(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4046941号
(P4046941)

(45)発行日　平成20年2月13日(2008.2.13)　(24)登録日　平成19年11月30日(2007.11.30)

(51)Int.Cl.　F I
G06K 9/20 (2006.01)　G06K 9/20 340C

請求項の数 11 (全 22 頁)

(21)出願番号　特願2000-367675(P2000-367675)
(22)出願日　平成12年12月1日(2000.12.1)
(65)公開番号　特開2002-170079(P2002-170079A)
(43)公開日　平成14年6月14日(2002.6.14)
審査請求日　平成16年6月11日(2004.6.11)

前置審査

(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　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
(74)代理人　100130409
弁理士　下山　治
(74)代理人　100134175
弁理士　永川　行光

最終頁に続く

(54)【発明の名称】文書書式識別装置および識別方法

(57)【特許請求の範囲】

【請求項１】

文書画像の文書書式を識別する文書書式識別装置であって、

文書書式を識別すべき文書画像から抽出された複数のブロックそれぞれの位置座標を含む文書書式データを作成する作成手段と、

前記作成手段により作成された前記識別すべき文書画像の文書書式データと、保存手段に保存されているマスター文書画像の文書書式データとを比較することにより、相似関係があるか否かを判断するとともに、当該判断結果と、相似関係があると判断した前記識別すべき文書画像と前記マスター文書画像との間の変倍率と、を含む相似情報を抽出する相似情報抽出手段と、

前記相似情報抽出手段にて抽出した相似情報及び前記文書書式データに基づいて、前記識別すべき文書画像の前記マスター文書画像に対する類似度を計算することにより、前記識別すべき文書画像の文書書式を識別する識別手段と、を備え、

前記相似情報抽出手段は、

前記識別すべき文書画像から抽出されたブロックの個数と前記マスター文書画像のブロックの個数とが所定数以上でかつ互いに等しい場合に、

前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＸ座標値を所定の順序に基づいて整列させることにより得たＸ座標値列（この座標値列をＹｉとし、その平均値をＹａｖｅとする）と、前記マスター文書画像から抽出された複数のブロックそれぞれから得たＸ座標値を所定の順序に基づいて整列させることにより得たＸ座標

値列（この座標値列をＸｉとし、その平均値をＸａｖｅとする）との間の相関係数を、
｛Σ（Ｘｉ－Ｘａｖｅ）（Ｙｉ－Ｙａｖｅ）｝／｛（（Σ（Ｘｉ－Ｘａｖｅ）＾２）×（Σ（Ｙｉ－Ｙａｖｅ）＾２））＾（１／２）｝
を用いて求め、

当該求めた相関係数が所定値より大きい場合に、前記識別すべき文書画像に関するＸ座標値列と前記マスター文書画像に関するＸ座標値列の傾きに基づいてＸ座標方向の変倍率を求める一方、当該求めた相関係数が所定値以下の場合には相似関係がないと判断し、

更に、前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＹ座標値を所定の順序に基づいて整列させることにより得たＹ座標値列（この座標値列をＹｉ’とし、その平均値をＹａｖｅ’とする）と、前記マスター文書画像から抽出された複数のブロックそれぞれから得たＹ座標値を所定の順序に基づいて整列させることにより得たＹ座標値列（この座標値列をＸｉ’とし、その平均値をＸａｖｅ’とする）との間の相関係数を、
｛Σ（Ｘｉ’－Ｘａｖｅ’）（Ｙｉ’－Ｙａｖｅ’）｝／｛（（Σ（Ｘｉ’－Ｘａｖｅ’）＾２）×（Σ（Ｙｉ’－Ｙａｖｅ’）＾２））＾（１／２）｝
を用いて求め、

当該求めた相関係数が所定値より大きい場合に、前記識別すべき文書画像に関するＹ座標値列と前記マスター文書画像に関するＹ座標値列の傾きに基づいてＹ座標方向の変倍率を求める一方、当該求めた相関係数が所定値以下の場合には相似関係がないと判断し、

前記Ｘ座標方向の変倍率と前記Ｙ座標方向の変倍率の両方が求められた場合に相似関係があると判断することを特徴とする文書書式識別装置。

【請求項２】

前記識別手段は、

前記相似情報に基づいて、前記類似度の計算に用いる文書書式データを補正し、類似度の計算を行うことを特徴とする請求項１記載の文書書式識別装置。

【請求項３】

前記識別手段は、

前記変倍率を前記類似度の計算に用いる文書書式データに乗算することを特徴とする請求項２記載の文書書式識別装置。

【請求項４】

前記相似情報抽出手段は、

前記識別すべき文書画像から抽出されたブロックの個数と前記マスター文書画像のブロックの個数とが所定数以上でかつ互いに等しい場合に、

前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＸ座標値を所定の順序に基づいて整列させることにより得たＸ座標値列（この座標値列をＹｉとし、その平均値をＹａｖｅとする）と、前記マスター文書画像から抽出された複数のブロックそれぞれから得たＸ座標値を所定の順序に基づいて整列させることにより得たＸ座標値列（この座標値列をＸｉとし、その平均値をＸａｖｅとする）との間の相関係数を、
｛Σ（Ｘｉ－Ｘａｖｅ）（Ｙｉ－Ｙａｖｅ）｝／｛（（Σ（Ｘｉ－Ｘａｖｅ）＾２）×（Σ（Ｙｉ－Ｙａｖｅ）＾２））＾（１／２）｝
を用いて求め、

当該求めた相関係数が所定値より大きい場合に、前記識別すべき文書画像に関するＸ座標値列と前記マスター文書画像に関するＸ座標値列の傾き及びその切片に基づいて、Ｘ座標方向の変倍率とＸ座標方向の原点ずれ量とを求める一方、当該求めた相関係数が所定値以下の場合には相似関係がないと判断し、

更に、前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＹ座標値を所定の順序に基づいて整列させることにより得たＹ座標値列（この座標値列をＹｉ’とし、その平均値をＹａｖｅ’とする）と、前記マスター文書画像から抽出された複数のブロックそれぞれから得たＹ座標値を所定の順序に基づいて整列させることにより得たＹ座標値列（この座標値列をＸｉ’とし、その平均値をＸａｖｅ’とする）との間の

相関係数を、
｛Σ（Ｘｉ’－Ｘａｖｅ’）（Ｙｉ’－Ｙａｖｅ’）｝／｛（（Σ（Ｘｉ’－Ｘａｖｅ’）＾２）×（Σ（Ｙｉ’－Ｙａｖｅ’）＾２））＾（１／２）｝
を用いて求め、
当該求めた相関係数が所定値より大きい場合に、前記識別すべき文書画像に関するＹ座標値列と前記マスター文書画像に関するＹ座標値列の傾き及びその切片に基づいて、Ｙ座標方向の変倍率とＹ座標方向の原点ずれ量とを求める一方、当該求めた相関係数が所定値以下の場合には相似関係がないと判断し、
前記Ｘ座標方向の変倍率と前記Ｙ座標方向の変倍率の両方が求められた場合に相似関係があると判断し、
前記識別手段は、
前記原点ずれ量を前記類似度の計算に用いる文書書式データに加算することを特徴とする請求項３記載の文書書式識別装置。
【請求項５】
前記相似情報は、前記変倍率に基づいて算出したペナルティを含み、前記識別手段は、該ペナルティを前記類似度の計算に用いることを特徴とする請求項２記載の文書書式識別装置。
【請求項６】
前記相似情報抽出手段は、
前記変倍率の適正を判定する判定手段を更に備え、
前記識別手段は、
前記判定手段にて変倍率が不適正であると判定した場合、前記相似情報を用いずに前記類似度計算を行うことを特徴とする請求項１記載の文書書式識別装置。
【請求項７】
前記相似情報抽出手段は、
前記識別すべき文書画像から抽出されたブロックの個数と前記マスター文書画像から抽出されたブロックの個数とが等しくない場合、相似関係がないと判断することを特徴とする請求項１に記載の文書書式識別装置。
【請求項８】
前記相似情報抽出手段は、
前記識別すべき文書画像から抽出されたブロックの個数と前記マスター文書画像から抽出されたブロックの個数とが前記所定数より小さくかつ互いに等しい場合、
前記識別すべき文書画像の幅と高さ、ならびに前記マスター文書画像の幅と高さとに基づいて、Ｘ座標方向の変倍率とＹ座標方向の変倍率とを求め、
前記求めたＸ座標方向の変倍率とＹ座標方向の変倍率、前記識別すべき文書画像から抽出されたブロックの幅と高さ、ならびに前記マスター文書画像から抽出されたブロックの幅と高さが、所定の関係式をみたすか否かを判断し、
当該所定の関係式を満たさない場合に相似関係がないと判断し、当該所定の関係式を満たす場合には相似関係があると判断することを特徴とする請求項１記載の文書書式識別装置。
【請求項９】
前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＸ座標値は、各ブロックの左上角のＸ座標値であり、
前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＹ座標値は、各ブロックの左上角のＹ座標値であることを特徴とする請求項１記載の文書書式識別装置。
【請求項１０】
文書画像の文書書式を識別する文書書式識別方法であって、
作成手段が、文書書式を識別すべき文書画像から抽出された複数のブロックそれぞれの位置座標を含む文書書式データを作成する作成工程と、

相似情報抽出手段が、前記作成工程で作成された前記識別すべき文書画像の文書書式データと、保存手段に保存されているマスター文書画像の文書書式データとを比較することにより、相似関係があるか否かを判断するとととともに、当該判断結果と、相似関係があると判断した前記識別すべき文書画像と前記マスター文書画像との間の変倍率と、を含む相似情報を抽出する相似情報抽出工程と、

