号の記載がないため，ＩＤ番号読み取りにより帳票を識別する従来手法を用いることはできない。さらに，これらの帳票は，同じ種類であっても枠の形状が異なる非定型帳票であるため，枠の構造から帳票を識別する従来手法を用いることはできない。また，表題部の文字を読み取ることにより帳票を識別する従来手法を用いる場合には，帳票の識別精度は文字認識の精度に大きく依存するという問題がある。登記済通知書の帳票名は，「権利に関する土地登記済通知書」，「権利に関する建物登記済通知書（一般）」，「権利に関する建物登記済通知書（専有）」，「表示に関する土地登記済通知書」，「表示に関する建物登記済通知書（一般）」，「表示に関する建物登記済通知書（一棟）」，「表示に関する建物登記済通知書（専有）」の７種類である。このうち，「表示に関する建物登記済通知書（一般）」と「表示に関する建物登記済通知書（一棟）」は，一字しか違わないため，この二種類に対する識別精度が低くなる可能性がある。
【０００４】
そこで，本発明の第１の目的は，帳票の種類が多様な読み取り対象に対して，高精度な帳票識別手段を有する帳票認識手段を提案することである。
【０００５】
従来の下線検出方法では，枠線以外の罫線を下線としていたため，文字の横方向のストローク等のノイズ成分を下線として誤抽出する可能性があった。そこで，本発明の第２の目的は，高精度な下線検出手段を有する帳票認識手段を提案することである。
【０００６】
【課題を解決するための手段】
第１の観点では、この発明は、登記済通知書の表面画像を入力し文字を読み取る登記情報の認識方法であって，登記済通知書の画像から文字行を抽出する文字行抽出手段と，抽出した複数の文字行と枠との位置関係から帳票名の文字行を選択し文字行選択手段と，帳票名の文字行を読み取る文字識別手段から，登記済通知書の種類を識別する第１の帳票識別手段と，登記済通知書の画像から罫線を抽出する罫線抽出手段と，抽出した罫線から表の特徴を抽出する表特徴抽出手段と，表の特徴から登記済通知書の種類を識別する第２の帳票識別手段と，登記済通知書の画像から文字行を抽出する文字行抽出手段と，抽出した文字行を読み取る文字識別手段と，読み取り結果の中から帳票の項目名を選択する項目名選択手段と，項目名の組み合わせから登記済通知書の種類を識別する第３の帳票識別手段とを具備し，当該３つの手段の結果を組み合わせることにより，登記済通知書の種類を識別する帳票認識方法を提供する。
【０００７】
第２の観点では、この発明は、登記済通知書の表面画像を入力し文字を読み取る登記情報の認識方法であって，登記済通知書の画像から文字行と罫線を抽出する文字行抽出手段と，抽出した罫線から枠罫線と枠罫線でない罫線を区別する罫線種判定手段と，枠罫線でない罫線が含まれる枠内の文字行を検出する手段と，当該枠内の文字行と当該枠内の枠罫線でない罫線との位置関係から，当該枠内の枠罫線でない罫線が下線か否かを判定する下線検出手段を具備する帳票認識方法を提供する。
【０００８】
【発明の実施の形態】
以下、本発明の一実施例を詳細に説明する。なお、これにより本発明が限定されるものではない。
【０００９】
図１は、本発明の一実施例である登記情報システムの構成図である。登記情報の認識を行う認識部１０１と認識結果の修正を行う修正部１０５がネットワーク１０４により接続さ

れており，入力センタ１１１において認識と修正を並行して行うことができる。処理の過程は，まずスキャナ１０２により登記済通知書１００の画像を入力する。次に，認識用計算機１０３では，文字および罫線の認識を行い，修正用計算機１０６において認識結果の修正確認を行う。また，辞書やコード表と照合チェックし，コードデータを出力する。認識結果は，通信制御用計算機１０７を介して，遠隔地にある計算センタ１１０にあるホスト計算機１０８に接続された登記情報データベース１０９に格納される。修正部１０５では，認識結果の一部を利用し，登記情報データベース１０９をアクセスし，登録済の登記情報を読み出す。当該読み出した登録情報と認識結果の一部を照合し，矛盾がないかどうかの検定を行う。

【００１０】

図２は，登記情報認識の処理過程を示すブロック図である。認識部１０１では，帳票画像を読み取り，修正部１０５に縮小画像２４８，枠座標２５０，下線座標２５２，文字行座標２５４，帳票種類２５６，認識結果ラティス２５８，文字座標２６０を出力する。修正部１０５では，これらの入力データをもとに，操作者が認識結果を修正する。画像入力部２００では，帳票表面の画像を白黒２値化して入力する。

【００１１】

入力した画像は，画像縮小部２０２と文字行画像抽出部２１８に出力される。画像縮小部２０２では，後続の処理の高速化のため帳票画像を縮小し，縮小画像２４８を出力する。縮小処理は，細い罫線が縮小後かすれないよう，画素ごとのＯＲ処理を行う。縮小した画像に対し，罫線抽出部２０４において実線と点線の罫線を抽出する。実線は，黒画素の連続するつながりをもとに抽出される。点線は，黒画素の連結成分の外接矩形の配置，サイズの拘束条件をもとに抽出される。枠抽出部２０６では，２０４で抽出した罫線から罫線が四方を取り囲む枠を求め，枠の頂点座標２５０を出力する。表特徴抽出部２０８では，２０６で抽出された枠の情報から，枠の集まりである表の特徴量を抽出する。この特徴量とは，縦横の罫線の本数や，罫線同士の接続関係，枠の位置関係等である。

