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

特許第4885112号
(P4885112)

(45)発行日　平成24年2月29日(2012.2.29)　(24)登録日　平成23年12月16日(2011.12.16)

(51)Int.Cl.　F I
G06T 7/00 (2006.01)　G06T 7/00 300F
G06F 17/30 (2006.01)　G06F 17/30 170B
G06F 17/30 340B

請求項の数 18　(全 24 頁)

(21)出願番号　特願2007-293392 (P2007-293392)
(22)出願日　平成19年11月12日 (2007.11.12)
(65)公開番号　特開2009-122758 (P2009-122758A)
(43)公開日　平成21年6月4日 (2009.6.4)
審査請求日　平成22年6月3日 (2010.6.3)

(73)特許権者　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(74)代理人　100089118
弁理士　酒井　宏明
(72)発明者　大黒　慶久
東京都大田区中馬込１丁目３番６号　株式会社リコー内

審査官　鹿野　博嗣

(56)参考文献　特開２００５－２４２５７９（ＪＰ，Ａ）
特開２００３－２０８４３３（ＪＰ，Ａ）

最終頁に続く

(54)【発明の名称】文書処理装置、文書処理方法及び文書処理プログラム

(57)【特許請求の範囲】
【請求項１】
文書画像間の照合を行う文書処理装置において、
前記文書画像に含まれた文字画像毎の外接矩形に基づいて、当該外接矩形を連結した文字行を切り出す文字行切出手段と、
前記文字行内における前記外接矩形の特性を表す配置情報を固定段階に量子化する量子化手段と、
前記量子化された配置情報の各々を固定種類のシンボルにシンボル化するシンボル生成手段と、
所定個の前記シンボルの組合せからなるシンボル系列の出現頻度を算出する出現頻度算出手段と、
照合対象の文書画像と、当該文書画像の被照合対象となる複数の文書画像とについて、前記出現頻度算出手段により算出された出現頻度を照合し、より高い相関を有した被照合対象の文書画像を所定数選定する被照合対象選定手段と、
前記照合対象の文書画像と、前記被照合対象選定手段により選定された被照合対象の文書画像の各々とで一致した前記シンボル系列に対応する各配置情報に基づいて、当該各配置情報の何れか又は全てが表す外接矩形の出現位置の分布状態を文書画像毎に導出する分布状態導出手段と、
前記分布状態導出手段により導出された前記照合対象の文書画像についての分布状態と、前記被照合対象の文書画像についての分布状態との類似度を判定し、最も高い類似度を

有した被照合対象の文書画像を照合結果として選定する照合結果選定手段と、
を備えたことを特徴とする文書処理装置。
【請求項２】
前記分布状態導出手段は、前記文書画像の水平方向及び／又は垂直方向について、前記外接矩形の出現位置の分布状態を導出することを特徴とする請求項１に記載の文書処理装置。
【請求項３】
前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を度数分布ヒストグラムとして導出することを特徴とする請求項１又は２に記載の文書処理装置。
【請求項４】
前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を正規分布とみなし、当該正規分布の平均、標準偏差、歪度及び尖度を導出することを特徴とする請求項１又は２に記載の文書処理装置。
【請求項５】
前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とにおける前記文字行に含まれた各文字画像の前記外接矩形のサイズを集計し、当該サイズの平均値又は最頻値により前記正規分布を規定する数値を正規化することを特徴とする請求項４に記載の文書処理装置。
【請求項６】
前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とにおいて一致した前記シンボル系列に対応する配置情報が表す外接矩形のサイズを集計することを特徴とする請求項５に記載の文書処理装置。
【請求項７】
文書画像間の照合を行う文書処理装置で実行される文書処理方法であって、
文字行切出手段が、前記文書画像に含まれた文字画像毎の外接矩形に基づいて、当該外接矩形を連結した文字行を切り出す文字行切出ステップと、
量子化手段が、前記文字行内における前記外接矩形の特性を表す配置情報を固定段階に量子化する量子化ステップと、
シンボル系列生成手段が、前記量子化された配置情報の各々を固定種類のシンボルにシンボル化するシンボル生成ステップと、
出現頻度算出手段が、所定個の前記シンボルの組合せからなるシンボル系列の出現頻度を算出する出現頻度算出ステップと、
被照合対象選定手段が、照合対象の文書画像と、当該文書画像の被照合対象となる複数文書画像とについて、前記出現頻度算出手段により算出された出現頻度を照合し、より高い相関を有した被照合対象の文書画像を所定数選定する被照合対象選定ステップと、
分布状態導出手段が、前記照合対象の文書画像と、前記被照合対象選定ステップで選定された被照合対象の文書画像の各々とで一致した前記シンボル系列に対応する各配置情報に基づいて、当該各配置情報の何れか又は全てが表す外接矩形の出現位置の分布状態を文書画像毎に導出する分布状態導出ステップと、
照合結果選定手段が、前記分布状態導出ステップで導出された前記照合対象の文書画像についての分布状態と、前記被照合対象の文書画像についての分布状態との類似度を判定し、最も高い類似度を有した被照合対象の文書画像を照合結果として選定する照合結果選定ステップと、
を含むことを特徴とする文書処理方法。
【請求項８】
前記分布状態導出手段は、前記分布状態導出ステップにおいて、前記文書画像の水平方向及び／又は垂直方向について、前記外接矩形の出現位置の分布状態を導出することを特徴とする請求項７に記載の文書処理方法。
【請求項９】
前記分布状態導出手段は、前記分布状態導出ステップにおいて、前記外接矩形の出現位

置の分布状態を度数分布ヒストグラムとして導出することを特徴とする請求項７又は８に記載の文書処理方法。

【請求項１０】

前記分布状態導出手段は、前記分布状態導出ステップにおいて、前記外接矩形の出現位置の分布状態を正規分布とみなし、当該正規分布の平均、標準偏差、歪度及び尖度を導出することを特徴とする請求項７又は８に記載の文書処理方法。

【請求項１１】

前記分布状態導出手段は、前記分布状態導出ステップにおいて、前記照合対象の文書画像と、前記被照合対象の文書画像とにおける前記文字行に含まれた各文字画像の前記外接矩形のサイズを集計し、当該サイズの平均値又は最頻値により前記正規分布を規定する数値を正規化することを特徴とする請求項１０に記載の文書処理方法。

【請求項１２】

前記分布状態導出手段は、前記分布状態導出ステップにおいて、前記照合対象の文書画像と、前記被照合対象の文書画像とにおいて一致した前記シンボル系列に対応する配置情報が表す外接矩形のサイズを集計することを特徴とする請求項１１に記載の文書処理方法。

【請求項１３】

文書画像間の照合を行うコンピュータを、

前記文書画像に含まれた文字画像毎の外接矩形に基づいて、当該外接矩形を連結した文字行を切り出す文字行切出手段と、

前記文字行内における前記外接矩形の特性を表す配置情報を固定段階に量子化する量子化手段と、

前記量子化された配置情報の各々を固定種類のシンボルにシンボル化するシンボル生成手段と、

前記シンボル系列内における、所定個のシンボルの組合せからなるシンボル系列の出現頻度を算出する出現頻度算出手段と、

照合対象の文書画像と、当該文書画像の被照合対象となる複数の文書画像とについて、前記出現頻度算出手段により算出された出現頻度を照合し、より高い相関を有した被照合対象の文書画像を所定数選定する被照合対象選定手段と、

前記照合対象の文書画像と、前記被照合対象選定手段により選定された被照合対象の文書画像の各々とで一致した前記シンボル系列に対応する各配置情報に基づいて、当該各配置情報の何れか又は全てが表す外接矩形の出現位置の分布状態を文書画像毎に導出する分布状態導出手段と、

前記分布状態導出手段により導出された前記照合対象の文書画像についての分布状態と、前記被照合対象の文書画像についての分布状態との類似度を判定し、最も高い類似度を有した被照合対象の文書画像を照合結果として選定する照合結果選定手段と、

して機能させることを特徴とする文書処理プログラム。

【請求項１４】

前記分布状態導出手段は、前記文書画像の水平方向及び／又は垂直方向について、前記外接矩形の出現位置の分布状態を導出することを特徴とする請求項１３に記載の文書処理プログラム。

【請求項１５】

前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を度数分布ヒストグラムとして導出することを特徴とする請求項１３又は１４に記載の文書処理プログラム。

【請求項１６】

前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を正規分布とみなし、当該正規分布の平均、標準偏差、歪度及び尖度を導出することを特徴とする請求項１３又は１４に記載の文書処理プログラム。

【請求項１７】

前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とに

おける前記文字行に含まれた各文字画像の前記外接矩形のサイズを集計し、当該サイズの平均値又は最頻値により前記正規分布を規定する数値を正規化することを特徴とする請求項１６に記載の文書処理プログラム。

【請求項１８】

前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とにおいて一致した前記シンボル系列に対応する配置情報が表す外接矩形のサイズを集計することを特徴とする請求項１７に記載の文書処理プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、文書画像間の照合を行う文書処理装置、文書処理方法および文書処理プログラムに関する。