識別手段が、前記相似情報抽出工程にて抽出した相似情報及び前記文書書式データに基づいて、前記識別すべき文書画像の前記マスター文書画像に対する類似度を計算することにより、前記識別すべき文書画像の文書書式を識別する識別工程と、を備え、

前記相似情報抽出工程では、

前記識別すべき文書画像から抽出されたブロックの個数と前記マスター文書画像のブロックの個数とが所定数以上でかつ互いに等しい場合に、

前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＸ座標値を所定の順序に基づいて整列させることにより得たＸ座標値列（この座標値列をＹｉとし、その平均値をＹａｖｅとする）と、前記マスター文書画像から抽出された複数のブロックそれぞれから得たＸ座標値を所定の順序に基づいて整列させることにより得たＸ座標値列（この座標値列をＸｉとし、その平均値をＸａｖｅとする）との間の相関係数を、

｛Σ（Ｘｉ－Ｘａｖｅ）（Ｙｉ－Ｙａｖｅ）｝／｛（（Σ（Ｘｉ－Ｘａｖｅ）＾２）×（Σ（Ｙｉ－Ｙａｖｅ）＾２））＾（１／２）｝

を用いて求め、

当該求めた相関係数が所定値より大きい場合に、前記識別すべき文書画像に関するＸ座標値列と前記マスター文書画像に関するＸ座標値列の傾きに基づいてＸ座標方向の変倍率を求める一方、当該求めた相関係数が所定値以下の場合には相似関係がないと判断し、

更に、前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＹ座標値を所定の順序に基づいて整列させることにより得たＹ座標値列（この座標値列をＹｉ’とし、その平均値をＹａｖｅ’とする）と、前記マスター文書画像から抽出された複数のブロックそれぞれから得たＹ座標値を所定の順序に基づいて整列させることにより得たＹ座標値列（この座標値列をＸｉ’とし、その平均値をＸａｖｅ’とする）との間の相関係数を、

｛Σ（Ｘｉ’－Ｘａｖｅ’）（Ｙｉ’－Ｙａｖｅ’）｝／｛（（Σ（Ｘｉ’－Ｘａｖｅ’）＾２）×（Σ（Ｙｉ’－Ｙａｖｅ’）＾２））＾（１／２）｝

を用いて求め、

当該求めた相関係数が所定値より大きい場合に、前記識別すべき文書画像に関するＹ座標値列と前記マスター文書画像に関するＹ座標値列の傾きに基づいてＹ座標方向の変倍率を求める一方、当該求めた相関係数が所定値以下の場合には相似関係がないと判断し、

前記Ｘ座標方向の変倍率と前記Ｙ座標方向の変倍率の両方が求められた場合に相似関係があると判断することを特徴とする文書書式識別方法。

【請求項１１】

請求項１０に記載の文書書式識別方法をコンピュータによって実行させるための制御プログラムを格納する記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、大量の帳票を処理する分野で、帳票の書式ごとに分類する装置を構築する際に、自動分類を可能にする帳票識別に関するものである。

【０００２】

【従来の技術】

帳票内に記載された情報をＯＣＲ等の光学式文字認識装置で読みとるためには、帳票の書式を識別し、帳票内の情報記載位置を正確に把握する必要がある。帳票の書式を識別する方法として、あらかじめ登録されたマスタ帳票と、識別したい帳票の帳票内のテーブルや文字を比較し、識別したい帳票と一致するマスタ帳票を抽出する方法がある。

【０００３】
帳票内のテーブルや文字を比較するには、識別したい帳票のテーブルブロックおよびテキストブロックに、最も近い座標位置にあるマスタ帳票のブロックを検出し、テーブルブロックおよびテキストブロックごとの詳細情報のマッチングを取る手法が一般的である。そのブロック位置の検出には、帳票ページの左上角を原点として、各ブロックの左上角位置の座標値を使用していた。
【０００４】
【発明が解決しようとする課題】
しかしながら、識別したい帳票がＦＡＸ等により送信された場合、ＦＡＸ等の給紙制約から、識別したい帳票が拡大または縮小されることがある。そして、図２の（Ａ）、（Ｂ）に示すようにマスタ帳票Ｂに対して、拡大または縮小した識別したい帳票Ａを比較すると、識別したい帳票の各ブロックの左上角位置は変倍されているため、識別したい帳票のブロックに対応するマスタ帳票のブロックを正確に検出することができない。また、ブロックごとの詳細構造のマッチングでも、テーブルブロックのサイズ、テーブル内の罫線の位置情報も、前記と同様に変倍されているので、相似形のテーブルブロックでもテーブル構造のマッチング計算では、異なるテーブルだと識別されることになる。その結果、拡大または縮小した帳票は、類似度が非常に低くなり、異なる帳票として判断されることになる。
【０００５】
本発明は、上記課題を鑑みてなされたものであり、異なる変倍率で拡大または縮小された複数の文書が混在する環境でも、文書書式を正しく識別することを目的とする。
【０００６】
【課題を解決するための手段】
かかる課題を解決するため、例えば本発明の文書書式識別装置は以下の構成を備える。すわなち、
文書画像の文書書式を識別する文書書式識別装置であって、
文書書式を識別すべき文書画像から抽出された複数のブロックそれぞれの位置座標を含む文書書式データを作成する作成手段と、
前記作成手段により作成された前記識別すべき文書画像の文書書式データと、保存手段に保存されているマスター文書画像の文書書式データとを比較することにより、相似関係があるか否かを判断するとともに、当該判断結果と、相似関係があると判断した前記識別すべき文書画像と前記マスター文書画像との間の変倍率と、を含む相似情報を抽出する相似情報抽出手段と、
前記相似情報抽出手段にて抽出した相似情報及び前記文書書式データに基づいて、前記識別すべき文書画像の前記マスター文書画像に対する類似度を計算することにより、前記識別すべき文書画像の文書書式を識別する識別手段と、を備え、
前記相似情報抽出手段は、
前記識別すべき文書画像から抽出されたブロックの個数と前記マスター文書画像のブロックの個数とが所定数以上でかつ互いに等しい場合に、
前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＸ座標値を所定の順序に基づいて整列させることにより得たＸ座標値列（この座標値列をＹｉとし、その平均値をＹａｖｅとする）と、前記マスター文書画像から抽出された複数のブロックそれぞれから得たＸ座標値を所定の順序に基づいて整列させることにより得たＸ座標値列（この座標値列をＸｉとし、その平均値をＸａｖｅとする）との間の相関係数を、
｛Σ（Ｘｉ－Ｘａｖｅ）（Ｙｉ－Ｙａｖｅ）｝／｛（（Σ（Ｘｉ－Ｘａｖｅ）＾２）×（Σ（Ｙｉ－Ｙａｖｅ）＾２））＾（１／２）｝
を用いて求め、
当該求めた相関係数が所定値より大きい場合に、前記識別すべき文書画像に関するＸ座標値列と前記マスター文書画像に関するＸ座標値列の傾きに基づいてＸ座標方向の変倍率を求める一方、当該求めた相関係数が所定値以下の場合には相似関係がないと判断し、

更に、前記識別すべき文書画像から抽出された前記複数のブロックそれぞれから得たＹ座標値を所定の順序に基づいて整列させることにより得たＹ座標値列（この座標値列をＹｉ’とし、その平均値をＹａｖｅ’とする）と、前記マスター文書画像から抽出された複数のブロックそれぞれから得たＹ座標値を所定の順序に基づいて整列させることにより得たＹ座標値列（この座標値列をＸｉ’とし、その平均値をＸａｖｅ’とする）との間の相関係数を、

｛Σ（Ｘｉ’－Ｘａｖｅ’）（Ｙｉ’－Ｙａｖｅ’）｝／｛（（Σ（Ｘｉ’－Ｘａｖｅ’）＾２）×（Σ（Ｙｉ’－Ｙａｖｅ’）＾２））＾（１／２）｝

を用いて求め、

当該求めた相関係数が所定値より大きい場合に、前記識別すべき文書画像に関するＹ座標値列と前記マスター文書画像に関するＹ座標値列の傾きに基づいてＹ座標方向の変倍率を求める一方、当該求めた相関係数が所定値以下の場合には相似関係がないと判断し、

前記Ｘ座標方向の変倍率と前記Ｙ座標方向の変倍率の両方が求められた場合に相似関係があると判断することを特徴とする。

【０００７】

【発明実施の形態】

［実施形態１］

以下、図面を参照して本発明の実施の形態を詳細に説明する。

【０００８】

図１は、本発明の実施の形態に係る帳票書式自動識別装置の概略構成を示すブロック図である。

【０００９】

１１はスキャナーであり、帳票イメージを光学的に読み取り、帳票イメージデータを出力する。１２はプロセッサでありメモリ１５に格納された制御プログラム１５ｄを実行することにより、画像特徴量抽出手段１２ａ、書式データ作成手段１２ｂ、類似度計算手段１２ｃとして機能する。スキャナー１１で読み取った画像は、帳票イメージ１５ｃとしてメモリ１５に格納される。帳票イメージ１５ｃは、２値化処理されて画像特徴量抽出手段１２ａに送られ、黒ドットのヒストグラム法などの手法により、テーブル、テキスト、ピクチャなどブロックごとに属性分類される。テーブルブロックについては、さらに罫線追跡手法などで、テーブルの詳細構造を求める。また、テキストブロックについては、さらに文字コードに変換するなどの処理を行う。

【００１０】

このようにして取得した情報から、書式データ作成手段１２ｂにて、図３に示す帳票のページ書式およびテーブル書式を作成し、メモリ１５およびディスク１４に保存する。図３は、画像特徴量抽出手段１２ａで抽出した後の帳票サンプル３１を示す。テーブル・ブロック３個（３１１～３１３）、ピクチャ・ブロック１個（３１４）が抽出されている。この帳票の書式データ３２は、ページ書式３２１とテーブル書式３２２に階層化して保存する。ページ書式３２１は、ヘッダ部３２１ａに帳票ページ幅、帳票ページ高さを所有する。