【００１２】

一方，文字行抽出部２１２では，２０２から出力された縮小画像から，文字の集合である文字行を抽出する。ここでは，黒画素の連結成分うち，文字と推定される大きさの連結成分の外接矩形の頂点座標をもとに，文字の並びと推定される外接矩形を融合することにより，文字行を生成する。行―枠対応部２１４では，２１２で抽出した文字行の頂点座標と２０６で抽出した枠の頂点座標を比較することにより，各文字行がどの枠内に存在するか，もしくは枠外にあるかを判定し，枠ごとに含まれる文字行の頂点座標と枠外の文字行の頂点座標２５４を出力する。　また，下線抽出部２１６では，２０４で抽出した罫線座標と，２０６で抽出した枠の頂点座標と，２１４で抽出した枠内の文字行座標とをもとに，下線を抽出して，下線の座標２５２を出力する。さらに，文字行画像抽出部２１８では，２１４で抽出された文字行座標をもとに，２００で入力された画像から文字行部分の画像を切り出す。文字切り出し・文字識別部２２０では，文字切り出し部２２２と文字識別部２２４が協調して，文字を１文字ずつ切り出し，その文字座標２６０を出力する。さらに，文字識別部２２４では，切り出した１文字分の画像パターンに対して，識別辞書２２６を用いて文字を識別する。帳票名照合部２２８では，文字識別部２２４の出力である文字識別結果を入力し，単語照合部２３０により帳票名辞書２３２に格納された帳票名単語と照合することにより帳票名についての認識結果の誤りを修正して帳票名を求める。

【００１３】

帳票名辞書２３２に格納された単語は，認識対象の帳票名である。認識対象の帳票名はあらかじめわかっており，帳票名は帳票の種類に１対１に対応する。さらに，項目照合部２３４では，２２８で照合されなかった文字認識結果を入力し，単語照合部２３６により項目辞書２３８に格納された項目名単語と照合することにより項目名についての認識結果の誤りを修正して項目名を求める。項目辞書２３８にされた単語は，認識対象の帳票内に記載された項目である。内容照合部２４０では，２３４で照合されなかった文字認識結果を入力し，単語照合部２４２により内容辞書２４４に格納された内容単語と照合することに

より内容についての認識結果の誤りを修正する。ここで，「内容」とは帳票において，項目名に対して記載されている内容をさす。例えば，「地目」という項目に対する内容には「居宅」や「公園」などがある。内容辞書２４４に格納された単語は，認識対象の帳票内に記載された内容を記載する単語のうち，あらかじめ使用が決められている単語である。２４０の処理の結果出力される認識結果ラティス２５８は，１文字ごとに文字識別処理の結果である候補文字を類似度が高い順に並べたものである。この文字識別結果は，帳票名照合，項目照合，内容照合により誤りを修正してある。

【００１４】

一方，帳票識別部２４６では，表特徴抽出部２０８と帳票名照合部２２８と項目名照合部２３４の出力結果を入力し，表特徴と帳票名，項目名から帳票の種類を識別し，帳票種類２５６を出力する。

【００１５】

図３は、図２で示した登記情報認識の処理フローを示す図である。ステップ３００で画像を入力し，ステップ３０２で当該画像を縮小する。次いで，ステップ３０４で画像から罫線を抽出し，ステップ３０６で罫線から枠を抽出する。さらに，ステップ３０８で表の特徴を抽出する。また，ステップ３１０で当該縮小画像から文字行を抽出し，ステップ３１２で，抽出した行と枠とを対応付ける。また，ステップ３１４で，罫線と枠と文字行の座標から下線を抽出する。さらに，ステップ３１６で，文字行の座標値に基づいて帳票画像から文字行部分の画像のみを抽出する。ステップ３１８で，当該文字行画像を１文字ずつの画像に分割し，ステップ３２０で切り出された画像パターンに対して文字識別を実行する。ステップ３２２では，文字識別結果を帳票名の単語と照合して帳票名を識別する。ステップ３２４では，文字識別結果を項目名の単語と照合して項目名を識別する。ステップ３２６では，文字識別結果を内容の単語と照合して内容を識別する。ステップ３２８では，ステップ３０８の処理結果である表の特徴とステップ３２２の処理結果である帳票名とステップ３２４の処理結果である項目名から帳票の種類を識別する。ステップ３３０では，３００から３２８の処理で得た結果を出力する。