【背景技術】

【０００２】

従来、文字列が画像（文字行）として記録された文書画像中から文字列を抽出する方法として、種々の技術が提案されている。例えば、文書画像に含まれた文字行に外接する矩形の形状及び位置に関する特徴（大きさ、間隔等）について、複数の制約を適用することにより文字行を文字列として認識することが可能な技術が提案されている（例えば、特許文献１、２参照）。

【０００３】

【特許文献１】特開平１１－２１９４０７号公報

【特許文献２】国際公開第００／６２２４３号パンフレット

【発明の開示】

【発明が解決しようとする課題】

【０００４】

しかしながら、特許文献１、２に記載の技術では、文字行の認識を精度よく行うために外接矩形に関数する複数の制約を人手によって最適値に調整する必要がある。また、文字行らしさを判定することはできるものの、文字行の内容に関する特徴を認識することはできないため、文書画像間の照合に用いたとしても十分な精度を得ることができない可能性がある。また、複数行間の相対的な位置関係の利用については何等言及されていないため、文書画像の一部分となる部分画像を照合対象とした場合には、対応することができないという問題がある。

【０００５】

本発明は、上記に鑑みてなされたものであって、文書画像間の類似性をより効率的且つ高精度に判定することが可能な文書処理装置、文書処理方法及び文書処理プログラムを提供することを目的とする。

【課題を解決するための手段】

【０００６】

上述した課題を解決し、目的を達成するために、請求項１に係る発明は、文書画像間の照合を行う文書処理装置において、前記文書画像に含まれた文字画像毎の外接矩形に基づいて、当該外接矩形を連結した文字行を切り出す文字行切出手段と、前記文字行内における前記外接矩形の特性を表す配置情報を固定段階に量子化する量子化手段と、前記量子化された配置情報の各々を固定種類のシンボルにシンボル化するシンボル生成手段と、所定個の前記シンボルの組合せからなるシンボル系列の出現頻度を算出する出現頻度算出手段と、照合対象の文書画像と、当該文書画像の被照合対象となる複数の文書画像とについて、前記出現頻度算出手段により算出された出現頻度を照合し、より高い相関を有した被照合対象の文書画像を所定数選定する被照合対象選定手段と、前記照合対象の文書画像と、前記被照合対象選定手段により選定された被照合対象の文書画像の各々とで一致した前記シンボル系列に対応する各配置情報に基づいて、当該各配置情報の何れか又は全てが表す外接矩形の出現位置の分布状態を文書画像毎に導出する分布状態導出手段と、前記分布状

態導出手段により導出された前記照合対象の文書画像についての分布状態と、前記被照合対象の文書画像についての分布状態との類似度を判定し、最も高い類似度を有した被照合対象の文書画像を照合結果として選定する照合結果選定手段と、を備えたことを特徴とする。

【０００７】

また、請求項２に係る発明は、請求項１に係る発明において、前記分布状態導出手段は、前記文書画像の水平方向及び／又は垂直方向について、前記外接矩形の出現位置の分布状態を導出することを特徴とする。

【０００８】

また、請求項３に係る発明は、請求項１又は２に係る発明において、前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を度数分布ヒストグラムとして導出することを特徴とする。

【０００９】

また、請求項４に係る発明は、請求項１又は２に係る発明において、前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を正規分布とみなし、当該正規分布の平均、標準偏差、歪度及び尖度を導出することを特徴とする。

【００１０】

また、請求項５に係る発明は、請求項４に係る発明において、前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とにおける前記文字行に含まれた各文字画像の前記外接矩形のサイズを集計し、当該サイズの平均値又は最頻値により前記正規分布を規定する数値を正規化することを特徴とする。

【００１１】

また、請求項６に係る発明は、請求項５に係る発明において、前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とにおいて一致した前記シンボル系列に対応する配置情報が表す外接矩形のサイズを集計することを特徴とする。

【００１２】

また、請求項７に係る発明は、文書画像間の照合を行う文書処理装置で実行される文書処理方法であって、文字行切出手段が、前記文書画像に含まれた文字画像毎の外接矩形に基づいて、当該外接矩形を連結した文字行を切り出す文字行切出ステップと、量子化手段が、前記文字行内における前記外接矩形の特性を表す配置情報を固定段階に量子化する量子化ステップと、シンボル系列生成手段が、前記量子化された配置情報の各々を固定種類のシンボルにシンボル化するシンボル生成ステップと、出現頻度算出手段が、所定個の前記シンボルの組合せからなるシンボル系列の出現頻度を算出する出現頻度算出ステップと、被照合対象選定手段が、照合対象の文書画像と、当該文書画像の被照合対象となる複数文書画像とについて、前記出現頻度算出手段により算出された出現頻度を照合し、より高い相関を有した被照合対象の文書画像を所定数選定する被照合対象選定ステップと、分布状態導出手段が、前記照合対象の文書画像と、前記被照合対象選定ステップで選定された被照合対象の文書画像の各々とで一致した前記シンボル系列に対応する各配置情報に基づいて、当該各配置情報の何れか又は全てが表す外接矩形の出現位置の分布状態を文書画像毎に導出する分布状態導出ステップと、照合結果選定手段が、前記分布状態導出ステップで導出された前記照合対象の文書画像についての分布状態と、前記被照合対象の文書画像についての分布状態との類似度を判定し、最も高い類似度を有した被照合対象の文書画像を照合結果として選定する照合結果選定ステップと、を含むことを特徴とする。

【００１３】

また、請求項８に係る発明は、請求項７に係る発明において、前記分布状態導出手段は、前記文書画像の水平方向及び／又は垂直方向について、前記外接矩形の出現位置の分布状態を導出することを特徴とする。

【００１４】

また、請求項９に係る発明は、請求項７又は８に係る発明において、前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を度数分布ヒストグラムとして導出すること

を特徴とする。

【００１５】

また、請求項１０に係る発明は、請求項７又は８に係る発明において、前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を正規分布とみなし、当該正規分布の平均、標準偏差、歪度及び尖度を導出することを特徴とする。

【００１６】

また、請求項１１に係る発明は、請求項１０に係る発明において、前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とにおける前記文字行に含まれた各文字画像の前記外接矩形のサイズを集計し、当該サイズの平均値又は最頻値により前記正規分布を規定する数値を正規化することを特徴とする。

【００１７】

また、請求項１２に係る発明は、請求項１１に係る発明において、前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とにおいて一致した前記シンボル系列に対応する配置情報が表す外接矩形のサイズを集計することを特徴とする。

【００１８】

また、請求項１３に係る発明は、文書画像間の照合を行うコンピュータを、前記文書画像に含まれた文字画像毎の外接矩形に基づいて、当該外接矩形を連結した文字行を切り出す文字行切出手段と、前記文字行内における前記外接矩形の特性を表す配置情報を固定段階に量子化する量子化手段と、前記量子化された配置情報の各々を固定種類のシンボルにシンボル化するシンボル生成手段と、前記シンボル系列内における、所定個のシンボルの組合せからなるシンボル系列の出現頻度を算出する出現頻度算出手段と、照合対象の文書画像と、当該文書画像の被照合対象となる複数の文書画像とについて、前記出現頻度算出手段により算出された出現頻度を照合し、より高い相関を有した被照合対象の文書画像を所定数選定する被照合対象選定手段と、前記照合対象の文書画像と、前記被照合対象選定手段により選定された被照合対象の文書画像の各々とで一致した前記シンボル系列に対応する各配置情報に基づいて、当該各配置情報の何れか又は全てが表す外接矩形の出現位置の分布状態を文書画像毎に導出する分布状態導出手段と、前記分布状態導出手段により導出された前記照合対象の文書画像についての分布状態と、前記被照合対象の文書画像についての分布状態との類似度を判定し、最も高い類似度を有した被照合対象の文書画像を照合結果として選定する照合結果選定手段と、して機能させることを特徴とする。

【００１９】

また、請求項１４に係る発明は、請求項１３に係る発明において、前記分布状態導出手段は、前記文書画像の水平方向及び／又は垂直方向について、前記外接矩形の出現位置の分布状態を導出することを特徴とする。

【００２０】

また、請求項１５に係る発明は、請求項１３又は１４に係る発明において、前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を度数分布ヒストグラムとして導出することを特徴とする。

【００２１】

また、請求項１６に係る発明は、請求項１３又は１４に係る発明において、前記分布状態導出手段は、前記外接矩形の出現位置の分布状態を正規分布とみなし、当該正規分布の平均、標準偏差、歪度及び尖度を導出することを特徴とする。

【００２２】

また、請求項１７に係る発明は、請求項１６に係る発明において、前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とにおける前記文字行に含まれた各文字画像の前記外接矩形のサイズを集計し、当該サイズの平均値又は最頻値により前記正規分布を規定する数値を正規化することを特徴とする。