【００１１】

また、データ部３２２ａにはブロック毎に各種情報が記憶される。例えば、ブロック属性がテーブルの場合には、位置情報としてブロック左端位置、ブロック上端位置、大きさ情報としてブロック幅、ブロック高さの情報を所有する。また、比較帳票をピックアップするのに使用するためのページ原点からの距離および類似度の計算に使用するための当該ブロックの面積を全テーブル・ブロックで割った値も所有する。さらに、テーブル詳細情報とリンクするためにテーブルＩＤを所有する。このテーブルＩＤにリンクしたテーブルのセルの詳細構造をテーブル書式３２２で示す。テーブル内のセル個数、セルの位置、大きさ情報を所有する。

【００１２】

帳票書式識別装置は、キーボードから帳票の登録、帳票の識別などの命令が入力されると

、各々の命令に対応する処理をプロセッサ１２が上記の書式データ３２を使用して行う。そして、その識別結果をディスプレイ１６に表示する。

【００１３】

図４を参照して、本実施形態の帳票書式識別装置、特に図１のプロセッサ１２が実行する各種制御処理の動作を説明する。

【００１４】

図４は、帳票書式識別装置による書式識別処理の概略フローチャートを示す。ステップＳ１０１にて、識別したい帳票をスキャナーで読み取り、ステップＳ１０３で、テーブルブロック、テキストブロックの座標値等の特徴量データを抽出する。ステップＳ１０５では、これらの特徴量データを類似度を計算するための書式データに変換する。この書式データを元にして、ステップＳ１０７にてマスター帳票の中から、当該識別したい帳票の書式データと同じ帳票である可能性のある帳票を絞り込む。ステップＳ１０９で絞り込んだ帳票のすべてについて書式の類似度を計算する（ステップＳ１１１）。計算の結果、類似度の高い方の所定数のマスター帳票を類似帳票の候補とし、その識別コードと類似度を出力する（ステップＳ１１３）。

【００１５】

本発明の帳票レイアウトの相似形チェックは、ステップＳ１０８にて行う。図５～図８を使用して、詳細にその処理を説明する。

【００１６】

ステップＳ１０８で帳票レイアウト相似形チェック処理が開始され、ステップＳ２０３で識別したい帳票およびマスタ帳票のフレームおよびテーブルブロックを整列する。本実施形態では、ブロック左上角のＸ座標の昇順にブロック情報を並べる。但し、図３に示すような、ブロック左上角座標のＸ成分がほとんど同じ位置にある場合には、Ｘ座標の誤差により識別したい帳票とマスタ帳票のブロック情報の並びを必ずしも対応付けることができない。識別したい帳票のブロック情報がテーブル１（３１１）→テーブル２（３１２）→ピクチャブロック（３１４）→テーブルブロック３（３１３）と整列しても、マスタ帳票のブロック情報がテーブル２（３１２）→テーブル１（３１１）→ピクチャブロック（３１４）→テーブルブロック３（３１３）のように整列する可能性は十分にある。そこで、Ｘ座標の位置が５ピクセル以内の差のブロックに関しては、別途Ｙ成分の昇順に整列させる。この結果、識別したい帳票およびマスタ帳票のブロック情報をテーブル１（３１１）→テーブル２（３１２）→ピクチャブロック（３１４）→テーブルブロック３（３１３）の順に整列することが保証される。

【００１７】

ステップＳ２０５では、比較するブロックの個数が同じであるかをチェックしている。ブロックの個数が異なる場合は、相似形でないと判断して、帳票レイアウト相似形チェック処理を終了し、ステップＳ１０９へ戻る。

【００１８】

ステップＳ２０５でブロック個数が同じであった場合には、ステップＳ２０７にて、ブロック個数が３個以上かをチェックしている。３個以上の場合と、２個以下の場合では、相似形判定プログラムが異なるからである。

【００１９】

３個以上の場合は、ステップＳ２０８＿１に進み、ブロック左上角Ｘ座標の比較処理を行う。ステップＳ２０８＿１の詳細を図６のフローチャートを用いて説明する。

【００２０】

ステップＳ２０９では、ブロック情報のＸ成分の相似形チェック処理をする。すなわち、識別したい帳票の左上角Ｘ座標を縦軸に、マスタ帳票の左上角Ｘ座標を横軸にして、下式より相関係数を計算する。

【００２１】

【数１】

$$\text{相関係数} = \frac{\Sigma(Xi - Xave)(Yi - Yave)}{\sqrt{\Sigma(Xi - Xave)^2 \times \Sigma(Yi - Yave)^2}}$$

Xi, Yi：ＸおよびＹ座標値
Xave, Yave：ＸおよびＹ座標の平均値

【００２２】
ここで、相関係数の算出にあたっては、上式の相関係数の分子の値をチェックし（ステップＳ２１１）、相関係数の分子が１４以下であれば、別プログラムで変倍率δＸを求めている。これは、以下に述べるように相関係数の誤差が大きくなるからである。
【００２３】
図３の帳票では、テーブルブロック１（３１１）、テーブルブロック２（３１２）、ピクチャブロック３１４の各左上角のＸ座標はほとんど同じ位置にあるため、Ｘｉ、ＹｉともにＸａｖｅ、Ｙａｖｅに近い値となる。従って、Ｘｉ、Ｙｉが誤差の為に変動した場合、相関係数の変動も大きくなる。このため、Ｘｉ、Ｙｉの誤差が大きいと考えられる環境では、相関係数の信頼度が落ちる。誤差の要因としては、スキャナで読み取るときに発生する誤差、傾斜補正等の画像処理を行ったときに発生する誤差、マッチング（すべての画像を１００ｄｐｉに正規化してマッチングを行っている）の為に解像度変換を行ったときに発生する誤差などがあげられる。それらの誤差を考慮して、上式の相関係数の分子が１４以下の場合は、信頼できないとして（識別したい帳票のページ幅）／（マスタ帳票のページ幅）＝変倍率δＸとしている（ステップＳ２１１＿２）。ただし、帳票をはさみ等で切り取った場合にできるズレによる誤差を考慮して、帳票ページ幅の差分が１０ピクセル以内であれば、変倍率δＸを１．０とする（ステップＳ２１１＿１、Ｓ２１１＿３）。
【００２４】
ステップＳ２１１にて相関係数の分子が１４以上であり、かつ相関係数が０．９９９６以上あれば、Ｘ成分に関しては相似形と判断する（ステップＳ２１３）。
【００２５】
一方、ステップＳ２１１にて相関係数の分子が１４以上であっても、相関係数が０．９９９６より小さい場合には、相似形でないと判断して相似形チェック処理を終了し、ステップＳ１０９へもどる。
【００２６】
相関係数が１に近いほど、前記の情報が直線上に並んでいるとみなすことができる。つまり、変倍されている可能性はあるが、Ｘ成分の並びは相似形であると考える。
【００２７】
次に、これらのデータが直線上に並んでいるので、その傾きを計算する。この傾きがＸ成分の変倍率δＸを示す（ステップＳ２１５）。傾き＝１の時は、識別したい帳票とマスタ帳票のＸ成分は同じであり、傾きが１より小さいと、マスタ帳票のＸ成分の増加率が識別したい帳票の比べて大きいことになり、マスタ帳票の幅が識別したい帳票の幅に比べて拡大しているといえる。つまり、傾き＝（識別したい帳票のページ幅）／（マスタ帳票のページ幅）の関係が成り立つ。
【００２８】
ステップＳ２１７では、前記の誤差を考慮して、変倍率δＸが１±０．０２８以内であれば変倍がないとみなし、変倍率δＸを１にリセットする。
【００２９】
次に、ステップＳ２１９でこの変倍率δＸが、テーブルブロックのサイズまで適用できる

かをチェックしている。すなわち、ステップＳ２０３で整列した帳票のブロック情報を順に１個ずつ、「（識別したい帳票のブロック幅）／（マスタ帳票のブロック幅）＜変倍率δＸ＋０．０２７＋１．９／（マスタ帳票のブロック幅）」を満たすかどうかをチェックし、満たさない場合には、ブロックレイアウトは相似でないとして当該処理を終了し、ステップＳ１０９にもどる。また、満たす場合には、変倍率δＸのテーブルブロックサイズへの適用可と判断し、処理を終了する。

【００３０】

ステップＳ２０８＿１と同様に、ステップＳ２０８＿２ではブロック左上角Ｙ座標の比較処理を行う。ステップＳ２０８＿２の詳細フローチャートを図７に示す。

【００３１】

ステップＳ２１０では、ブロック情報のＹ成分の相似形チェック開始する。すなわち、識別したい帳票の左上角Ｙ座標を縦軸に、マスタ帳票の左上角Ｙ座標を横軸にして、相関係数を計算する。

【００３２】

ここで、相関係数の算出にあたっては、相関係数の分子の値をチェックし（ステップＳ２１２）、相関係数の分子が１４以下であれば、別プログラムで変倍率δＹを求めている。相関係数の分子が１４以下の場合は、信頼できないとして（識別したい帳票のページ高さ）／（マスタ帳票のページ高さ）＝変倍率δＹとしている（ステップＳ２１２＿２）。ただし、帳票をはさみ等で切り取った場合にできるズレによる誤差を考慮して、帳票ページ幅の差分が１０ピクセル以内であれば、変倍率δＹを１．０とする（ステップＳ２１２＿１、Ｓ２１２＿３）。

【００３３】

ステップＳ２１２にて相関係数の分子が１４以上であり、かつ相関係数が０．９９９６以上あれば、Ｙ成分に関しては相似形と判断する（ステップＳ２１４）。

【００３４】

一方、ステップＳ２１２にて相関係数の分子が１４以上であっても、相関係数が０．９９９６より小さい場合には、相似形でないと判断して相似形チェック処理を終了し、ステップＳ１０９へもどる。