【００１６】

図４は，認識対象である登記済通知書の画像を，説明のために簡略的に示した図である。帳票画像４００の例では，帳票名「権利に関する建物登記済通知書（専有）」４０１が記載されており。横罫線４０２，４０４，４０６，４０８と縦罫線４１０，４１２，４１４，４１６が印刷されている。また，項目として「符号」４１８と「所在」４２０，「地目」４２２がある。「符号」の内容としては「１」（４２４）と「２」（４２６），「所在」の内容としては４２８と４３０に「国分寺市東恋ヶ窪１丁目２８０番地」が記載されている。「地目」の内容としては，「宅地」（４３２）と「公園」（４３４）が記載されている。さらに，内容４２４「１」，４２８「国分寺市東恋ヶ窪１丁目２８０番地」，４３２「宅地」には，それぞれ下線４３６，４３８，４４０が印刷されている。

【００１７】

図５は，図４の帳票画像に対する，図３のステップ３０４の罫線抽出処理結果を示すものである。（ａ）の５００は横罫線の抽出結果であり，（ｂ）の５２０は縦罫線の抽出結果である。（ａ）では，図４の横罫線４０２から４０８に相当する罫線として，それぞれ，５０２から５０８が抽出されている。下線４３６，４３８，４４０に相当する下線として，それぞれ，５１０，５１４，５１６が抽出されている。５１２と５１８は，「市東恋」の横ストロークをつなげることによって，罫線として抽出したものである。この離れた横ストロークが接続される現象は，横罫線を抽出する際に黒画素を横方向に収縮・膨張処理することにより，接近した黒画素が接続されることに起因する。また，（ｂ）では，図４の縦罫線４１０から４１６に相当する罫線として，それぞれ，５２２から５２８が抽出されている。

【００１８】

図６は，図４の帳票画像に対する，図３のステップ３０６の枠抽出処理結果を示すものである。６００は枠抽出結果である。６０２から６１８の９個の枠が抽出されている。

【００１９】
図７は，図４の帳票画像に対する，図３のステップ３１０の文字行抽出処理結果を示すものである。７００は文字行抽出結果である。図４の文字行４０１，４１８，４２０，４２２，４２４から４３４の文字行に対して，それぞれ７０２から７２０の文字行の外接矩形が抽出されている。
【００２０】
図８は，図３のステップ３１４の下線抽出処理に関する処理フローである。罫線抽出処理３０４，枠抽出処理３０６，文字行抽出処理３１０の結果を用いて
，ステップ８００では，枠を構成しない罫線を抽出する。ステップ８０２では，ステップ８００で抽出した罫線の本数分だけ，以下の処理を繰り返す。ステップ８０４では，文字行の座標と罫線の座標を比較する。比較の方法については図９と図１０を用いて説明する。ステップ８０６では，比較した値が基準を満たすか否かを判定する。基準値を満たす場合，ステップ８０８で，比較対象の罫線を下線とする。なお，上記ステップ８０８において抽出された２本の下線について，端点同士がが微小な間隔で離れており，延長線上に存在する場合には，１本の下線であるとすることもできる。また，上記ステップ８０８において抽出した下線の長さが基準値以下であれば，下線とみなさないとすることもできる。
【００２１】
図９は，図８の処理フローを説明するための帳票の枠の例である。横罫線９００と９０２，縦罫線９０４と９０６，文字行９０８，下線９１０が印刷されている。
【００２２】
図１０は，図９の例から罫線と文字行を抽出した結果である。この図を用いて下線の判定を説明する。下線判定処理は，文字行と同一枠内にある罫線の中で，文字行の下に位置し，文字行とほぼ同じ長さの罫線を下線と判定する。図１０において，１００７は文字が印刷されていた領域であり，１００８は１００７の外接矩形である。図９の９００から９１０の罫線は，それぞれ１０００から１０１０として抽出されている。さらに，１０１２は文字の横ストロークを罫線として抽出したものである。抽出された罫線の中から，枠を構成していない罫線として，１０１０と１０１２が抽出される。以下，１０１０を例として下線と判定される場合について説明し，１０１２を例として下線と判定されない場合を説明する。
【００２３】
図１０の１０１０について判定する。まず，罫線の下端のｙ座標と文字行の下端のｙ座標との差ｄ１１（１０１４）を求める。次に，罫線の上端のｙ座標と文字行の上端のｙ座標との差ｄ１２（１０１６）を求める。さらに，罫線のｘ方向の長さＬ１（１０１８）と文字行のｘ方向の長さＬｃ（１０２０）との差を求める。この値を基準値，α，β，γ１，γ２と比較する。ｄ１１が文字行より下でα未満であり，ｄ１２がβ以上であり，Ｌ１－Ｌｃがγ１以上γ２以下であれば，この罫線を下線とする。上記の処理の判定基準であるα，β，γ１，γ２の値は経験的に求めることができる。
【００２４】
例えば，αは，文字行と下線との間隔が一定であればその値を用いることができる。一定でなければ，枠の高さと文字の高さの差の１／２を用いることができる。βは，文字行の下端と下線との間隔と，文字の高さとが一定であれば，この２つの値の和を用いることができる。γ１とγ２の値は，一文字程度のマージンを見込んで，γ１は文字幅に（－１）をかけた値，γ２は文字幅等を用いることができる。上記のα，β，γ１，γ２の値の設定にあたっては，帳票の傾きや，線のかすれやつぶれ等に対して頑健性をもたせるため，マージンをもたせて値を設定することができる。また，ｄ１１の値の許容値について，負の値を許容すれば，下線が文字と重なる場合にも対応できる。
【００２５】
次に，図１０の１０１２について判定する。まず，罫線の下端のｙ座標と文字行の下端のｙ座標との差ｄ２１（１０２２）を求める。次に，罫線の上端のｙ座標と文字行の上端のｙ座標との差ｄ２２（１０２４）を求める。さらに，罫線のｘ方向の長さＬ２（１０２６

