(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2007-58605
(P2007-58605A)
(43)公開日　平成19年3月8日(2007.3.8)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	210A	5B050
G06K 9/72	(2006.01)	G06F	17/30	170B	5B064
G06T 1/00	(2006.01)	G06F	17/30	380Z	5B075
		G06K	9/72	A	
		G06T	1/00	200E	

審査請求 未請求　請求項の数 4　OL　（全 8 頁）

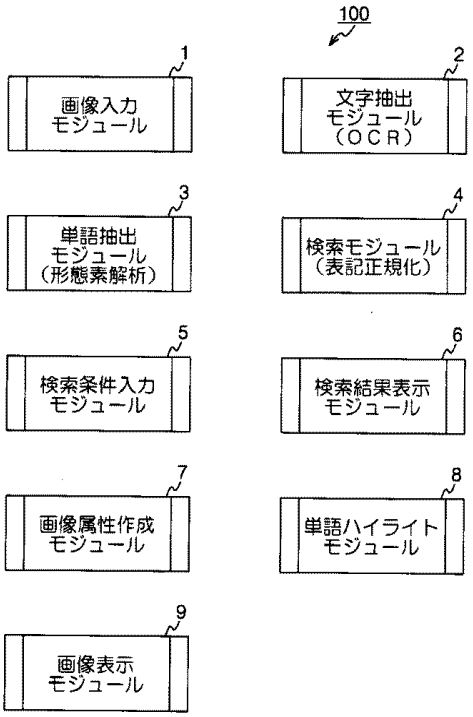
(21)出願番号　特願2005-243449 (P2005-243449)
(22)出願日　平成17年8月24日 (2005.8.24)

(71)出願人　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　100085660
弁理士　鈴木　均
(72)発明者　伊井　泰洋
東京都大田区中馬込１丁目３番６号
株式会社リコー内

Fターム(参考)　5B050 BA16 BA20 FA02 FA08 FA13 FA17 GA08
5B064 AA01 BA01 CA08 EA19 FA02 FA12
5B075 ND07 NK31 PP22 PQ02 PQ22

(54)【発明の名称】文書管理システム

(57)【要約】　（修正有）
【課題】　検索結果ハイライトが確実にできる文書管理システムの提供。
【解決手段】　画像を入力する画像入力モジュール１と、入力された画像中から文字列を抽出する文字抽出モジュール２と、抽出された文字列から単語を抽出する単語抽出モジュール３と、抽出された単語をインデックスとして登録し文書検索を行う検索モジュール４と、キーワードを入力する検索条件入力モジュール５と、キーワードによる検索結果から文書表示を行う際にキーワードもしくは正規化されたキーワードによるヒット文字列を抽出して表示する検索結果表示モジュール６と、抽出したヒット文字列と正規化されたキーワードに基づいて、キーワードの位置情報を計算して画像に登録する属性情報を作成する画像属性作成モジュール７と、画像上の単語をハイライト表示する単語ハイライトモジュール８と、検索結果より文書を選択して表示する画像表示モジュール９と、を備えて構成される。
【選択図】　図１