【００３５】

相関係数が１に近いほど、前記の情報が直線上に並んでいるとみなすことができる。つまり、変倍されている可能性はあるが、Ｙ成分の並びは相似形であると考える。

【００３６】

次に、これらのデータが直線上に並んでいるので、その傾きを計算する。この傾きがＹ成分の変倍率δＹを示す（ステップＳ２１６）。傾き＝１の時は、識別したい帳票とマスタ帳票のＹ成分は同じであり、傾きが１より小さいと、マスタ帳票のＹ成分の増加率が識別したい帳票の比べて大きいことになり、マスタ帳票の幅が識別したい帳票の幅に比べて拡大しているといえる。つまり、傾き＝（識別したい帳票のページ高さ）／（マスタ帳票のページ高さ）の関係が成り立つ。

【００３７】

ステップＳ２１８では、誤差を考慮して、変倍率δＹが１±０．０２８以内であれば変倍がないとみなし、変倍率δＹを１にリセットする。

【００３８】

次に、ステップＳ２２０でこの変倍率δＹが、テーブルブロックのサイズまで適用できるかをチェックしている。すなわち、ステップＳ２０３で整列した帳票のブロック情報を順に１個ずつ、「（識別したい帳票のブロック高さ）／（マスタ帳票のブロック高さ）＜変倍率δＹ＋０．０２７＋１．９／（マスタ帳票のブロック高さ）」を満たすかどうかをチェックし、満たさない場合には、ブロックレイアウトは相似でないとして当該処理を終了し、ステップＳ１０９にもどる。また、満たす場合には、変倍率δＹのテーブルブロックサイズへの適用可と判断し、処理を終了する。

【００３９】

ステップＳ２２１では、Ｘ成分、Ｙ成分両方ともに相似形であり、変倍率（δＸ、δＹ）を取得できた帳票のみ、レイアウトが相似形であると判断し、それ以外の場合には、相似でないとして当該処理を終了し、ステップＳ１０９にもどる。

【００４０】

さて、ステップＳ２０７でブロックの個数が２個以下の場合は、図８のブロック個数が２個以下の場合の処理を行う（ステップＳ３０１）。

【００４１】

すなわち、ステップＳ３０１でブロック個数が２個以下の場合の処理が開始され、ステップＳ３０２にて識別したい帳票のページ幅とマスター帳票のページ幅を比較し、差分が１０ピクセルより大きい場合には、変倍率δＸ＝（識別したい帳票のページ幅）／（マスタ帳票のページ幅）とし、差分が１０ピクセル以内の場合には変倍率δＸ＝１．０とする（ステップＳ３０２、Ｓ３０３、Ｓ３０４）。

【００４２】

さらにステップＳ３０５にてこの変倍率δＸが、テーブルブロックのサイズまで適用できるかをチェックしている。すなわち、ステップＳ２０３で整列した帳票のブロック情報を順に１個ずつ、「（識別したい帳票のブロック幅）／（マスタ帳票のブロック幅）＜変倍率δＸ＋０．０２７＋１．９／（マスタ帳票のブロック幅）」を満たすかどうかをチェックし、満たさない場合には、ブロックレイアウトは相似でないとして当該処理を終了し、ステップＳ１０９にもどる。また、満たす場合には、変倍率δＸのテーブルブロックサイズへの適用可と判断する。

【００４３】

同様に、ステップＳ３０６にて識別したい帳票のページ高さとマスター帳票のページ高さを比較し、差分が１０ピクセルより大きい場合には、変倍率δＹ＝（識別したい帳票のページ高さ）／（マスタ帳票のページ高さ）とし、差分が１０ピクセル以内の場合には変倍率δＹ＝１．０とする（ステップＳ３０６、Ｓ３０７、Ｓ３０８）。

【００４４】

さらにステップＳ３０９にてこの変倍率δＹが、テーブルブロックのサイズまで適用できるかをチェックしている。すなわち、ステップＳ２０３で整列した帳票のブロック情報を順に１個ずつ、「（識別したい帳票のブロック高さ）／（マスタ帳票のブロック高さ）＜変倍率δＹ＋０．０２７＋１．９／（マスタ帳票のブロック高さ）」を満たすかどうかをチェックし、満たさない場合には、ブロックレイアウトは相似形でないとして当該処理を終了し、ステップＳ１０９にもどる。また、満たす場合には、変倍率δＹのテーブルブロックサイズへの適用可と判断する。

【００４５】

ステップＳ３０５およびステップＳ３０９にてテーブルブロックサイズへの適用可能と判断された変倍率（δＸ、δＹ）について、ステップＳ３１０で、ブロックの左上角、「マスタ帳票のＸ座標×変倍率（δＸ）≦識別したい帳票のＸ座標±１０、かつマスタ帳票のＹ座標×変倍率（δＹ）≦識別したい帳票のＹ座標±１０」の条件式を満たしていれば、識別したい帳票とマスタ帳票のレイアウトは相似形であると判断し、帳票レイアウト相似形チェック処理のステップＳ２２３へ進む（ステップＳ３１１）。また、条件を満たさない場合には、相似形でないとして当該処理を終了し、ステップＳ１０９にもどる。

【００４６】

ステップＳ２２１またはステップＳ３１１にて相似形であると判断された場合には、ステップＳ２２３でページレイアウトの変倍によるペナルティを以下の式で決定する。

【００４７】

【数２】

$$PX = \sin\left(\frac{\pi}{4} - \tan^{-1}\text{変倍率}\delta X\right) \times 60.0$$

$$PY = \sin\left(\frac{\pi}{4} - \tan^{-1}\text{変倍率}\delta Y\right) \times 60.0$$

$$PXY = \sin\left(\tan^{-1}\text{変倍率}\delta X - \tan^{-1}\text{変倍率}\delta Y\right) \times 120.0$$

【００４８】
ＰＸ、ＰＹは各成分の変倍によるペナルティ、ＰＸＹは、両成分の変形度によるペナルティをあらわす。
【００４９】
変倍率δＸ、δＹが１ならば、ＰＸ、ＰＹは０である。つまり、変倍していないのでペナルティを課せないことを意味する。
【００５０】
ＰＸＹは、Ｘ、Ｙ成分が均等に変倍したときには０になるし、Ｘ成分が１より大きく、Ｙ成分が１より小さく変倍するように、変倍によるレイアウトの変形が大きくなればペナルティが大きくなるように調整する式である。
【００５１】
レイアウトが相似形である場合は、以上の計算式でペナルティを与えて、従来のページ書式のマッチングによるペナルティを０とする。逆に、相似形でない場合は、従来通りのページ書式のマッチングによるペナルティを与える。
【００５２】
ページ書式のマッチングの次に、テーブルブロックの詳細構造、その次にテキストブロックの文字比較を行うが、これらのブロックを検出する際には、相似形チェックで求めた変倍率を使用した計算式を使用する。
【００５３】
例えば、図２の（Ａ）、（Ｂ）に示すような識別したい帳票とマスタ帳票について、本実施形態による帳票レイアウト相似形チェック処理で、相似形だと判定され、変倍率（δＸ、δＹ）が得られたとする。
【００５４】
マスタ帳票の（Ｘ１、Ｙ１）のブロックに対応する識別したい帳票のブロックは、（δＸ×Ｘ１、δＹ×Ｙ１）で正確な位置を求めることができる。
【００５５】
この計算式で検出したブロックがテーブルブロックの場合は、テーブルの各罫線情報が帳票ページの変倍率と同様に変倍されているので、マスタ帳票の罫線情報（Ｌｘ、Ｌｙ）を（Ｌｘ×δＸ、Ｌｙ×δｙ）に変倍して、識別したい帳票の罫線情報と比較することで、正確なテーブルブロックの詳細構造のマッチングを行うことができる。
【００５６】
以上、記述した中での数値は、数多くの帳票サンプルを使用した統計値であり、帳票識別の環境によっては、変更してもかまわない。
【００５７】
［実施形態２］
原点ずれが生じると、変倍のみでは正しく認識できない恐れがある。
【００５８】
以下に図面を参照して本発明の実施形態のうち、識別したい帳票とマスタ帳票の原点位置がずれた場合の識別処理について詳細を説明する。
【００５９】
なお、帳票書式識別装置は図１と同様のものを使用し、図３と同様の書式データを作成す