）と文字行のｘ方向の長さＬｃ（１０２０）との差を求める。これらの値を上記α，β，γ１，γ２と比較した場合，ｄ２１は負の大きな値となり，ｄ２２はβより小さな値になるため，下線ではないと判定される。

【００２６】

なお，ここで用いたｄ１１，ｄ１２は文字の高さや枠の高さ等で正規化してもよい。また，Ｌ１とＬｃの差の代わりに比を比較してもよい。α，β，γ１，γ２の値は，比較対象の定義に合わせて設定する。

【００２７】

また，ここでは，罫線の下端のｙ座標と文字行の下端のｙ座標との差１０１４と，罫線の上端のｙ座標と文字行の上端のｙ座標との差１０１６，罫線のｘ方向の長さ（１０１８）と文字行のｘ方向の長さ（１０２０）との差の３つの評価値を用いたが，必要に応じてこの中の１つもしくは２つのみを用いていもよい。

【００２８】

図１１は，図３のステップ３１４下線抽出処理において，文字行の座標の代わりに文字の座標を用いた例である。図１０で説明した判定基準を用いて，枠を構成しない罫線１１０８と文字の外接矩形１１１２を比較することにより，１１０８は下線であると判定できる。また，枠を構成しない罫線１１１０と文字の外接矩形１１１４を比較することにより，１１１０は下線でないと判定できる。

【００２９】

図１２は，文字行内の一部の文字に対してのみ下線が印刷されている例である。枠１２００内に，文字行１２０２と下線１２０４が記載されている。図１１の方法を用いれば，文字行中の「１丁目２８０番」の文字のみに下線が印刷されていることを判定できる。

【００３０】

図１３は，図３のステップ３１４の下線抽出処理に関する別の処理フローである。登記済通知書では，図４の４３６，４３８，４４０のように同一線上に複数の下線が存在することが多い。一方，下線４３６は短いので，文字内の横方向のストロークと長さが変わらないため，罫線抽出の際に抽出もれする可能性がある。この処理では，罫線抽出の際に抽出もれする可能性のある短い下線を正しく抽出することを目的とする。このため，まず長い下線を抽出し，この下線の延長上にある罫線を下線と判定する。

【００３１】

以下，図１３の各ステップについて説明する。ステップ１３００では，長い下線のみを抽出する。この処理は，図８で示した処理等を用いて実現できる。ステップ１３０２では，横方向のランレングスデータのうち枠線を構成しないランレングスデータを抽出する。ステップ１３０４では，抽出したランレングスデータの個数分についてステップ１３０６と１３０８の処理を繰り返す。ステップ１３０６では，対象とするランレングスデータが下線の延長線上にあるか否かを判定する。延長線上にあれば，ステップ１３０８で下線を構成するランレングスデータであるとして抽出する。ステップ１３１０では，ステップ１３０８で下線を構成すると判定されたランレングスデータから構成される罫線を下線として抽出する。なお，上記ステップ１３１０において抽出された２本の下線について，端点同士がが微小な間隔で離れており，延長線上に存在する場合には，１本の下線であるとすることもできる。また，上記ステップ１３１０において抽出した下線の長さが基準値以下であれば，下線とみなさないとすることもできる。

【００３２】

図１４は，図１３の処理フローを説明するための帳票の枠の例である。横罫線１４００と１４０２，縦罫線１４０４から１４１０，下線１４１２から１４１６，文字行１４１８から１４２２が印刷されている。

【００３３】

図１５は，図１４の画像から枠を構成しない横方向のランレングスデータと長い下線とを抽出した結果である。１５００は図１３のステップ１３００で抽出された長い下線である。横方向のランレングスデータの連結成分のうち，１５０２と１５０４は１５００の延長

線上１５０８から許容範囲ｗ（１５１０）以内にあるので，下線であると判定する。１５０６はｗよりも外にあるので，下線はないと判定する。

【００３４】

図１６は，図３のステップ３１４の文字行抽出処理に関する別の処理フローである。この処理では，枠を構成しない横方向のランレングスデータの長さの値をランの中点から傾き方向に投影して作成したヒストグラムを用いて下線を抽出する。以下，図１６の各ステップにてついて説明する。ステップ１６００では，横方向のランレングスデータのうち枠線を構成しないランレングスデータを抽出する。ステップ１６０２では，抽出したランレングスデータの長さの値を，ランの中点から傾き方向に投影してヒストグラムを作成する。ステップ１６０４では，ヒストグラムの山の数だけステップ１６０６とステップ１６０８の処理を繰り返す。ステップ１６０６では，投影値が基準値以上であるか否かを判定する。基準値以上であれば，ステップ１６０８で投影されたランレングスデータは下線を構成すると判定する。ステップ１６１０では，ステップ１６０８で下線を構成すると判定されたランレングスデータから下線を抽出する。なお，上記ステップ１６１０において抽出された２本の下線について，端点同士がが微小な間隔で離れており，延長線上に存在する場合には，１本の下線であるとすることもできる。また，上記ステップ１６１０において抽出した下線の長さが基準値以下であれば，下線とみなさないとすることもできる。