【００２３】

また、請求項１８に係る発明は、請求項１７に係る発明において、前記分布状態導出手段は、前記照合対象の文書画像と、前記被照合対象の文書画像とにおいて一致した前記シ

ンボル系列に対応する配置情報が表す外接矩形のサイズを集計することを特徴とする。
【発明の効果】
【００２４】
本発明によれば、照合対象の文書画像と被照合対象の文書画像とについて、文字行内における外接矩形の特徴を表した配置情報を抽出し、これらを固定段階に量子化してシンボルを生成することにより、文字認識することなく文字行の特徴の抽出が可能となり、被照合対象の文書画像から、照合対象の文書画像と相関の高い被照合対象の文書画像を所定の数だけ選定することができる。また、照合対象の文書画像と、選定された被照合対象の文書画像とについて、一致するシンボル系列の出現位置の分布状態を照合することで、当該シンボル系列の相対的な位置関係の類似性を判定することができるため、照合対象の文書画像と被検照合対象の文書画像との類似性を高精度に判定することができる。これにより、文書画像中の部分画像が照合対象の文書画像とされた場合であっても、この部分画像に含まれた文字画像の外接矩形の位置関係に基づいて、当該部分画像と類似する文書画像を高精度に検索することが可能となる。
【発明を実施するための最良の形態】
【００２５】
以下に添付図面を参照して、本発明に係る文書処理装置、文書処理方法及び文書処理プログラムの最良な実施の形態を詳細に説明する。
【００２６】
（文書処理装置のハードウェア構成）
図１は、本発明の第１の実施形態にかかる文書処理装置１００のハードウェア構成を示したブロック図である。図１に示したように、文書処理装置１００は、ＰＣ（Personal Computer）などのコンピュータであり、文書処理装置１００の各部を制御するＣＰＵ（Central Processing Unit）１、ＣＰＵ１を起動するためのプログラムが記憶されるＲＯＭ（Read Only Memory）２、後述する画像入力部２１により入力された文書画像やオペレーティングシステム、種々のプログラム等を記憶するハードディスク３、ＣＰＵ１のワークエリアとして機能するＲＡＭ（Random Access Memory）４、オペレータからの各種入力を受け付けるキーボード５、入力状況等を表示する表示装置６、ＣＤ－ＲＯＭなどの各種光情報記録メディア（図示せず）に記憶されたプログラム等を読み取る光ディスクドライブ７、インターネットやＬＡＮ（Local Area Network）等の電気通信回線を介して文書画像を送受信する通信装置８、原稿画像の光学的な読み取りを行うスキャナ９等から構成されており、これらの各部間で入出力されるデータをバスコントローラ１０が調停して動作する。
【００２７】
文書処理装置１００では、オペレータが電源を投入するとＣＰＵ１がＲＯＭ２内のローダーというプログラムを起動させ、ハードディスク３よりオペレーティングシステムというコンピュータのハードウェアとソフトウェアとを管理するプログラムをＲＡＭ４に読み込み、このオペレーティングシステムを起動させる。このようなオペレーティングシステムは、オペレータの操作に応じてプログラムを起動したり、情報を読み込んだり、保存を行ったりする。オペレーティングシステムのうち代表的なものとしては、Ｗｉｎｄｏｗｓ（登録商標）、ＵＮＩＸ（登録商標）等が知られている。これらのオペレーティングシステム上で走る動作プログラムをアプリケーションプログラムと呼んでいる。
【００２８】
ここで、文書処理装置１００は、ＣＰＵ１が実行するプログラムとして、後述する文書照合処理）にかかる文書処理プログラムをハードディスク３に記憶している。この意味で、ハードディスク３は、文書処理プログラムを記憶する記憶媒体として機能する。
【００２９】
また、一般的には、文書処理装置１００のハードディスク３にインストールされるプログラムは、ＣＤ－ＲＯＭなどの各種光情報記録メディアやＦＤ等の磁気メディア等の記憶媒体に記録され、この記憶媒体に記録されたプログラムがハードディスク３にインストー

ルされる。このため、ＣＤ－ＲＯＭなどの各種光情報記録メディアやＦＤ等の磁気メディア等の可搬性を有する記憶媒体も、文書処理プログラムを記憶する記憶媒体となり得る。さらには、文書処理プログラムは、例えば通信装置８を介して外部から取り込まれ、ハードディスク３にインストールされても良い。

【００３０】

ＣＰＵ１は、オペレーティングシステム上で動作する文書処理プログラムが起動すると、この文書処理プログラムとの協働により後述する各機能部を実現させる。以下、文書処理装置１００の機能的構成について説明する。

【００３１】

（文書処理装置の機能的構成）

図２は、文書処理装置１００の機能的構成を示したブロック図である。図２に示したように、文書処理装置１００は機能部として、画像入力部２１、照合画像選択部２２、矩形抽出部２３、行切出部２４、量子化部２５、シンボル生成部２６、出現頻度集計部２７、候補画像選定部２８、出現位置分布導出部２９、照合結果選定部３０及び表示部３１を含み構成される。

【００３２】

画像入力部２１は、外部から入力される文書画像を受け付け、ハードディスク３に記憶する。具体的に、画像入力部２１の機能は、図１に示した光ディスクドライブ７、通信装置８、スキャナ９により実現することができる。

【００３３】

照合画像選択部２２は、画像入力部２１から入力される文書画像や、キーボード５を介して指定されたハードディスク３に記憶された文書画像を、照合対象の文書画像として選択する。以下、照合対象の文書画像を「照合画像」という。なお、照合画像選択部２２は、文書画像中の特定の領域がキーボード５を介して指定された場合には、この領域内に含まれる部分的な文書画像（部分画像）を照合画像として選択するものとする。

【００３４】

また、照合画像選択部２２は、照合画像の照合先となる被照合対象の文書画像を選択する。ここで、被照合対象の文書画像は、例えば、ハードディスク３に予め記憶された一部又は全ての文書画像としてもよいし、キーボード５を介して指定された文書画像を被照合対象の文書画像としてもよい。以下、被照合対象の文書画像を「被照合画像」という。

【００３５】

矩形抽出部２３は、文書画像に含まれた各文字画像の外接矩形を抽出する。ここで「文字画像」とは、所定の言語からなる文字が画像として表されたものを意味する。行切出部２４は、矩形抽出部２３で抽出された外接矩形を連結することで文字行の切り出しを行う。以下、文字行に含まれる外接矩形を「行内矩形」という。

【００３６】

量子化部２５は、行切出部２４で切り出された文字行に含まれる各行内矩形の特性を表す配置情報を固定段階に量子化する。ここで、行内矩形の特性とは、各行内矩形に対応する文字画像の黒画素密度や文字行内における行内矩形の高さ、始点位置等のパラメータ群であって、行内矩形に固有の配置状態を表すものである。なお、配置情報の量子化については後述する。

【００３７】

シンボル生成部２６は、量子化部２５により量子化された配置情報の各々を固定種類のシンボルにシンボル化し、文書画像を構成する各文字行に対応する一連のシンボル系列を生成する。以下、文書画像全体についてのシンボル系列を全体シンボル系列という。

【００３８】

出現頻度集計部２７は、全体シンボル系列内において、所定個のシンボルの組合せからなるシンボル系列が出現する頻度（出現頻度）を算出する。候補画像選定部２８は、照合画像と、当該照合画像の照合先となる被照合画像とについて、出現頻度集計部２７により算出された出現頻度を照合し、より高い相関を有した被照合画像を所定個数選定する。以

下、候補画像選定部２８により選定された被照合画像を「候補画像」という。

【００３９】

出現位置分布導出部２９は、照合画像と候補画像との各文書画像において、両文書画像で一致した各シンボル系列に対応する配置情報の何れか又は全てが表す行内矩形に基づき、当該行内矩形の出現位置の分布状態を文書画像毎に夫々導出する。また、出現位置分布導出部２９は、照合画像についての分布状態と、候補画像についての分布状態との類似度を算出し、算出した類似度を対応する候補画像と対応付けてＲＡＭ４等に保持する。

【００４０】

照合結果選定部３０は、出現位置分布導出部２９により算出された類似度に基づいて、最も高い類似度を有した候補画像を照合結果として選定する。

【００４１】

表示部３１は、画像入力部２１から入力された文書画像や各処理の経過状況等の表示を行うとともに、照合結果選定部３０により選定された候補画像の表示を行う。なお、表示部３１の機能は、図１に示した表示装置６により実現できる。

【００４２】

以下、文書処理装置１００が実行する各種の処理のうち、本実施の形態に特長的な処理である文書照合処理について以下に説明する。

【００４３】

図３は、文書照合処理の手順を示したフローチャートである。まず、照合画像選択部２２は、画像入力部２１から入力される文書画像や、キーボード５を介して指定された文書画像を照合画像として選択する（ステップＳ１）。次いで、照合画像選択部２２は、ステップＳ１で選択した照合画像の照合先となる、被照合画像を選択する（ステップＳ２）。

【００４４】

続いて、矩形抽出部２３、行切出部２４、量子化部２５、シンボル生成部２６及び出現頻度集計部２７は、ステップＳ１、Ｓ２で選択された各文書画像について、出現頻度集計処理を実行する（ステップＳ３）。以下、図４を参照して、ステップＳ３の出現頻度集計処理について説明する。なお、出現頻度集計処理は、照合画像及び被照合画像の各々について行われるものとするが、以下の説明では「文書画像」と総称して説明する。

【００４５】

図４は、出現頻度集計処理の手順を示したフローチャートである。まず、矩形抽出部２３は、文書画像に含まれた各文字画像の黒画素に外接する外接矩形を抽出する（ステップＳ３１）。続いて、行切出部２４は、水平方向に隣接する外接矩形同士を連結して文字行に成長させた後、この文字行を夫々切り出す（ステップＳ３２）。