る。したがって、図１と図３の内容は実施形態１と重複することから説明は省略する。
【００６０】
本実施形態の帳票書式識別装置、特に図１のプロセッサ１２が実行する各種制御処理のうち、実施形態１と異なる処理を中心に説明する。
【００６１】
本実施形態の帳票レイアウトの相似形チェックは、ステップＳ１０８にて行う。図１０～図１３を使用して、詳細にその処理を説明する。
【００６２】
図１０のステップＳ４０３からＳ４０７までは、図５のステップＳ２０３からＳ２０７までと同じ処理を行う。
【００６３】
ステップＳ４０８＿１ではブロック左上角Ｘ座標の比較処理を行う。この処理の詳細を図１１を参照して説明する。
【００６４】
すなわち、ステップＳ４０９では、ブロック情報のＸ成分の相似形チェックをすべく、識別したい帳票の左上角Ｘ座標を縦軸に、マスタ帳票の左上角Ｘ座標を横軸にして、相関係数を計算する。
【００６５】
ここで、相関係数の算出にあたっては、相関係数の分子の値をチェックし（ステップＳ４１１）、相関係数の分子が１４以下であれば、別プログラムで変倍率δＸを求めている。これは、相関係数の誤差が大きくなるからで、詳細は実施形態１と同じであるため、説明は省略する。
【００６６】
相関係数の分子が１４以下の場合は、信頼できないとして（識別したい帳票のページ幅）／（マスタ帳票のページ幅）＝変倍率δＸとしている（ステップＳ４１１＿２）。ただし、帳票をはさみ等で切り取った場合にできるズレによる誤差を考慮して、帳票ページ幅の差分が１０ピクセル以内であれば、変倍率δＸを１．０とする（ステップＳ４１１＿１、Ｓ４１１＿３）。
【００６７】
上記でもとめた変倍率δＸを用いて、ステップＳ４１１＿４にて、原点ずれ量ｓｈｉｆｔＸを「（識別したい帳票の先頭ブロックの左上角Ｘ座標）－（マスタ帳票の先頭ブロックの左上角Ｘ座標）×変倍率δＸ」より算出する。
【００６８】
ステップＳ４１１にて相関係数の分子が１４以上であり、かつ相関係数が０．９９９６以上あれば、Ｘ成分に関しては相似形と判断する（ステップＳ４１３）。
【００６９】
一方、ステップＳ４１１にて相関係数の分子が１４以上であっても、相関係数が０．９９９６より小さい場合には、相似形でないと判断して相似形チェック処理を終了し、ステップＳ１０９へもどる。
【００７０】
相関係数が１に近いほど、前記の情報が直線上に並んでいるとみなすことができる。つまり、変倍されている可能性はあるが、Ｘ成分の並びは相似形であると考える。
【００７１】
次に、これらのデータが直線上に並んでいるので、その傾きを計算する（回帰直線の傾き）。この傾きがＸ成分の変倍率δＸを示す（ステップＳ４１５）。傾き＝１の時は、識別したい帳票とマスタ帳票のＸ成分は同じであり、傾きが１より小さいと、マスタ帳票のＸ成分の増加率が識別したい帳票の比べて大きいことになり、マスタ帳票の幅が識別したい帳票の幅に比べて拡大しているといえる。一方、回帰直線の縦軸との切片がＸ座標の原点ずれ量ｓｈｉｆｔＸになる（ステップＳ４１７）。
【００７２】

ステップＳ４１９でこの変倍率δＸが、テーブルブロックのサイズまで適用できるかをチェックしている。すなわち、ステップＳ２０３で整列した帳票のブロック情報を順に１個ずつ、「（識別したい帳票のブロック幅）／（マスタ帳票のブロック幅）＝変倍率δＸ」を満たすかどうかをチェックし、満たさない場合には、ブロックレイアウトは相似でないとして当該処理を終了し、ステップＳ１０９にもどる。また、満たす場合には、変倍率δＸのテーブルブロックサイズへの適用可と判断し、処理を終了する。

【００７３】

ステップＳ４０８＿１と同様に、ステップＳ４０８＿２では、ブロック左上角Ｙ座標の比較処理を行う。ステップＳ４０８＿２の詳細フローチャートを図１２に示す。

【００７４】

すなわち、ステップＳ４１０では、ブロック情報のＸ成分の相似形チェックをすべく、識別したい帳票の左上角Ｙ座標を縦軸に、マスタ帳票の左上角Ｙ座標を横軸にして、相関係数を計算する。

【００７５】

ここで、相関係数の算出にあたっては、相関係数の分子の値をチェックし（ステップＳ４１２）、相関係数の分子が１４以下であれば、別プログラムで変倍率δＹを求めている。これは、相関係数の誤差が大きくなるからで、詳細は実施形態１と同じであるため、説明は省略する。

【００７６】

相関係数の分子が１４以下の場合は、信頼できないとして（識別したい帳票のページ高さ）／（マスタ帳票のページ高さ）＝変倍率δＹとしている（ステップＳ４１２＿２）。ただし、帳票をはさみ等で切り取った場合にできるズレによる誤差を考慮して、帳票ページ幅の差分が１０ピクセル以内であれば、変倍率δＹを１とする（ステップＳ４１２＿１、Ｓ４１２＿３）。

【００７７】

上記でもとめた変倍率δＹを用いて、ステップＳ４１２＿４にて、原点ずれ量ｓｈｉｆｔＹを「（識別したい帳票の先頭ブロックの左上角Ｙ座標）－（マスタ帳票の先頭ブロックの左上角Ｙ座標）×変倍率δＹ」より算出する。

【００７８】

ステップＳ４１２にて相関係数の分子が１４以上であり、かつ相関係数が０．９９９６以上あれば、Ｙ成分に関しては相似形と判断する（ステップＳ４１４）。

【００７９】

一方、ステップＳ４１２にて相関係数の分子が１４以上であっても、相関係数が０．９９９６より小さい場合には、相似形でないと判断して相似形チェック処理を終了し、ステップＳ１０９へもどる。

【００８０】

相関係数が１に近いほど、前記の情報が直線上に並んでいるとみなすことができる。つまり、変倍されている可能性はあるが、Ｙ成分の並びは相似形であると考える。

【００８１】

次に、これらのデータが直線上に並んでいるので、その傾きを計算する（回帰直線の傾き）。この傾きがＹ成分の変倍率δＹを示す（ステップＳ４１６）。傾き＝１の時は、識別したい帳票とマスタ帳票のＹ成分は同じであり、傾きが１より小さいと、マスタ帳票のＹ成分の増加率が識別したい帳票の比べて大きいことになり、マスタ帳票の高さが識別したい帳票の高さに比べて拡大しているといえる。一方、回帰直線の縦軸との切片がＹ座標の原点ずれ量ｓｈｉｆｔＹになる（ステップＳ４１８）。

【００８２】

ステップＳ４２０でこの変倍率δＹが、テーブルブロックのサイズまで適用できるかをチェックしている。すなわち、ステップＳ２０３で整列した帳票のブロック情報を順に１個ずつ、「（識別したい帳票のブロック高さ）／（マスタ帳票のブロック高さ）＝変倍率δＹ」を満たすかどうかをチェックし、満たさない場合には、ブロックレイアウトは相似で

ないとして当該処理を終了し、ステップＳ１０９にもどる。また、満たす場合には、変倍率δＹのテーブルブロックサイズへの適用可と判断し、処理を終了する。

【００８３】

ステップＳ４２１では、Ｘ成分、Ｙ成分両方ともに相似形であり、変倍率（δＸ、δＹ）を取得できた帳票のみ、レイアウトが相似形であると判断し、それ以外の場合には相似でないとして、当該処理を終了し、ステップＳ１０９にもどる。

【００８４】

さて、ステップＳ４０７でブロックの個数が２個以下の場合は、図１３のブロック個数が２個以下の場合の処理を行う（ステップＳ５０１）。

【００８５】

すなわち、ステップＳ５０１でブロック個数が２個以下の場合の処理が開始され、ステップＳ５０２にて識別したい帳票のページ幅とマスター帳票のページ幅を比較し、差分が１０ピクセルより大きい場合には、変倍率δＸ＝（識別したい帳票のページ幅）／（マスタ帳票のページ幅）とし、差分が１０ピクセル以内の場合には変倍率δＸ＝１．０とする（ステップＳ５０２、Ｓ５０３、Ｓ５０４）。

【００８６】

さらにステップＳ５０５にてこの変倍率δＸが、テーブルブロックのサイズまで適用できるかをチェックしている。すなわち、ステップＳ４０３で整列した帳票のブロック情報を順に１個ずつ、「（識別したい帳票のブロック幅）／（マスタ帳票のブロック幅）＝変倍率δＸ」を満たすかどうかをチェックし、満たさない場合には、ブロックレイアウトは相似でないとして当該処理を終了し、ステップＳ１０９にもどる。また、満たす場合には、変倍率δＸのテーブルブロックサイズへの適用可と判断する。

【００８７】

同様に、ステップＳ５０７にて識別したい帳票のページ高さとマスター帳票のページ高さを比較し、差分が１０ピクセルより大きい場合には、変倍率δＹ＝（識別したい帳票のページ高さ）／（マスタ帳票のページ高さ）とし、差分が１０ピクセル以内の場合には変倍率δＹ＝１．０とする（ステップＳ５０７、Ｓ５０８、Ｓ５０９）。

【００８８】

さらにステップＳ５１０にてこの変倍率δＹが、テーブルブロックのサイズまで適用できるかをチェックしている。すなわち、ステップＳ４０３で整列した帳票のブロック情報を順に１個ずつ、「（識別したい帳票のブロック高さ）／（マスタ帳票のブロック高さ）＝変倍率δＹ」を満たすかどうかをチェックし、満たさない場合には、ブロックレイアウトは相似形でないとして当該処理を終了し、ステップＳ１０９にもどる。また、満たす場合には、変倍率δＹのテーブルブロックサイズへの適用可と判断する。

【００８９】

ステップＳ５０６およびステップＳ５１１にてテーブルブロックサイズへの適用可能と判断された変倍率（δＸ、δＹ）について、ステップＳ５１２で、ブロックの左上角が、「マスタ帳票のＸ座標×変倍率（δＸ）＋原点ずれ量ｓｈｉｆｔＸ≦識別したい帳票のＸ座標±１０、かつマスタ帳票のＹ座標×変倍率（δＹ）＋原点ずれ量ｓｈｉｆｔＹ≦識別したい帳票のＹ座標±１０」の条件式を満たしていれば、識別したい帳票とマスタ帳票のレイアウトは相似形であると判断し、帳票レイアウト相似形チェック処理のステップＳ４２２へ進む（ステップＳ５１３）。また、条件を満たさない場合には、相似形でないとして当該処理を終了し、ステップＳ１０９にもどる。