【００３５】

図１７は，図１４の画像から枠を構成しない横方向のランレングスデータを抽出し，ヒストグラムを作成した結果である。１７００から１７０６は図１６のステップ１６００で抽出された横方向のランレングスデータの連結成分である。ヒストグラム１７０８と１７１０は，ステップ１６０２で投影された結果である。ステップ１６０６において，１７０８と１７１０について，許容範囲ｗ（１７１２）の範囲内の面積を基準値と比較する。この場合，１７０８は基準値以上，１７１０は基準値未満であるとすると，１７００，１７０２，１７０４は下線であり，１７０６は下線ではないと判定できる。

【００３６】

図１８は，図３のステップ３２８の帳票識別処理に関する処理フローである。ステップ３０８では表の特徴量を抽出する。ステップ３２２では帳票名の単語照合結果を求める。ステップ３２４では項目名の単語照合結果を求める。ステップ１８００では，３０８，３２２，３２４の結果からそれぞれ導出される帳票の種類を用いて，多数決により帳票種類を識別する。

【００３７】

ステップ３０８で抽出する表の特徴としては，罫線の接続関係，枠の個数，枠の配置関係，縦罫線の本数，横罫線の本数等がある。罫線の接続関係が帳票の種類ごとに異なる場合には，特開平７－１４１４６２号公報に記載されている技術を用いて帳票の種類を特定できる。

【００３８】

【表１】

表１

帳票名	縦罫線（実線）の本数
権利に関する土地登記済通知書	１０
権利に関する建物登記済通知書（一般）	１０
権利に関する建物登記済通知書（専有）	１１もしくは１６
表示に関する土地登記済通知書	8
表示に関する建物登記済通知書（一般）	8もしくは１０
表示に関する建物登記済通知書（一棟）	7
表示に関する建物登記済通知書（専有）	8もしくは１２

【００３９】
表１では，ステップ３０８で抽出する表の特徴の例として，認識対象である登記済通知書の縦の実線罫線の本数を示している。これにより，縦の実線罫線は７，８，１０，１１，１２，１６本のうちのいずれかでることがわかる。このうち，８本と１０本の場合を除けば，帳票の種類が一意に決定する。８本と１０本の場合も帳票種類の候補を挙げることができる。
【００４０】
また，ステップ３２２で照合する帳票名の単語は，帳票名全てを一つの単語として登録してもよく，「権利」「表示」，「建物」「土地」，「一般」，「専有」，「一棟」など特徴的な単語のみを登録してもよい。
【００４１】
【表２】

表２

帳票名	所在	地積	床面積	一棟の建物番号	主	所	棟	表
権利に関する土地登記済通知書	○	○						
権利に関する建物登記済通知書（一般）	○		○					
権利に関する建物登記済通知書（専有）	○		○	○				
表示に関する土地登記済通知書						○		○
表示に関する建物登記済通知書（一般）					○	○		
表示に関する建物登記済通知書（一棟）							○	
表示に関する建物登記済通知書（専有）					○	○		

【００４２】
表２は，ステップ３０８で照合する項目名の中から一部を抜粋して示したものである。表２より，「所在」や「所」のように複数の帳票に共通する項目名や，「地積」や「一棟の建物番号」，「棟」，「表」のように帳票固有の項目名などがある。帳票固有の項目名をもたない種類の帳票でも，複数の項目を組み合わせて存在を判定することにより，「表示に関する建物登記済通知書（一般）」と「表示に関する建物登記済通知書（専有）」を除く５種類の帳票の種類を識別することができる。例えば，「床面積」の項目が存在し，「一棟の建物番号」の項目が存在しなければ「権利に関する建物登記済通知書（一般）」と識別することができる。
【００４３】
ステップ１８００では，ステップ３０８，３２２，３２４の結果を統合して帳票の種類を識別する。統合の手段としては，上記３つの結果の多数決を用いることができる。
【００４４】
ステップ１８００において，３０８，３２２，３２４の各ステップで，一意に帳票の種類

を識別できない場合でも，各ステップの処理結果を相互に補完することによって，帳票の種類を識別することもできる。例えば，ステップ３０８において，縦の実線罫線の本数が８本抽出された場合，表１より帳票の種類は「表示に関する土地登記済通知書」，「表示に関する建物登記済通知書（一般）」，「表示に関する建物登記済通知書（専有）」の３種類が考えられる。しかし，ステップ３２４において，項目名「表」が抽出されれば，「表示に関する土地登記済通知書」であると一意に決定できる。

【００４５】

なお，ステップ１８００において，３０８，３２２，３２４の３つのステップの結果を用いるのではなく，２つのみを用いることもできる。

【００４６】

なお，ステップ１８００において，３０８，３２２，３２４の各ステップの結果を同等に扱うのではなく，一つのステップで得た結果から帳票を識別し，他のステップで得た結果は，帳票識別の結果を検証するために用いることもできる。