【００４６】

ここで、文書画像の行の切り出しについて、図５－１～図５－３を参照して説明する。矩形抽出部２３は、文書画像（図５－１）について、黒画素の連結成分を求め、それと外接する外接矩形Ａ，Ｂ，Ｃ・・・を求める（図５－２）。そして、行切出部２４は、矩形抽出部２３により求められた外接矩形を、水平方向に隣接する外接矩形同士を連結して文字行Ｚに成長させる（図５－３）。行内矩形の生成及び文字行の切り出しにかかる処理自体は、公知の手法を用いることができるため詳細な説明は省略する。

【００４７】

なお、文書画像から一つの文字行として切り出す単位は、行単位や段落単位、章単位等で切り出すことが好ましい。一般的に文書画像に含まれる文字画像のサイズは、行単位や段落単位、章単位で均一となるため、このような纏まりで文字行を切り出すことで、当該文字行内に含まれる文字画像のサイズ（文字サイズ）を揃えることが可能となる。また、本実施形態では、外接矩形の成長を水平方向で実施する態様としたが、これに限らず、文字方向等に応じて垂直方向、或いは、水平方向及び垂直方向の両方で実施する態様としてもよい。

【００４８】

図４に戻り、量子化部２５及びシンボル生成部２６は、ステップＳ３２で切り出した各

文字行について、シンボル生成処理を実行する。以下、図６を参照してステップＳ３３のシンボル生成処理について説明する。

【００４９】

図６は、シンボル生成処理の手順を示したフローチャートである。まず、量子化部２５は、ステップＳ３２で切り出された各文字行の高さを計測する（ステップＳ３３１）。

【００５０】

次いで、量子化部２５は、各文字行に含まれる各行内矩形の水平方向の始点（Ｘｓ）に基づいて、当該行内矩形を昇順にソートすることで配置順序を整列する（ステップＳ３３２）。続いて、量子化部２５は、整列した各行内矩形の配置状態を表す配置情報を夫々取得し、この配置情報を固定段階に量子化する（ステップＳ３３３）。以下、図７－１、図７－２、図８および図９を参照して、ステップＳ３３２、Ｓ３３３の処理を説明する。

【００５１】

図７－１および図７－２は、行内矩形の配置例を示す説明図である。欧米系文字行は、図７－１に示すように、大文字と小文字とが混在していることに加え、アポストロフィー、アクサンテギュ、ウムラウトなど、記号類の有無が存在するので、行内矩形の始点の高さは、図７－１のａの位置とｂの位置との２カ所に集中することは明らかである。つまり、矩形の配置位置は上下に対称ではない。一方、アジア系文字行は、図７－２に示すように、漢字、ひらがな、カタカナ、ハングルなど、文字の構造が複雑であり、行内矩形の始点の高さは、欧米系文字行で見られるような、２カ所への明確な集中はない。しかし、矩形の配置位置が上下左右、対称ではないことは、欧米系行と同じである。

【００５２】

図７－１の欧文文字の行内矩形と、図７－２のアジア系文字の行内矩形とを比較してみると、行内矩形の並び方は言語の種類に関わらず、その文字行の内容に応じて変化していることがわかる。そこで、文字の外接矩形を抽出することで、文字の大まかな特徴を捉えることができる。すなわち、文字そのものを特定しなくても、例えば図８に示すように、矩形座標の始点（Ｘｓ，Ｙｓ）と終点（Ｘｅ，Ｙｅ）を求め、これを利用した文字画像の外接矩形の配置状態を表す特徴を取得するだけで各文字行の画像特徴を捉えることができる。

【００５３】

行内矩形の配置位置が同じであっても、欧米系文字は構造が単純なためアジア系文字と較べて矩形内の黒画素密度は低くなる。なお、アジア系文字においても、構造が簡単なひらがな、カタカナの黒画素密度は低く、構造が複雑な漢字の黒画素密度が高くなることは言うまでもない。

【００５４】

このように、文字行内における一つの矩形の配置状態は、行内矩形の始点の高さ、矩形サイズ（幅、高さ）行内矩形中の黒画素密度等を計測することによって唯一に定義することができる。ステップＳ３３３の処理では、これら計測結果を配置情報として各文字行の行内矩形毎に取得し、固定段階に量子化する。

【００５５】

以下では、行内矩形の始点の高さを基準にして行内矩形の配置状態を定義する一例を示す。図９は、行内矩形の配置状態を示す特徴を量子化する方法を示す説明図である。原稿を特定していない状況下では、行高さは可変であり、処理が行高さの値に依存しないように、行内矩形の高さを次式で正規化する。なお、ｙｓは行内矩形始点の高さ、Ｈはステップ３３２で取得した行高を意味する。

$$YsRate = ys / H \quad \cdots (1)$$

【００５６】

ここで、$0 < YsRate \leqq 1$であるから、ＹｓＲａｔｅを固定段階に量子化することは容易である。例えば、Ｎ段階に量子化するなら、

$$YsVal = INT(YsRate * (N-1)) \quad \cdots (2)$$

（ただし、ＩＮＴ（）：小数点以下切捨て）

とすればよい。各段階は、０～（Ｎ－１）とラベル付けされる。矩形幅ｗおよび矩形高さｈも同様の手順で量子化される。

【００５７】

ところで、記憶容量節約および演算量低減のためなどの理由で、画像処理においては原画像そのものではなく圧縮画像を処理対象にする場合が多い。圧縮画像は、画素数が減るために文字画像の細部に関する情報は失われる。本発明は、図９に示すように、文字画像の外接矩形に注目するものであり、画像そのものの詳細な特徴に基づくものではない。したがって、原画像だけでなく、圧縮画像に対しても有効に機能しうる。

【００５８】

なお、上記では文字行画像の特徴として行内矩形の始点の高さを基準としたが、これに限定されない。例えば、文字行画像の特徴として行内矩形の高さを用いる場合は、図９において、次のとおりである。

ＨｅｉｇｈｔＲａｔｅ＝ｈ／Ｈ　・・・（３）

ＨｅｉｇｈｔＶａｌ

＝ＩＮＴ（ＨｅｉｇｈｔＲａｔｅ＊（Ｎ－１））＋０．５　・・・（４）

（ただし、ＩＮＴ（）：小数点以下切捨て）

各段階は、０～（Ｎ－１）とラベル付けされる。

【００５９】

また、文字行画像の特徴として行内矩形の幅を用いる場合は、次のとおりである。

ＷｉｄｔｈＲａｔｅ＝ｗ／Ｈ　・・・（５）

ＷｉｄｔｈＶａｌ

＝ＩＮＴ（ＷｉｄｔｈＲａｔｅ＊（Ｎ－１））＋０．５　・・・（６）

（ただし、ＩＮＴ（）：小数点以下切捨て）

各段階は、０～（Ｎ－１）とラベル付けされる。

【００６０】

図５に戻り、続いて、シンボル生成部２６は、ステップＳ３３３で量子化された配置情報の各々を固定種類のシンボルにシンボル化した後（ステップＳ３３４）、図４のステップＳ３４の処理に移行する。

【００６１】

以下、図１０および図１１を参照し、ステップＳ３３４の処理について説明する。上述したとおり、ステップＳ３３３で取得された配置情報は、対応する行内矩形の配置状態を特徴付けるものとなっている。ステップＳ３３４の処理では、量子化された配置情報に含まれる複数種類の測定結果を一つにまとめてシンボル化することで、一つの行内矩形を一つのシンボルに対応させる。

【００６２】

例えば、矩形の始点の高さ、矩形高さ、矩形幅の３種の情報をまとめる。仮に、前述の処理で、矩形の始点の高さ（ｙｓ／Ｈ）を１５段階、矩形高さ（ｈ／Ｈ）を８段階、矩形幅（ｗ／Ｈ）を２段階に量子化するとする。この結果、図１０に示すように、各情報は、矩形の始点の高さ（ｙｓ／Ｈ）は１５段階であるから４ｂｉｔｓ、矩形高さ（ｈ／Ｈ）は８段階であるから３ｂｉｔｓ、矩形幅（ｗ／Ｈ）は２段階であるから１ｂｉｔで表現することができる。また、

４ｂｉｔｓ＋３ｂｉｔｓ＋１ｂｉｔ＝８ｂｉｔｓ

であるから、１ｂｙｔｅの各ビットに全情報を格納することができる。そして、これらの３種の情報を一つにまとめたシンボルの種類は、

１５段階×８段階×２段階＝２４０種

となる。

【００６３】

ところで、矩形の配置状態を表す複数の特徴を多次元ベクトルの各次元とみなせば、矩形は、その各特徴を用いて一つのベクトルデータに変換（ベクトル量子化）できる。ベクトル量子化とは、周知のように、ベクトルデータの多数のバラエティから、それらを代表

する少数のベクトルデータを求めることである。求められた代表ベクトルに順にラベル付けすれば、ベクトルデータの系列を単なる一次元のシンボルデータの系列に変換することができる。ベクトル量子化に関しては、「ベクトル量子化と情報圧縮」（コロナ社）Ａｌｌｅｎ　Ｇｅｒｓｈｏ，Ｒｏｂｅｒｔ　Ｍ．Ｇｒａｙ著、田崎三郎ほか訳、に詳しい。