【００９０】

ステップＳ４２１またはステップＳ５１３にて相似形であると判断された場合には、ステップＳ４２２でページレイアウトの変倍によるペナルティを以下の式で決定する。

【００９１】

【数３】

$$PX = \sin\left(\frac{\pi}{4} - \tan^{-1}\text{変倍率}\delta X\right) \times 60.0$$

$$PY = \sin\left(\frac{\pi}{4} - \tan^{-1}\text{変倍率}\delta Y\right) \times 60.0$$

$$PXY = \sin\left(\tan^{-1}\text{変倍率}\delta X - \tan^{-1}\text{変倍率}\delta Y\right) \times 120.0$$

ＰＸ、ＰＹは各成分の変倍によるペナルティ、ＰＸＹは、両成分の変形度によるペナルティをあらわす。

【００９２】

変倍率δＸ、δＹが１ならば、ＰＸ、ＰＹは０である。つまり、変倍していないのでペナルティを課せないことを意味する。

【００９３】

ＰＸＹは、Ｘ、Ｙ成分が均等に変倍したときには０になるし、Ｘ成分が１より大きく、Ｙ成分が１より小さく変倍するように、変倍によるレイアウトの変形が大きくなればペナルティが大きくなるように調整する式である。

【００９４】

また、原点ずれ量によるペナルティは、ＰＸ１＝原点ずれ量（ｓｈｉｆｔＸ）×０．２２、ＰＹ１＝原点ずれ量（ｓｈｉｆｔＹ）×０．２２よりもとめる。

【００９５】

レイアウトが相似形である場合は、以上の計算式でペナルティを与えて、従来のページ書式のマッチングによるペナルティを０とする。逆に、相似形でない場合は、従来通りのページ書式のマッチングによるペナルティを与える。

【００９６】

ページ書式のマッチングの次に、テーブルブロックの詳細構造、その次にテキスト・ブロックの文字比較を行うが、これらのブロックを検出する際には、相似形チェックで求めた変倍率を使用した計算式を使用する。

【００９７】

例えば、図９のような帳票ＡとＢがあり、帳票Ａが識別したい帳票、帳票Ｂがマスタ帳票と仮定する。

【００９８】

本実施形態による帳票レイアウト相似形チェック処理で、相似形だと判定され、変倍率（δＸ、δＹ）、原点ずれ量（ｓｈｉｆｔＸ、ｓｈｉｆｔＹ）が得られたとする。

【００９９】

帳票Ｂの（Ｘ１、Ｙ１）のブロックに対応する帳票Ａのブロックは、（δＸ×Ｘ１＋ｓｈｉｆｔＸ、δＹ×Ｙ１＋ｓｈｉｆｔＹ）で正確な位置を求めることができる。

【０１００】

この計算式で検出したブロックがテーブルブロックの場合は、テーブルの各罫線情報が帳票ページの変倍率と同様に変倍されているので、帳票Ｂの罫線情報（Ｌｘ、Ｌｙ）を（Ｌｘ×δＸ、Ｌｙ×δｙ）に変倍して、帳票Ａの罫線情報と比較することで、正確なテーブルブロックの詳細構造のマッチングを行うことができる。罫線情報は、テーブルブロックの左上角を原点にしているので、帳票ページ原点ずれ量は、テーブルブロックの詳細構造には影響を与えない。

【０１０１】

なお、上述した中での数値は、数多くの帳票サンプルを使用した統計値であり、帳票識別の環境によっては、変更してもかまわない。

【０１０２】

［実施形態３］

図１４に示すように、レイアウト構造が変倍されている場合には、その帳票内のテーブルの罫線情報も同じ率で変倍されている。従って、テーブルブロックの詳細構造である罫線情報の比較に、この変倍率を使用することで、より正確な詳細構造の比較を行うことができる。

【０１０３】

罫線情報をＬｉとすれば、識別したい帳票のテーブルブロックの詳細構造Ｌｉ＝（マスタ帳票のテーブルの詳細構造Ｌｉ）×変倍率（δＸ、δＹ）の関係が成り立つ。

【０１０４】

テキストブロックの位置の検出する際に、変倍率（δＸ、δＹ）を使用することで、比較すべき文字列を正確に知ることができる。しかし、文字の比較は、単なる文字コードの照合だから、変倍率（δＸ、δＹ）は不要ではあるが、ペナルティ要素として、文字の大きさを取り入れている場合には、マスタ帳票の文字の大きさに変倍率をかけることで、より正確なマッチングをおこなえる。

【０１０５】

［実施形態４］

識別したい帳票の一部分だけ文字認識を行うために、帳票認識を利用する方法がある。図１４（Ａ）、（Ｂ）に示すように、マスタ帳票に文字認識を行う領域をあらかじめ設定しているとする。図１４（Ａ）、（Ｂ）では、網掛け部分の銀行口座に登録している氏名欄が文字認識する領域である。

【０１０６】

まず、識別したい帳票を帳票認識することで、マスタ帳票のＩＤを取得できる。そのＩＤには、文字認識する領域が対応づけられている。本実施形態では、帳票のＸ、Ｙ方向の変倍率と帳票ページ原点ずれ量をＩＤとともに出力することができるので、識別したい帳票の文字認識する領域は、下記の式から修正することができる。

【０１０７】

ＩＤに対応つけてマスタに登録している文字認識領域を左上角座標（Ｘ、Ｙ）、幅Ｗ、高さＨとする。

【０１０８】

受け取った変倍率が（δＸ、δＹ）、原点ずれ量（ｓｈｉｆｔＸ、ｓｈｉｆｔＹ）であれば、識別したい帳票の文字認識領域は、左上角座標（Ｘ×δＸ＋ｓｈｉｆｔＸ、Ｙ×δＹ＋ｓｈｉｆｔＹ）、幅（Ｗ×δＸ）、高さ（Ｈ×δＹ）となる。

【０１０９】

【他の実施形態】

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憶媒体を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記憶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。

【０１１０】

この場合、記憶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憶媒体は本発明を構成することになる。

【０１１１】

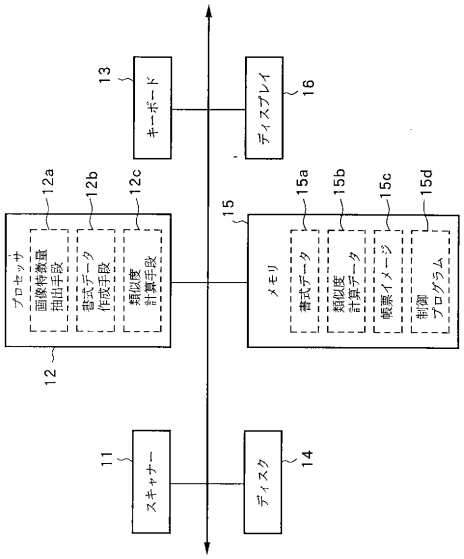
プログラムコードを供給するための記憶媒体としては、例えば、フロッピディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、磁気テープ、不揮発性のメモリカード、ＲＯＭなどを用いることができる。

【０１１２】

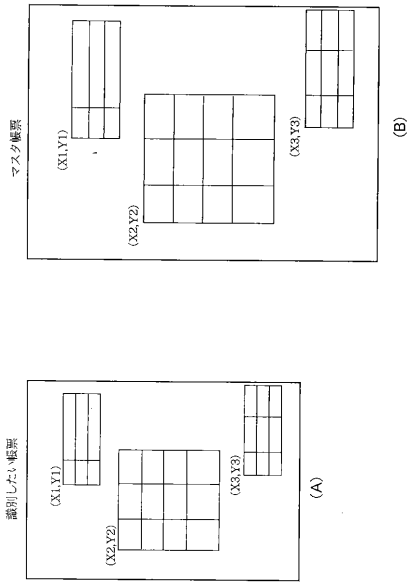
また、コンピュータが読出したプログラムコードを実行することにより、前述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているＯＳ（オペレーティングシステム）などが実際の処理の一部または全部を行い、その処理によって前述した実施形態の機能が実現される場合も含まれることは言

うまでもない。
【０１１３】
さらに、記憶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または全部を行い、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。
【０１１４】
【発明の効果】
以上説明したように、本発明によれば、異なる変倍率で拡大または縮小された複数の文書が混在する環境でも文書書式を正しく識別することができる。
【図面の簡単な説明】
【図１】本発明の実施の一形態に係わる帳票識別装置の概略構成を示すブロック図である。
【図２】本発明の相似形帳票の一例を示す図である。
【図３】本発明のマッチング対象となる帳票の一例を示す図である。
【図４】本発明の処理の概要を示すフローチャートである。
【図５】本発明の相似形チェック処理を示すフローチャートである。
【図６】本発明の相似形チェック処理で、ブロック左上角Ｘ座標の比較処理を示すフローチャートである。
【図７】本発明の相似形チェック処理で、ブロック左上角Ｙ座標の比較処理を示すフローチャートである。
【図８】本発明の相似形チェック処理で、ブロック個数が２個以下の場合の処理を示すフローチャートである。
【図９】本発明の相似形帳票の一例を示す図である。
【図１０】本発明の相似形チェック処理を示すフローチャートである。
【図１１】本発明の相似形チェック処理で、ブロック左上角Ｘ座標の比較処理を示すフローチャートである。
【図１２】本発明の相似形チェック処理で、ブロック左上角Ｙ座標の比較処理を示すフローチャートである。
【図１３】本発明の相似形チェック処理で、ブロック個数が２個以下の場合の処理を示すフローチャートである。
【図１４】本発明の相似形帳票の一例を示す図である。