【００４７】

図１９は，本発明の一実施例である登記情報システムの構成図である。１０１から１０９の構成は図１に同じである。ソータ１９００は，認識部１０１で認識し，修正部１０５で修正した結果に基づき，登記済通知書を記載内容の優先度順に帳票１００をソートする。以下にソートの例を２つ挙げる。第一は，所在と地番に該当する文字から，町ごとに丁目，番地，号の順にソートする。第二は，作成日，番号の順にソートする。また，ソートする対象は，登記済通知書の帳票でも，認識結果のデータでもよい。

【００４８】

【発明の効果】

本発明の帳票認識方法によれば，登記済通知書のような非定型帳票に対しても高精度に帳票の種類を識別することができる。

【００４９】

また，本発明の帳票認識方法によれば，下線を文字中のストロークなどと間違うことなく，高精度に抽出することができる。

【００５０】

また，本発明の帳票認識方法によれば，帳票の認識結果に基づいて，帳票をソートすることができる。

【図面の簡単な説明】

【図１】本発明の一実施例である登記情報認識システムの構成図である。

【図２】登記情報認識の処理過程を示すブロック図である。

【図３】図２で示した登記情報認識のＰＡＤ図である。

【図４】認識対象である登記済通知書の画像の説明図である。

【図５】図４の画像に対して図３のステップ３０４の罫線抽出処理をした結果を示す図である。

【図６】図４の画像に対して図３のステップ３０６の枠抽出処理をした結果を示す図である。

【図７】図４の画像に対して図３のステップ３１０の文字行抽出処理をした結果を示す図である。

【図８】図３のステップ３１４の下線抽出処理に関するＰＡＤ図である。

【図９】下線抽出対象画像の説明図である。

【図１０】図９の画像に対して罫線と文字行を抽出した結果を示す図である。

【図１１】図９の画像に対して罫線を抽出し，文字を切り出した結果を示す図である。

【図１２】文字行の一部の文字に対してのみ下線が印刷されている画像の説明図である。

【図１３】図３のステップ３１４の下線抽出処理に関するＰＡＤ図である。

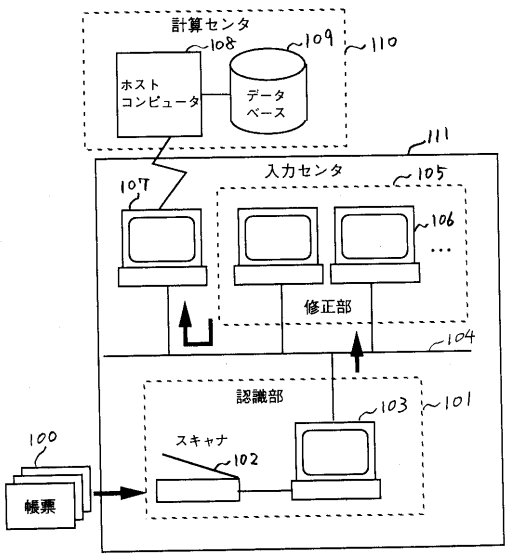
【図１４】下線抽出対象画像の説明図である。

【図１５】図１４の画像に対して枠線を構成しないランレングスデータと長い下線線を抽出した結果を示す図である。

【図１６】図３のステップ３１４の下線抽出処理に関するＰＡＤ図である。
【図１７】図１４の画像に対して枠線を構成しないランレングスデータを抽出し，ランレングスデータの長さを傾き方向に投影した結果を示す図である。
【図１８】図３のステップ３２８の帳票識別処理に関するＰＡＤ図である。
【図１９】本発明の一実施例である，ソート機能をもつ登記情報認識システムの構成図である。
【符号の説明】
２００…画像入力、２０４…罫線抽出、２０６…枠抽出、２０８…表特徴抽出、２４６…帳票識別、２２２…文字切り出し、２２４…文字識別、２３６…単語照合、２４０…内容照合