【００６４】

なお、まとめる情報の種類及びその格納のための記憶エリアは、記憶サイズは固定ではなく、識別対象である文字行を特定するのに好適な情報を適宜選択し、決定することが可能であることは言うまでもない。また、図１０では、矩形の始点の高さ、矩形高さ、矩形幅についてシンボル化する例を示したが、これに限らず、上述した黒画素密度などの配置情報を含めてシンボル化する態様としてもよい。

【００６５】

以上の作業を経ることによって、シンボル生成部２６は、各文字行に含まれる行内矩形を、固定個のシンボル（ラベル）に変換することができる。したがって、実際の行内矩形の配置は、図１１に示すような単なるシンボルの並びとみなすことができる。これで、シンボル系列の並び傾向を記録することができ、行内矩形の並び傾向を記録することと等価となる。

【００６６】

図４に戻り、出現頻度集計部２７は、ステップＳ３４でシンボル化した各配置情報に対して、所定個のシンボルの組合せからなるシンボル系列の出現頻度を照合画像及び被照合画像の各々について夫々算出、集計し（ステップＳ３４）、図３のステップＳ４の処理に移行する。

【００６７】

以下、ステップＳ３４の処理について説明する。配置情報がシンボル化された後には、テキスト検索と同様に、一般的な検索手法によって検索することが可能になる。つまり、照合画像と被照合画像についてシンボル系列間の完全一致を求めればよい。ただし、文字行画像の読み取り誤差によって、文字矩形の特徴の計測結果は異なるので、文字行が同一であっても、そのシンボル変換結果が同一にならない場合もある。よって、シンボル系列の完全一致を求めるのみでは、同一文字行画像を検索できない虞がある。

【００６８】

そこで、ステップＳ３４の処理では、シンボル系列の完全一致ではなく、シンボル系列の並び傾向の相関を求める。具体的には、照合画像及び被照合画像について生成された全シンボル系列の各々における、所定個のシンボルの組みからなるシンボル系列の出現頻度を算出し集計する。

以下、詳述する。

【００６９】

並びの傾向を記録する手段としては、ｎ－ｇｒａｍモデルがある。ｎ－ｇｒａｍモデルは、クロード・エルウッドシャノンによって提案された言語モデルである。このモデルでは、系列中のシンボルの出現が、直前のｎ個（ｎは自然数）のシンボルに影響されるとしている。現在の状態がｎ個前の入力に依存して決まる確率プロセスをｎ重マルコフ過程と呼び、ｎ－ｇｒａｍモデルは（ｎ－１）重マルコフモデルとも呼ばれる。特に、ｎ＝３の場合をｔｒｉｇｒａｍと呼び、広く使用されている。

【００７０】

具体的には、下記式（７）で示されるモデルである。さらに、式（８）にしたがって、照合画像及び被照合画像の各全シンボル系列から３つのシンボルの組みからなるシンボル系列（ｔｒｉｇｒａｍ）の出現頻度を夫々算出する。

【数１】

$$P(W)=\prod_{i=1}^{n}P(w_i \mid w_{i-2},w_{i-1}) \quad \cdots(7)$$

$P(w_i \mid w_{i-2},w_{i-1})$：$w_{i-2},w_{i-1}$ の後に w_i が出現する条件付き確率

$$P(w_i \mid w_{i-2},w_{i-1})=C(w_{i-2},w_{i-1},w_i)/C(w_{i-2},w_{i-1}) \quad \cdots(8)$$

$C(w_{i-2},w_{i-1})$　：系列w_{i-2},w_{i-1} の出現頻度

$C(w_{i-2},w_{i-1},w_i)$　：系列w_{i-2},w_{i-1},w_i の出現頻度

【００７１】

一方で、ｔｒｉｇｒａｍの出現頻度順位を求め、出現頻度の高い順にｔｒｉｇｒａｍを集計する。表１に、ｔｒｉｇｒａｍ集計結果の一例を示す。

【表１】

表1

順位	trigram	出現頻度
1	[s013,s045,s032]	324
2	[s013,s064,s033]	312
3	[s015,s005,s221]	302
4	[s016,s145,s203]	287
5	[s134,s002,s102]	279
6	[s134,s095,s244]	269
7	[s137,s187,s105]	256
8	[s138,s076,s228]	244
9	[s140,s097,s003]	235
10	[s141,s045,s013]	209
…		

【００７２】

表１において、出現頻度はｔｒｉｇｒａｍに示した３つ組みのシンボル系列、即ち三つの行内矩形を表す配置情報が、全体シンボル系列中にこの順序で出現する頻度を表している。例えば、ｔｒｉｇｒａｍ［s013，s045，s032］では、s013，s045の後ろにs032が出現する頻度が324であり、ｔｒｉｇｒａｍ［s013，s064，s033］では、s013，s064の後ろにs033が出現する頻度が312であることを示している。このように、文書画像の全シンボル系列に関して表１に示したようなｔｉｇｒａｍ集計結果を求めることが、各文書画像の特徴を求めること（学習）に相当する。

【００７３】

以上の動作を経ることによって、出現頻度集計部２７は、照合画像および被照合画像の各文書画像について、表１に示したようなｔｒｉｇｒａｍの出現確率の集計結果を導出する。

【００７４】

続いて、候補画像選定部２８は、ステップＳ３の処理で導出された照合画像に対応する集計結果（照合画像集計結果）と、被照合画像に対応する集計結果（被照合画像集計結果）とを照合し、より高い相関を有した上位ｎ個の被照合画像を候補画像として選定する（ステップＳ４）。ここで、「ｎ」は１以上の整数であって、任意の値を設定することが可能であるものとする。

【００７５】

照合画像集計結果と、被照合画像集計結果とを照合する場合、一つの文字行に含まれる行内矩形の個数は同値とならないことが多いため、出現頻度そのものを比較することは有意ではない。そのため、ステップＳ４では、下記式（９）に示した順位相関係数を用いることで、照合画像集計結果と、被照合画像集計結果との相関を判定する。

Ｒｘｙ＝１－（６＊Σ（Ｒｘｉ－Ｒｙｉ）＾２）／（ｎ＊（ｎ＾２－１））　・・・（９）

【００７６】

ここで、ｎはデータ数、Ｒｘｉは照合画像集計結果の順位毎の出現頻度、Ｒｙｉは被照合画像集計結果の順位毎の出現頻度を意味しており、各順位についてＲｘｉとＲｙｉとの差を二乗した値の総和がΣにより演算されるようになっている。なお、順位相関係数に関しては、「ノンパラメトリック法」（培風館）柳川尭著に詳しい。

【００７７】

候補画像選定部２８は、照合画像集計結果と、被照合画像集計結果とに含まれる各出現頻度について、順位相関係数Ｒｘｙを算出し、被照合画像のうち、Ｒｘｙの値が“１”に近いものからｎ個分の被照合画像を候補画像として選定する。なお、順位相関係数を統計的に検定し、最大の順位相関係数が有意な値を示さない場合には、照合画像に類似する被照合画像はない、と判断することとしてもよい。

【００７８】

ここで、図１２を参照して、上述したステップＳ１～Ｓ４迄の処理の概要を説明する。ステップＳ１、Ｓ２の処理において、照合画像Ｘと、複数の被照合画像Ｙとが選択されると、ステップＳ３の処理では、これら文書画像を構成する各文字行に含まれた行内矩形の各々が、配置情報に基づいてシンボル化され、照合画像Ｘについての全シンボル系列Ｘ１と、各被照合画像Ｙについての全シンボル系列Ｙ１とが夫々生成される。そして、全シンボル系列中における、ｔｒｉｇｒａｍの出現頻度が集計されることで照合画像Ｘに対応する照合画像集計結果Ｘ２と、被照合画像Ｙの夫々に対応する被照合画像集計結果Ｙ２とが導出される。続いて、照合画像集計結果Ａ２に含まれた順位毎の出現頻度と、被照合画像集計結果Ｂ２の夫々に含まれた順位毎の出現頻度と、に基づいて順位相関係数Ｒｘｙが算出される。

【００７９】

続くステップＳ４において、ステップＳ３で算出された被照合画像Ｙ毎の順位相関係数Ｒｘｙの値に基づいて、この値が“１”に近いものからｎ個分の被照合画像Ｙが候補画像として選定されることになる。

【００８０】

図３に戻り、出現位置分布導出部２９は、出現位置分布照合処理を実行する（ステップＳ５）。以下、ステップＳ５の出現位置分布照合処理について説明する。

【００８１】

図１３は、出現位置分布照合処理の手順を示したフローチャートである。まず、出現位置分布導出部２９は、ステップＳ４の処理で選定されたｎ個の候補画像から、本処理の対象とする候補画像を一つ選択する（ステップＳ５１）。

【００８２】

続いて、出現位置分布導出部２９は、ステップＳ５１で処理対象とした候補画像の被照合画像集計結果と、照合画像の照合画像集計結果とに基づいて、両文書画像の間で一致するｔｒｉｇｒａｍ、即ち三つのシンボルの組みからなるシンボル系列を選択する（ステップＳ５２）。ここで、選択するｔｒｉｇｒａｍの個数は特に問わないものとするが、より