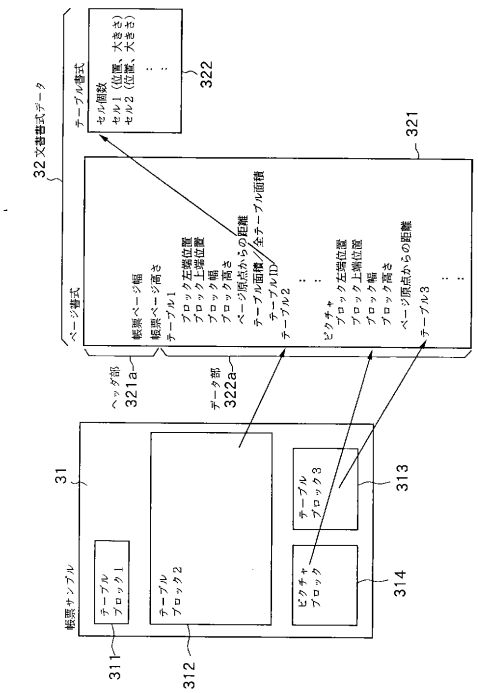
【図１】


プロセッサ
画像特徴量抽出手段
書式データ作成手段
類似度計算手段
メモリ
書式データ
類似度計算データ
帳票イメージ
制御プログラム
キーボード
ディスプレイ
スキャナー
ディスク
12
12a
12b
12c
15
15a
15b
15c
15d
11
13
14
16

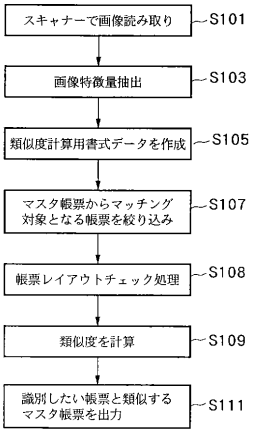
【図２】


マスタ帳票
(X1,Y1)
(X2,Y2)
(X3,Y3)
(B)
識別したい帳票
(X1,Y1)
(X2,Y2)
(X3,Y3)
(A)

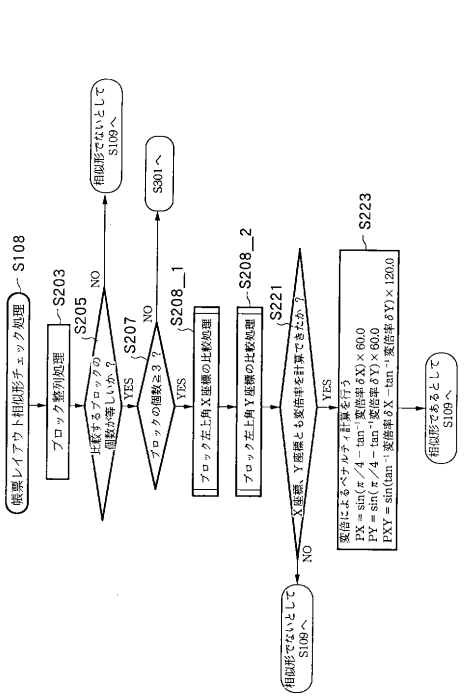
【図３】


32 文書書式データ
ページ書式
帳票ページ幅
帳票ページ高さ
テーブル1
ブロック左端位置
ブロック上端位置
ブロック幅
ブロック高さ
ページ原点からの距離
テーブル面積／全テーブル面積
テーブルID
テーブル2
ピクチャ
ブロック左端位置
ブロック上端位置
ブロック幅
ブロック高さ
ページ原点からの距離
テーブル3
テーブル書式
セル個数
セル1（位置、大きさ）
セル2（位置、大きさ）
322
321
ヘッダ部
321a
データ部
322a
帳票サンプル
テーブル ブロック1
テーブル ブロック2
テーブル ブロック3
ピクチャ ブロック
31
311
312
313
314

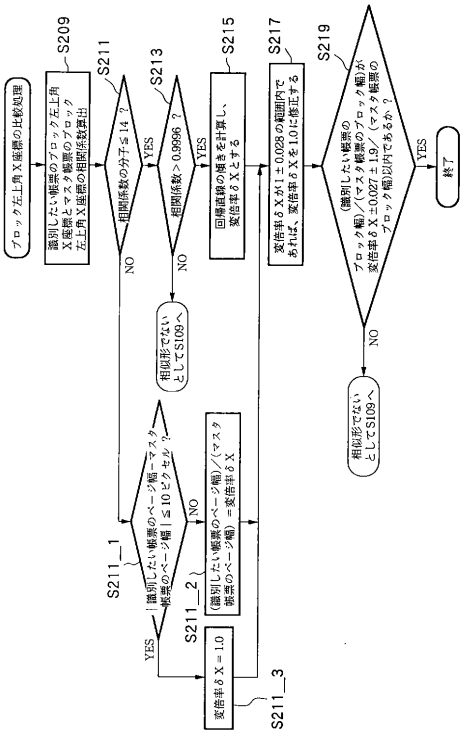
【図４】


スキャナーで画像読み取り
S101
画像特徴量抽出
S103
類似度計算用書式データを作成
S105
マスタ帳票からマッチング対象となる帳票を絞り込み
S107
帳票レイアウトチェック処理
S108
類似度を計算
S109
識別したい帳票と類似するマスタ帳票を出力
S111

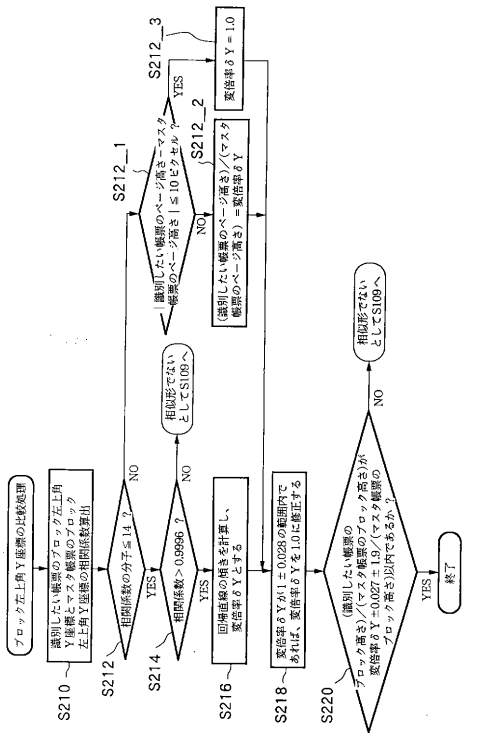
【図５】


帳票レイアウト相似形チェック処理
S108
ブロック整列処理
S203
比較するブロックの個数が等しいか？
S205
NO
相似形でないとしてS109へ
YES
ブロックの個数≧3？
S207
NO
S301へ
YES
ブロック左上角X座標の比較処理
S208_1
ブロック左上角Y座標の比較処理
S208_2
X座標、Y座標とも変倍率を計算できたか？
S221
NO
相似形でないとしてS109へ
YES
変倍率δX＝tan⁻¹変倍率δX)×60.0
PY＝sin(π／4－tan⁻¹変倍率δY)×60.0
PXY＝sin(tan⁻¹変倍率δX－tan⁻¹変倍率δY)×120.0
S223
相似形であるとしてS109へ

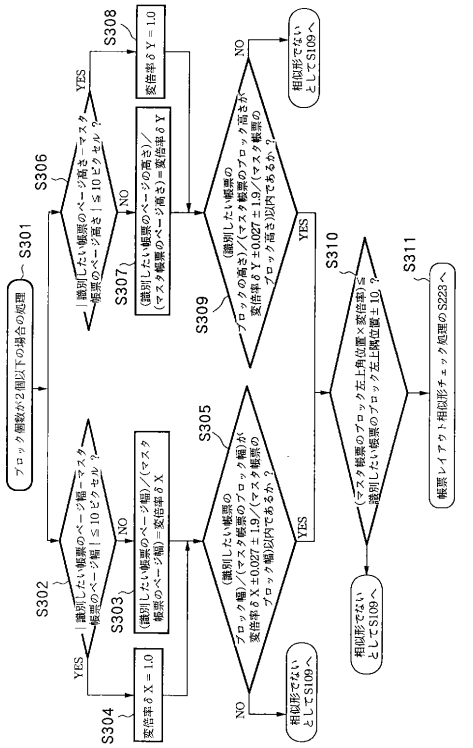
【図６】


ブロック左上角X座標の比較処理
識別したい帳票のブロック左上角X座標とマスタ帳票のブロック左上角X座標の相関係数算出
S209
相関係数の分子≦14？
S211
NO
YES
相関係数＞0.9996？
S213
NO
相似形でないとしてS109へ
YES
回帰直線の傾きを計算し、変倍率δXとする
S215
変倍率δXが1±0.028の範囲内であれば、変倍率δXを1.0に修正する
S217
(識別したい帳票のブロック幅)／(マスタ帳票のブロック幅)が変倍率δX±0.027±1.9／(マスタ帳票のブロック幅)以内であるか？
S219
NO
相似形でないとしてS109へ
YES
終了
識別したい帳票のページ幅－マスタ帳票のページ幅｜≦10ピクセル？
S211_1
NO
YES
(識別したい帳票のページ幅)／(マスタ帳票のページ幅)＝変倍率δX
S211_2
変倍率δX＝1.0
S211_3