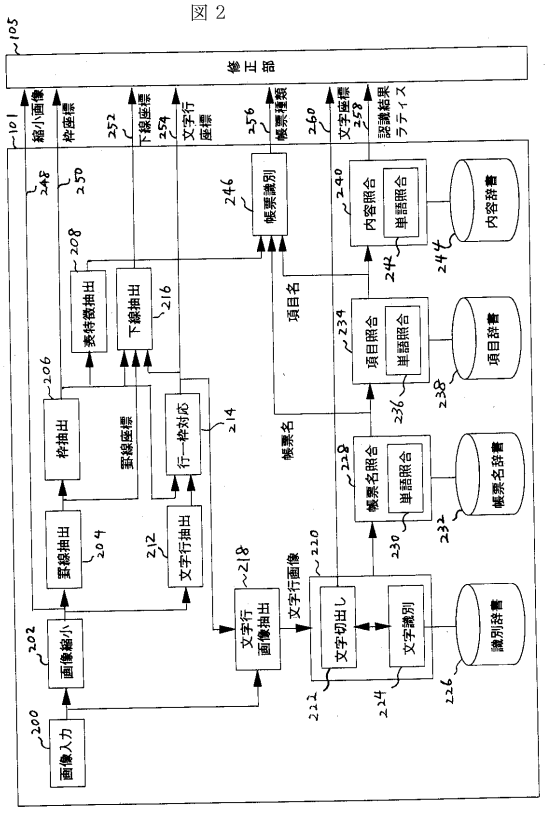
【図１】

図１



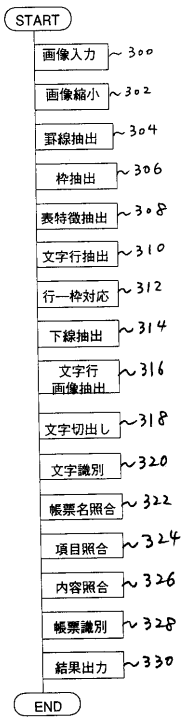
【図２】

図２



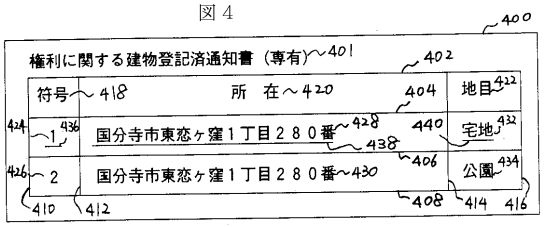
【図３】

図３



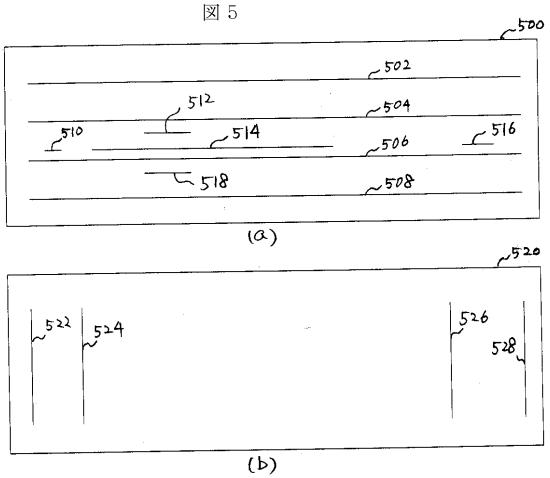
【図４】

図４



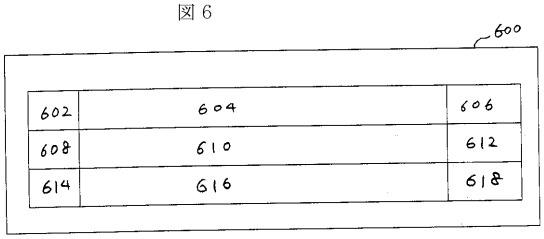
【図５】

図５



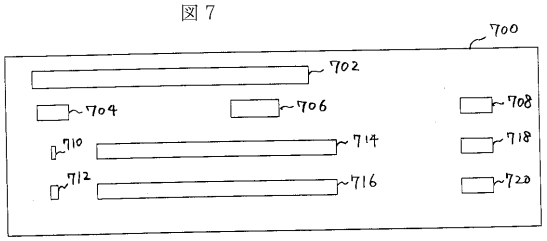
【図６】

図６



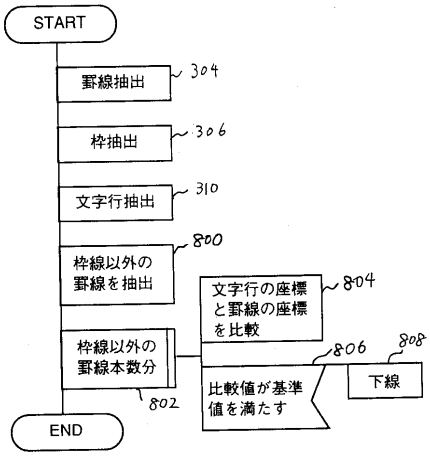
【図７】

図７



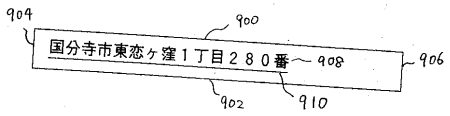
【図８】

図８



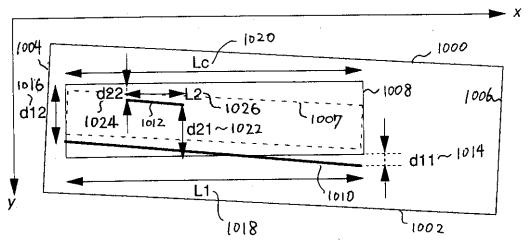
【図９】

図９



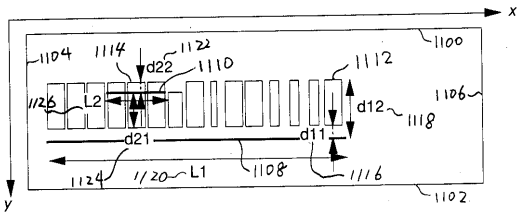
【図１０】

図１０



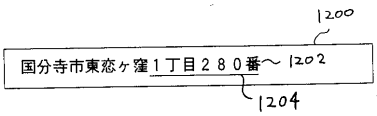
【図１１】

図１１



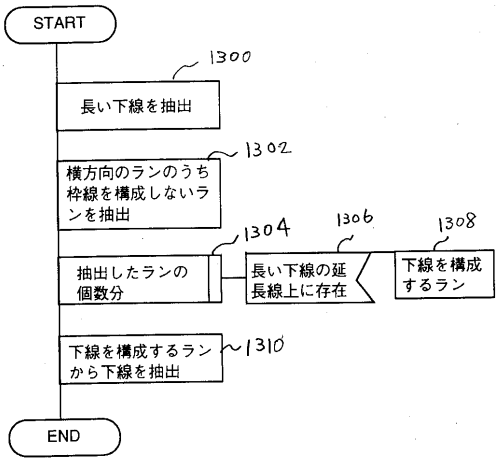
【図１２】