出現頻度の高いｔｒｉｇｒａｍを選択することが好ましい。また、ｔｒｉｇｒａｍを構成する三つのシンボルのうち、何れかのシンボルを選択する態様としてもよい。

【００８３】

次いで、出現位置分布導出部２９は、照合画像と処理対象の候補画像とについて、文書画像の水平方向および垂直方向における、ステップＳ５２で選択したシンボル系列に対応する行内矩形の出現位置の分布状態をヒストグラム（度数分布ヒストグラム）として導出する（ステップＳ５３）。

【００８４】

図８に示したように、行内矩形は始点（Ｘｓ、Ｙｓ）と終点（Ｘｅ、Ｙｅ）との２点により表現される。そのため、水平方向（Ｘ軸）に関して分布をとる場合、始点Ｘｓについてヒストグラムを生成すればよく、垂直方向（Ｙ軸）に関しては分布をとる場合、始点Ｙｓについてヒストグラムを生成すればよい。

【００８５】

図１４は、行内矩形の存在位置の分布状態をヒストグラムで表現した一例を示した図である。同図に示したように、照合画像と被照合画像との両文書画像の間で一致した行内矩形（図中Ｋ）について、文書画像の水平方向と垂直方向でのヒストグラムを夫々導出する。ヒストグラム集計にあたっての集計幅は、特に問わないものとするが、例えば、ステップＳ３の処理で切り出した各文字行の高さの平均値程度とすることとしてもよい。

【００８６】

図１３に戻り、次に出現位置分布導出部２９は、ステップＳ５３で求めた両ヒストグラムを照合し、その類似度を算出する（ステップＳ５４）。なお、本実施形態では両ヒストグラムの照合方法として、メジアン（中央値）、モード（最頻値）、平均の各々が属するデータ区間のヒストグラム値を、両ヒストグラムの間で比較するものとする。

【００８７】

具体的には、ヒストグラムのデータ区間を座標の小さいものから順次番号付けし、メジアン、モード、平均の所属するデータ区間の番号を求める。ここで、メジアン、モード、平均の所属するデータ区間番号を（MedianClassNo, ModeClassNo, AvClassNo）と表現すれば、以下の４種の組が求められる。

（MedianClassNoXaxQuery,ModeClassNoXaxQuery,AvClassNoXaxQuery）　・・・（１０）

（MedianClassNoYaxQuery,ModeClassNoYaxQuery,AvClassNoYaxQuery）　・・・（１１）

（MedianClassNoXaxDB,ModeClassNoXaxDB,AvClassNoXaxDB）　・・・（１２）

（MedianClassNoYaxDB,ModeClassNoYaxDB,AvClassNoYaxDB）　・・・（１３）

【００８８】

なお、「XaxQuery」は、照合画像の水平方向のヒストグラムを意味するものであり、上記（１０）式は、照合画像の水平方向のヒストグラムにおける、該当するデータ区間番号のヒストグラム値を夫々意味する。また、「YaxQuery」は、照合画像の垂直方向のヒストグラムを意味するものであり、上記（１１）式は、照合画像の垂直方向のヒストグラムにおける、該当するデータ区間番号のヒストグラム値を夫々意味する。また、「XaxDB」は、被照合画像の水平方向のヒストグラムを意味するものであり、上記（１２）式は、被照合画像の水平方向のヒストグラムにおける、該当するデータ区間番号のヒストグラム値を夫々意味する。また、「YaxDB」は、被照合画像の垂直方向のヒストグラムを意味するものであり、上記（１３）式は、被照合画像の垂直方向のヒストグラムにおける、該当するデータ区間番号のヒストグラム値を夫々意味する。

【００８９】

出現位置分布導出部２９は、上記４種の組の値を算出した後、下記（１４）～（１６）式を用いて、垂直方向についての照合画像のヒストグラムと、被照合画像のヒストグラムとの形状の類似度を算出する。

MedianClassNoXaxDB＋CA=MedianClassNoXaxQuery　・・・（１４）

ModeClassNoXaxDB＋CA=ModeClassNoXaxQuery ・・・（１５）
AvClassNoXaxDB＋CA=AvClassNoXaxQuery ・・・（１６）

【００９０】

上記（１４）～（１６）式において、「ＣＡ」は定数であって、最初に処理する１式（例えば（１４）式）から求まる値である。出現位置分布導出部２９は、この定数ＣＡの値が残りの２式にて成立するか否か、つまり、残り２式での定数ＣＡからのずれの度合いを、照合画像のヒストグラムと、被照合画像のヒストグラムとの形状の類似度として算出する。なお、定数ＣＡからのずれの度合いは、例えば、ＣＡ’／ＣＡを算出することで導出できる。ここで、ＣＡ’は、ＣＡ＋α（αは定数ＣＡからのずれ値）であり、完全一致する際のずれの度合い、即ち類似度は“１”となる。

【００９１】

また、同様に出現位置分布導出部２９は、下記（１７）～（１９）式を用いて、垂直方向についての、照合画像のヒストグラムと、被照合画像のヒストグラムとの形状の類似度を算出する。

MedianClassNoYaxDB＋CB=MedianClassNoYaxQuery ・・・（１７）
ModeClassNoYaxDB＋CB=ModeClassNoYaxQuery ・・・（１８）
AvClassNoYaxDB＋CB=AvClassNoYaxQuery ・・・（１９）

【００９２】

上記（１７）～（１９）式において、「ＣＢ」は定数であって、上述したＣＡと同様、最初に処理する１式（例えば（１７）式）から求まる値である。出現位置分布導出部２９は、この定数ＣＢの値が残りの２式にて成立するか否か、つまり、残り２式での定数ＣＢからのずれの度合いを、照合画像のヒストグラムと、被照合画像のヒストグラムとの形状の類似度として算出する。なお、定数ＣＢからのずれの度合いは、上述した定数ＣＡについてと同様に導出することができる。

【００９３】

出現位置分布導出部２９は、上記の手続きにより算出した水平方向および垂直方向での類似度を、処理対象の被照合画像と対応付けてＲＡＭ４等に保持する。ここで、水平方向（又は垂直方向）に対して導出されるずれの度合いの個数は、２式（或いは３式）分となるが、これらを個別に類似度として保持する態様としてもよいし、これらの平均値を類似度として保持する態様としてもよい。

【００９４】

なお、本実施形態では、文書画像の水平方向および垂直方向の両方向について、ヒストグラムの形状の類似度を算出したが、何れか一方向のみについて算出する態様としてもよい。また、本実施形態では、行内矩形の出現位置の分布状態をヒストグラムで表すものとしたが、これに限らず、例えば正規分布を用いて表すものとしてもよい。

【００９５】

図１５は、行内矩形の存在位置の分布状態を正規分布で表現した一例を示した図である。同図に示したように、正規分布を用いて表す場合には、各行内矩形の始点に基づいた集計結果から、水平方向（Ｘ軸）に関して、平均μｘ、標準偏差σｘ、歪度、尖度を算出し、また同様に垂直方向（Ｙ軸）に関して、平均μｙ、標準偏差σｙ、歪度、尖度を算出すればよい。

【００９６】

この場合、平均値については、照合画像と被照合画像とで画像サイズが異なる可能性があるため、直接比較することは有意ではない。正規分布の形状が一致しているか否かを求めるには、標準偏差、歪度、尖度が類似しているかを判定すればよい。例えば、検索画像の標準偏差、歪度、尖度と、被検索画像の標準偏差、歪度、尖度との各々を比較し、比率が１に近いものほど正規分布の形状が類似するものと判断することができる。

【００９７】

なお、照合画像の解像度と、被照合画像の解像度とが一致している場合には、同一文字を構成するドット数は同じになるが、解像度が異なる場合にはドット数は同じにならない

。つまり、ヒストグラムや正規分布の形状の一致を評価する場合にも、解像度が同じ場合には両者の数値をそのまま利用しても構わないが、解像度が異なる場合には、ドット数に基づく数値をそのまま利用することができない。

【００９８】

そこで、両文書画像の解像度が異なる場合、或いは解像度自体が未知の場合には、数値の正規化を行う必要がある。一般的な文書画像においては段落単位では文字のサイズは同一であるため、同じ段落に属する文字行は行高さが等しくなる。また、照合画像が被照合画像の一部分であれば、同じ行高さになる可能性が高いことは明らかである。よって、被検索画像および検索画像において、各文字行の行高さを集計し、最頻出となる行高さについて、ヒストグラムを規定する数値（平均、モード、メジアン）を除算する。なお、正規分布の場合も同様である。また、最頻出の行高さではなく、各文字行の行高さの平均値で除算してもよい。いずれを選択するかは設計事項であり、使用する環境に応じて決定すればよい。

【００９９】

また、照合画像が被照合画像の一部分であっても、その一部分の特異な部分だけが照合画像となった場合には、全体画像において最頻出する行高さが、部分画像において最頻出となる行高さと一致しないことが考えられる。例えば、本文行と見出し行とは行高さが大きく異なる文書画像において、全体画像の行数としては本文行が圧倒的に多いと予想される。その文書の部分画像には見出し行だけしか含まれていない場合には、最頻出行は見出し行となり、全体画像の最頻出行から推定した行高さとは一致しないため、この一致しない結果に基づいて正規化しても正しい比較結果を得ることができないのは明らかである。