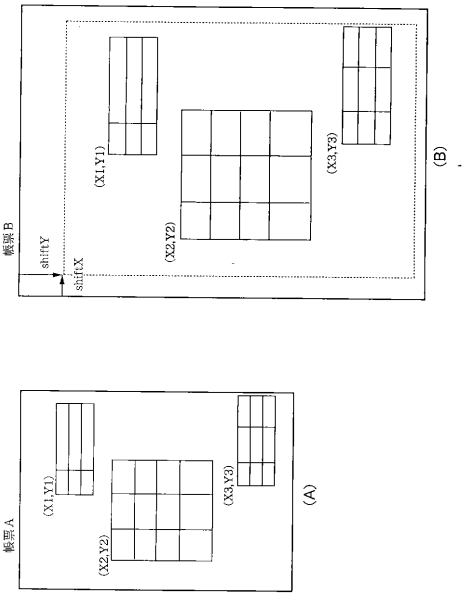
【図７】


ブロック左上角Y座標の比較処理
識別したい帳票のブロック左上角Y座標とマスタ帳票のブロック左上角Y座標の相関係数算出
S210
相関係数の分子≦14？
S212
NO
YES
相関係数＞0.9996？
S214
NO
相似形でないとしてS109へ
YES
回帰直線の傾きを計算し、変倍率δYとする
S216
変倍率δYが1±0.028の範囲内であれば、変倍率δYを1.0に修正する
S218
(識別したい帳票のブロック高さ)／(マスタ帳票のブロック高さ)が変倍率δY±0.027±1.9／(マスタ帳票のブロック高さ)以内であるか？
S220
NO
相似形でないとしてS109へ
YES
終了
識別したい帳票のページ高さ－マスタ帳票のページ高さ｜≦10ピクセル？
S212_1
NO
YES
(識別したい帳票のページ高さ)／(マスタ帳票のページ高さ)＝変倍率δY
S212_2
変倍率δY＝1.0
S212_3

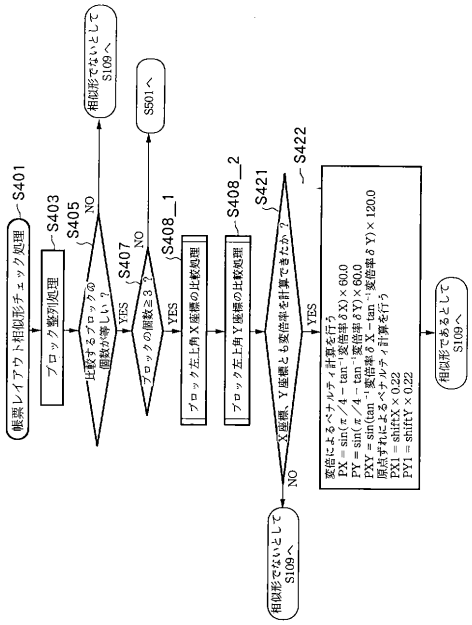
【図８】


ブロック個数が2個以下の場合の処理
S301
識別したい帳票のページ幅－マスタ帳票のページ幅｜≦10ピクセル？
S302
YES
NO
(識別したい帳票のページ幅)／(マスタ帳票のページ幅)＝変倍率δX
S303
変倍率δX＝1.0
S304
識別したい帳票のページ高さ－マスタ帳票のページ高さ｜≦10ピクセル？
S306
YES
NO
(識別したい帳票のページ高さ)／(マスタ帳票のページ高さ)＝変倍率δY
S307
変倍率δY＝1.0
S308
(識別したい帳票のブロック幅)／(マスタ帳票のブロック幅)が変倍率δX±0.027±1.9／(マスタ帳票のブロック幅)以内であるか？
S305
NO
YES
相似形でないとしてS109へ
(識別したい帳票のブロック高さ)／(マスタ帳票のブロック高さ)が変倍率δY±0.027±1.9／(マスタ帳票のブロック高さ)以内であるか？
S309
NO
YES
相似形でないとしてS109へ
(マスタ帳票のブロック左上角位置×変倍率)≦識別したい帳票のブロック左上角位置±1.0？
S310
NO
相似形でないとしてS109へ
YES
S311
帳票レイアウト相似形チェック処理のS223へ

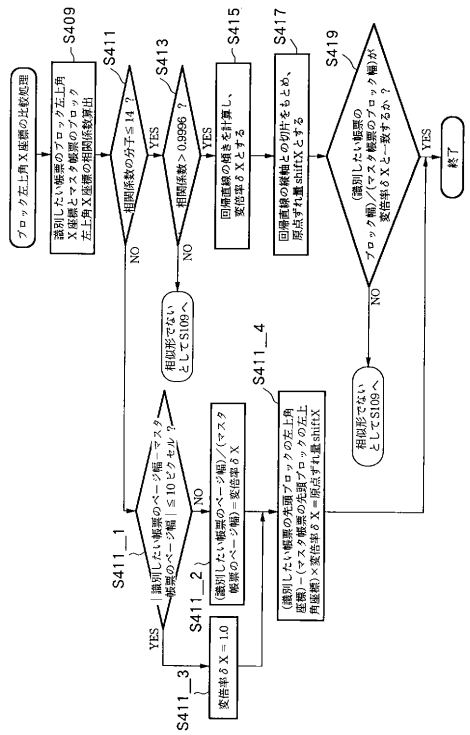
【図９】



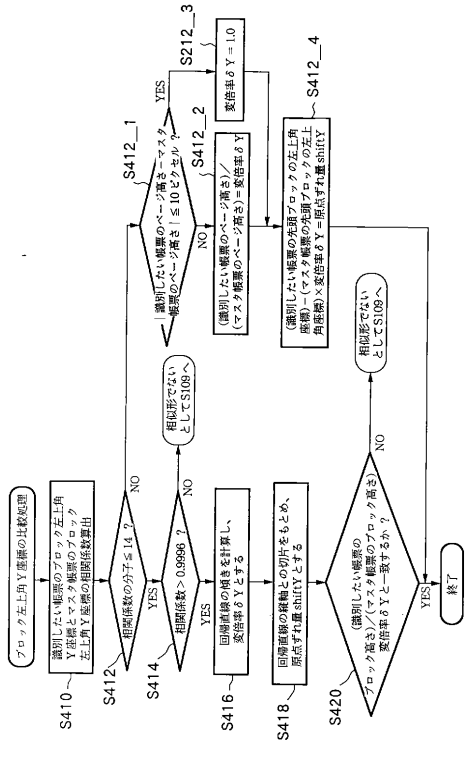
【図１０】



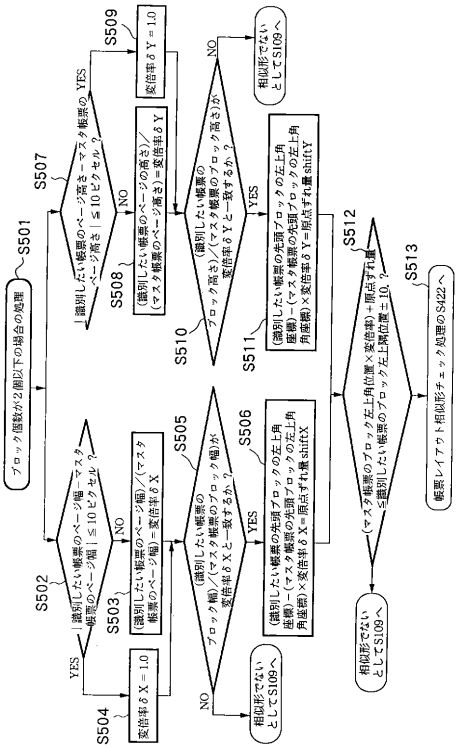
【図１１】



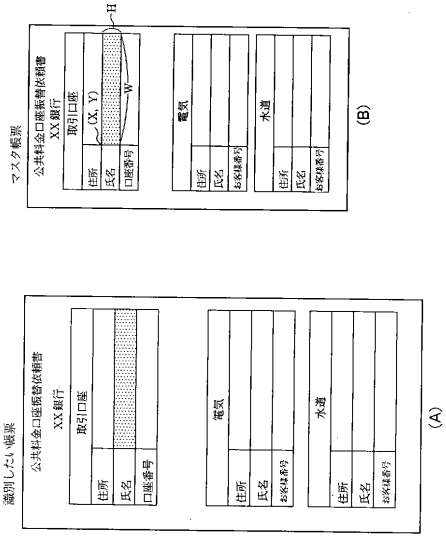
【図１２】



【図１３】


ブロック個数が２個以下の場合の処理
S501
S502
識別したい帳票のページ幅 − マスタ帳票のページ幅 | ≦ 10ピクセル ?
YES
NO
S503
(識別したい帳票のページ幅)／(マスタ帳票のページ幅) = 変倍率δX
S504
変倍率δX = 1.0
S505
(識別したい帳票のブロック幅)／(マスタ帳票のブロック幅) の変倍率δXと一致するか ?
YES
NO
S506
(識別したい帳票の先頭ブロックの左上角座標) − (マスタ帳票の先頭ブロックの左上角座標) × 変倍率δX = 原点ずれ量 shiftX
相似形でないとしてS109へ
S507
識別したい帳票のページ高さ − マスタ帳票のページ高さ | ≦ 10ピクセル ?
YES
NO
S508
(識別したい帳票のページ高さ)／(マスタ帳票のページ高さ) = 変倍率δY
S509
変倍率δY = 1.0
S510
(識別したい帳票のブロック高さ)／(マスタ帳票のブロック高さ) の変倍率δYと一致するか ?
YES
NO
S511
(識別したい帳票の先頭ブロックの左上角座標) − (マスタ帳票の先頭ブロックの左上角座標) × 変倍率δY = 原点ずれ量 shiftY
相似形でないとしてS109へ
S512
(マスタ帳票のブロック左上角位置 × 変倍率) + 原点ずれ量 ≦ 識別したい帳票のブロック左上端位置 ± 10 ?
YES
NO
相似形でないとしてS109へ
S513
帳票レイアウト相似形チェック処理のS422へ

【図１４】


識別したい帳票
公共料金口座振替依頼書
XX銀行
取引口座
住所
氏名
口座番号
電気
住所
氏名
お客様番号
水道
住所
氏名
お客様番号
(A)
マスタ帳票
公共料金口座振替依頼書
XX銀行
取引口座
(X, Y)
H
W
住所
氏名
口座番号
電気
住所
氏名
お客様番号
水道
住所
氏名
お客様番号
(B)

フロントページの続き

(72)発明者　数見　健一
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　佐藤　実

(56)参考文献　特開平１０－０２７２０８（ＪＰ，Ａ）
特開昭６４－００５１４１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06K　9/00～9/76