図１２



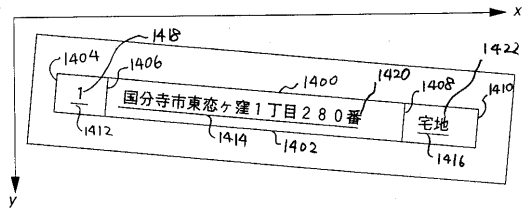
【図１３】

図１３



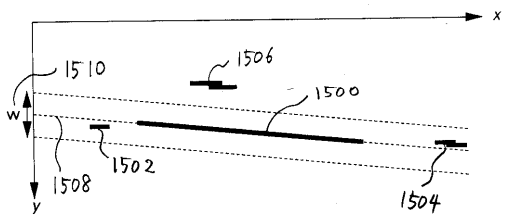
【図１４】

図１４



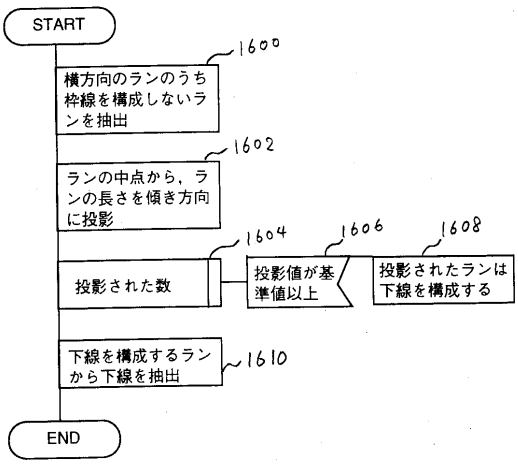
【図１５】

図１５



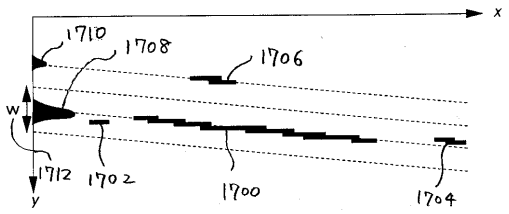
【図１６】

図１６



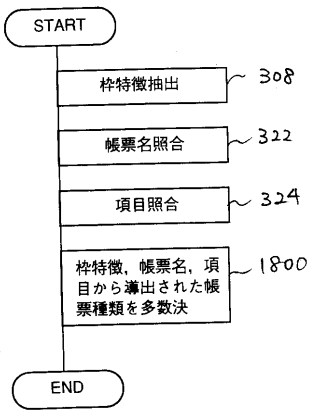
【図１７】

図１７



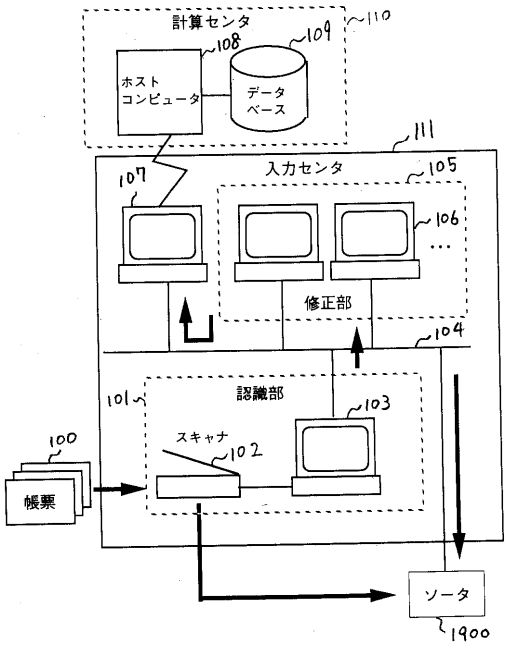
【図１８】

図１８



【図１９】

図１９



フロントページの続き

(72)発明者　丸川　勝美
東京都国分寺市東恋ケ窪１丁目２８０番地　株式会社日立製作所中央研究所内
(72)発明者　中島　和樹
東京都国分寺市東恋ケ窪１丁目２８０番地　株式会社日立製作所中央研究所内

合議体
審判長　田口　英雄
審判官　深沢　正志
審判官　佐藤　敬介

(56)参考文献　特開平７－３７０３６（ＪＰ，Ａ）
特開平５－１０１２１９（ＪＰ，Ａ）
特開平３－６２２８５（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.^{7}，ＤＢ名)
G06K9/00-9/82