【０１００】

このような場合、照合画像と被照合画像との両文書画像内において、一致した行内矩形（シンボル系列）だけを対象に矩形サイズの集計を行い、最頻出した矩形サイズのドット数に基づいて、数値（平均、モード、メジアン）を正規化することで対応することができる。

【０１０１】

図１３に戻り、出現位置分布導出部２９は、ステップＳ５４の処理で求めた類似度を、処理対象の候補画像に対応付けてハードディスク３又はＲＡＭ４に保持する（ステップＳ５５）。続いて、出現位置分布導出部２９は、ステップＳ４の処理で選定されたｎ個の被候補画像の全てに対して、本処理の処理対象としたか否かを判定する（ステップＳ５６）。ここで、本処理の対象としていない未処理の候補画像が存在すると判定した場合には（ステップＳ５６；Ｎｏ）、ステップＳ５１へと再び戻り、未処理の候補画像のうち一つを処理対象として選択する。

【０１０２】

一方、ステップＳ５６において、全ての候補画像を処理対象としたと判定した場合（ステップＳ５６；Ｙｅｓ）、図３のステップＳ６の処理に移行する。

【０１０３】

図３に戻り、照合結果選定部３０は、ステップＳ５の処理によりＲＡＭ４等に保持されたｎ個の候補画像の類似度に基づいて、最も高い類似度を有した候補画像、即ち類似度の値が“１”に最も近かった候補画像を照合結果として選定する（ステップＳ６）。

【０１０４】

続いて、表示部３１は、ステップＳ６の処理で照合結果に選定された文書画像を、照合画像に対する照合結果として表示装置６に表示し（ステップＳ７）、本処理を終了する。

【０１０５】

図１６は、上記文書照合処理の動作を説明するための図である。同図において、Ｄ１１は照合画像であって、特定の文書画像中の一部分となる部分画像が照合画像に選択された場合を示している。また、Ｄ２１～Ｄ２４は、ステップＳ４までの処理により選定された４つの候補画像を示している。なお、照合画像Ｄ１１は、候補画像Ｄ２４の部分画像となっている。即ち、候補画像Ｄ２４が照合画像Ｄ１１に最も類似する文書画像となっている

。

【０１０６】

上述したようにステップＳ４の処理では、行内矩形の配置情報に対応するシンボル系列を照合することで、照合画像Ｄ１１と相関関係にある文書画像として、候補画像Ｄ２１～Ｄ２４までを絞り込むことが可能である。なお、照合画像Ｄ１１、候補画像Ｄ２１～Ｄ２４中矩形Ｋで表した部分が、各文書画像で一致したシンボル系列（或いはシンボル）の行内矩形を意味している。

【０１０７】

しかしながら、ステップＳ４の処理ではシンボル系列の出現頻度に基づいて類似度を判断するのみであるため、候補画像Ｄ２４が照合画像Ｄ１１に最も類似する文書画像であること、即ち、照合画像Ｄ１１が候補画像Ｄ２４の一部分であることまでを判断することはできない。そのため、ステップＳ５の処理では、各文書画像で一致したシンボル系列の相対的な位置関係、即ち出現位置の分布状態を照合することで、候補画像Ｄ２４が照合画像Ｄ１１に最も類似する文書画像であることを特定することが可能となる。

【０１０８】

以上のように、本実施形態によれば、照合画像と被照合画像とについて、文字行内における外接矩形の特徴を表した配置情報を抽出し、これらを固定段階に量子化してシンボルを生成することにより、文字認識することなく文字行の特徴の抽出が可能となり、被照合画像から、照合画像と相関の高い被照合画像を所定の数だけ候補画像として選定することができる。また、照合画像と候補画像とについて、一致するシンボル系列の出現位置の分布状態を照合することで、当該シンボル系列の相対的な位置関係の類似性を判定することができるため、照合対象画像と候補画像との類似性を高精度に判定することができる。これにより、文書画像中の部分画像が照合対象の文書画像とされた場合であっても、この部分画像に含まれた文字画像の外接矩形の位置関係に基づいて、当該部分画像と類似する文書画像を高精度に検索することが可能となる。

【０１０９】

なお、本発明は、上記実施の形態そのままに限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で構成要素を変形して具体化することができる。また、上記実施の形態に開示されている複数の構成要素の適宜な組み合わせにより、種々の発明を形成することができる。例えば、実施の形態に示される全構成要素からいくつかの構成要素を削除してもよい。さらに、異なる実施の形態にわたる構成要素を適宜組み合わせても良い。

【０１１０】

例えば、本実施形態で実行される文書照合処理にかかるプログラムを、インターネット等のネットワークに接続されたコンピュータ上に格納し、ネットワーク経由でダウンロードさせることにより提供するように構成しても良い。また、本実施形態の文書処理装置１００で実行される文書照合処理にかかるプログラムをインターネット等のネットワーク経由で提供または配布するように構成しても良い。

【０１１１】

また、本実施形態で実行される文書照合処理にかかるプログラムを、ＲＯＭ等の記憶媒体に予め組み込んで提供するように構成してもよい。

【０１１２】

また、上記実施形態では、図２に示した各機能部をＣＰＵ１とＲＯＭ２に記憶された所定のプログラムとの協働により実現する態様としたが、これに限らず、ハードウェア構成により実現する態様としてもよい。具体的には、リアルタイム性が重要視される場合には、処理を高速化する必要があるため、論理回路（図示せず）を別途設け、論理回路の動作により各種の演算処理を実行するようにすることが好ましい。

【０１１３】

また、上記実施形態では、文字行よりも小さな単位として行内矩形に着目したが、これに限らず、他の単位でも適用可能である。例えば、文字（文字画像）単位や単語単位の画像特徴でも数値化し量子化することで、上記と同様にシンボル化することが可能であり、

照合することが可能である。この場合、黒画画素に基づいて文字画像を切り出したのち、当該文字画像の外接矩形を文字単位又は単語単位で用いることで対応することが可能である。なお、文字単位又は単語単位での分割は、ＯＣＲ（Optical Character Recognition）等で用いられる公知の文字切り出し手法を用いればよい。

【０１１４】

代表的な文字切り出し手法として、射影を利用する方法がある。この方法では、水平行について、垂直方向に黒画素数を集計し、その分布を求め、ある黒画素数がしきい値以下の部分を分割位置候補とする。また、分割位置候補に対しては、行高さから推定した文字幅、隣接する分割位置との距離、行全体に亘る分割位置の周期性等の観点から妥協点を評価し、適当な分割位置の選択を行う（垂直行も同様）。

【０１１５】

また、単語単位に分割する他の方法としては、欧文等分かち書きの習慣のある言語については、単語間の空白に基づいて容易に実現することが可能である。このように、文字単位、単語単位等の単位で分割された場合であっても、その範囲の画像に外接する矩形を求めることが可能であり、その外接矩形の開始位置、終点位置を用いることで行内矩形に対する場合と同様な手順で量子化を行うことができる。

【産業上の利用可能性】

【０１１６】

以上のように、本発明に係る文書処理装置、文書処理方法および文書処理プログラムは、文書画像間を照合する文字処理装置に有用であり、特に、文書画像の一部分となる部分画像を照合対象とし、この部分画像に類似する文書画像の検索を行う文書処理装置に適している。

【図面の簡単な説明】

【０１１７】

【図１】文書処理装置のハードウェア構成を示したブロック図である。
【図２】文書処理装置の機能的構成を示したブロック図である。
【図３】文書照合処理の手順を示したフローチャートである。
【図４】出現頻度集計処理の手順を示したフローチャートである。
【図５－１】文字行の切り出しを説明するための図である。
【図５－２】文字行の切り出しを説明するための図である。
【図５－３】文字行の切り出しを説明するための図である。
【図６】シンボル生成処理の手順を示したフローチャートである。
【図７－１】行内矩形の配置例を示した図である。
【図７－２】行内矩形の配置例を示した図である。
【図８】行内矩形に対する座標の設定例を説明するための図である。
【図９】行内矩形の配置状態を説明するための図である。
【図１０】配置情報の量子化を説明するための図である。
【図１１】量子化された配置情報をシンボル化した一例を示した図である。
【図１２】文書照合処理の概要を説明するための図である。
【図１３】出現位置分布照合処理の手順を示したフローチャートである。
【図１４】行内矩形の存在位置の分布状態をヒストグラムで表現した一例を示した図である。
【図１５】行内矩形の存在位置の分布状態を正規分布で表現した一例を示した図である。
【図１６】文書照合処理の概要を説明するための図である。

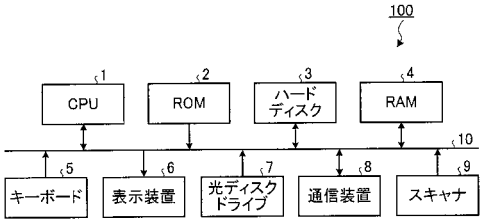
【符号の説明】

【０１１８】

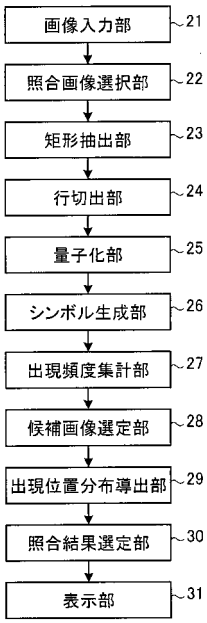
１００　文書処理装置
１　ＣＰＵ
２　ＲＯＭ
３　ハードディスク

４　ＲＡＭ
５　キーボード
６　表示装置
７　光ディスクドライブ
８　通信装置
９　スキャナ
１０　バスコントローラ
２１　画像入力部
２２　照合画像選択部
２３　矩形抽出部
２４　行切出部
２５　量子化部
２６　シンボル生成部
２７　出現頻度集計部
２８　候補画像選定部
２９　出現位置分布導出部
３０　照合結果選定部
３１　表示部

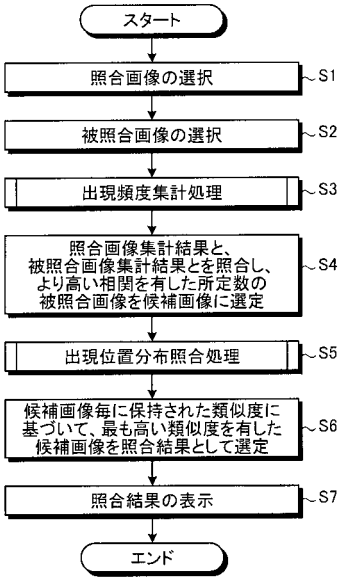
【図１】



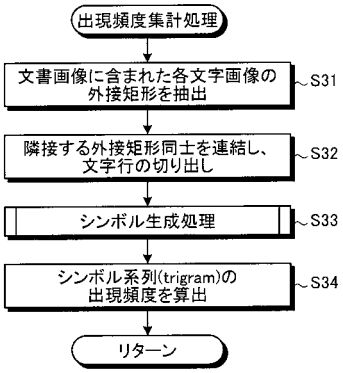
【図２】



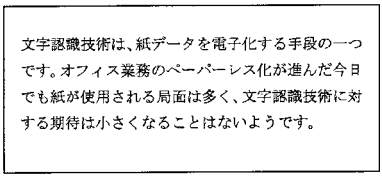
【図３】



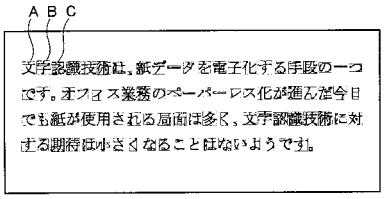
【図４】



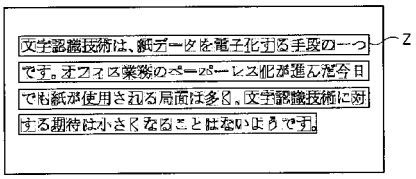
【図５－１】



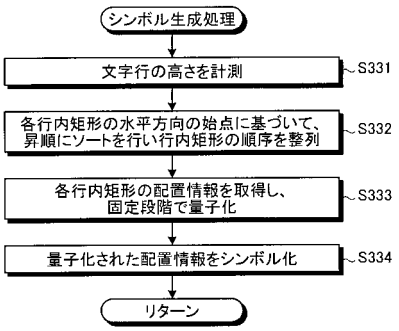
【図５－２】



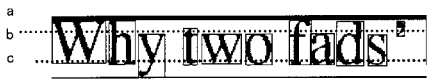
【図５－３】



【図６】

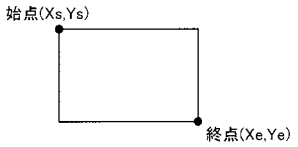


【図７－１】

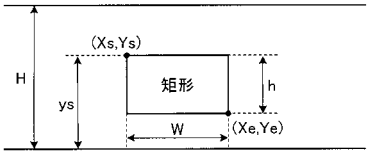


【図７－２】

【図８】

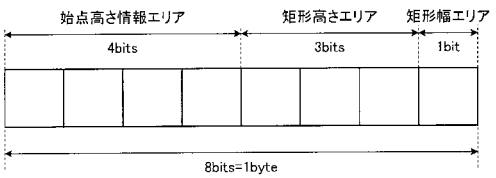


【図９】

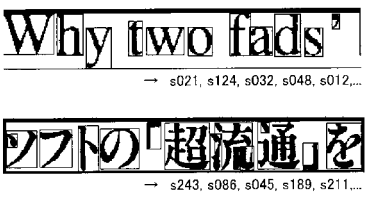


H: 行高さ
h: 行内矩形の高さ
w: 行内矩形の幅
ys: 行内矩形始点の高さ

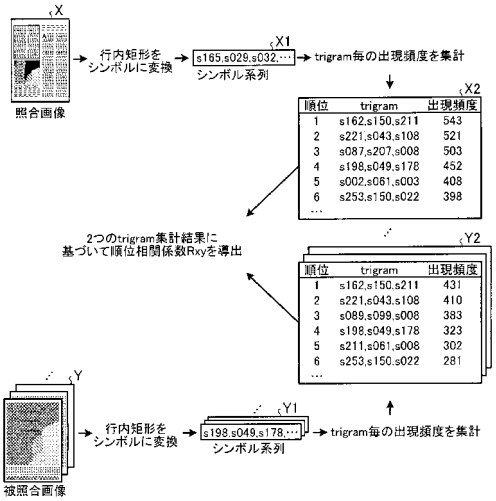
【図１０】



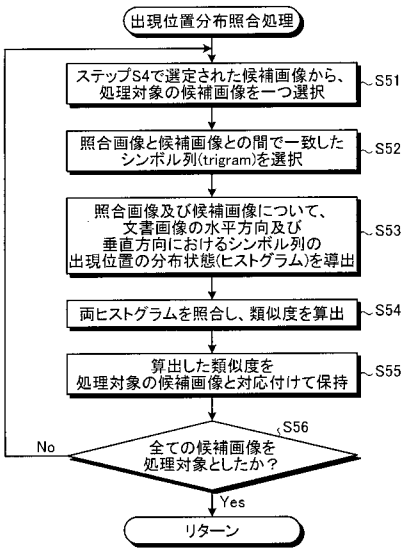
【図１１】



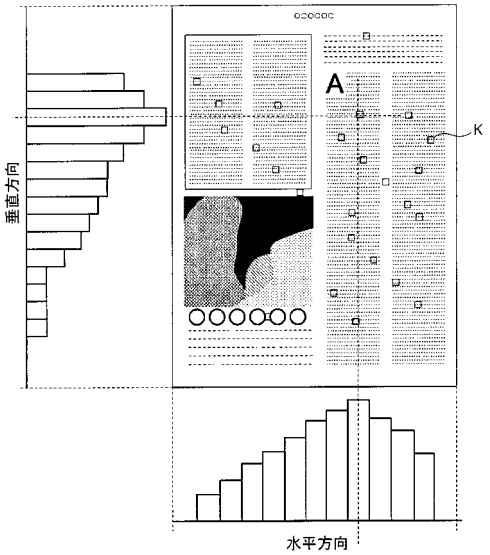
【図１２】



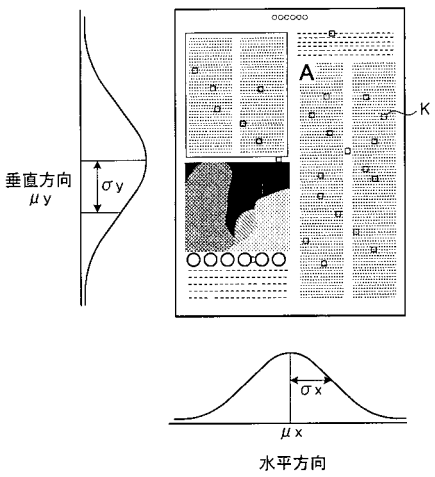
【図１３】



【図１４】



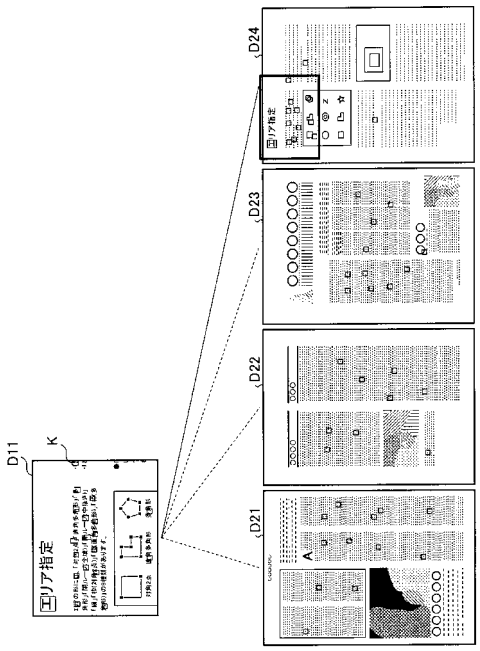
【図１５】



水平方向：X軸
標準偏差σx　平均μx
歪度、尖度

垂直方向：Y軸
標準偏差σy　平均μy
歪度、尖度

【図１６】



フロントページの続き

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｔ　　７／００
Ｇ０６Ｆ　１７／３０