【特許請求の範囲】
【請求項１】
画像を電子データとして入力する画像入力手段と、該画像入力手段により入力された画像中から文字列を抽出する文字抽出手段と、該文字抽出手段により抽出された文字列から単語を抽出する単語抽出手段と、該単語抽出手段により抽出された単語をインデックスとして登録し文書検索を行う文書検索手段と、検索のためのキーワードを入力する検索条件入力手段と、前記キーワードによる検索結果から文書表示を行う際に前記キーワードもしくは正規化されたキーワードによるヒット文字列を抽出して表示する検索結果表示手段を有する文書管理システムにおいて、
前記抽出したヒット文字列と正規化されたキーワードに基づいて、前記キーワードの位置情報を計算して画像に登録する属性情報を作成する属性情報作成手段を備え、前記属性情報として前記抽出したヒット文字列と正規化されたキーワードを保持することにより、ハイライト表示の抜けを防止することを特徴とする文書管理システム。
【請求項２】
前記検索条件入力手段により入力したキーワードと前記正規化されたキーワードとのハイライト表示方法を変えることにより、前記ヒット文字列が前記入力したキーワードと同一か、あるいは正規化されたキーワードかを区別することを特徴とする請求項１に記載の文書管理システム。
【請求項３】
前記単語抽出手段は、自然言語で書かれた文を意味を持つ最小単位の列に分割し、品詞を見分ける形態素解析により単語を抽出することを特徴とする請求項１又は２に記載の文書管理システム。
【請求項４】
前記文書検索手段は、複数の表記をまとめて一つの表記として扱う表記正規化法により正規化したキーワードに基づいて検索することを特徴とする請求項１又は２に記載の文書管理システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、文書管理システムに関し、さらに詳しくは、画像を文書として管理し、蓄積、表示、検索する文書管理システムに関するものである。
【背景技術】
【０００２】
紙原稿を電子化して保管、検索を行う電子ファイリングシステムにおいては、登録済の文書を検索する機能、及び検索条件に該当する文書に対し、検索該当箇所を操作者に明示する方法が一般的に提供されている。例えば、「特許公報」という文字列を検索キーワードにして検索を行い、ヒットした文書を開くと、文書内の「特許公報」の文字が反転、マーキングなどで強調表示される。これを検索結果のハイライト処理と呼ぶ。また、文書検索においては、いくつかの異なる単語を同一のものとして扱い、検索のヒット率を上げる手法が採用されている。例えば、「メモリー」「メモリ」のように長音の有無により異なる単語は、同一のものとして処理した方が所望の文書を探しやすい。また「Memory」「MEMORY」「ＭＥＭＯＲＹ」のように、大文字と小文字による差、あるいは全角文字と半角文字による差を持つ単語も、同一として処理した方が所望の文書を探しやすい。このため、複数の表記をまとめてひとつの表記として扱うことを異表記正規化と呼ぶ。これとは逆にひとつの単語から、複数の正規化単語を抽出することを異表記逆正規化と呼ぶ。異表記逆正規化の例としては、Memoryというキーワードから、MEMORY、Ｍｅｍｏｒｙ、ＭＥＭＯＲＹのような正規化が予測される単語を生成することである。
【０００３】
一般的に、検索結果のハイライト処理の流れは以下のようになる。操作者が検索したいキーワード（語句）を入力し、検索サブシステムが検索を行い、該当する文書のリストを

返し、リスト上にある文書を表示する。キーワードに対して異表記逆正規化を行い、該当する単語すべてを表示文書上でハイライトを行う。この方式での問題点は、異表記逆正規化を行った場合に、検索漏れがでてくることである。例えば、検索サブシステムが、長音をすべて省略するような正規化を行う場合を考える。このとき、「コンピューター」という単語は正規化により「コンピュタ」となり、検索インデックスに登録される。また、別文書に「コンピュータ」という単語が含まれていた場合も、正規化の結果は「コンピュタ」となる。つまり、「コンピューター」も「コンピュータ」も同一の検索インデックスになる。これは、表現が微妙に違っていても、検索にヒットするという点では効果的である。しかし、「コンピュタ」の異表記逆正規化を考えた場合、ひとつの語句から多くの逆正規化単語が抽出される。「コンピュタ」を例にとれば「コンピュータ」「コンピューター」の他に「コーンピュータ」「コーンーピュータ」「コンーピュータ」他多くの抽出が行われる。これは、削除した長音を元に戻そうとする場合、それぞれの音の後に長音を付加するためである。単語が長くなればなるほど、異表記逆正規化の出力単語の種類は増えていく。即ち、異表記逆正規化の出力単語の膨大な増加は、処理時間の遅延をもたらすため、実際にはある程度の個数の単語を出力したら、そこで異表記逆正規化処理を止め、単語数の上限を設けている。しかし、この取りやめ処理により、異表記逆正規化により正規化前の単語を１００％取得することはできなくなる。つまり、検索にヒットしても、ハイライトが行われない文書が生成されることになる。

尚、従来技術として特許文献１には、ＯＣＲと文書処理装置を分離し、ＯＣＲの出力形態として、文字行抽出及び文字切出及び文字識別の多重仮説を保持するデータ（読取仮説データ）と、文書画像の罫線情報や枠情報や文字行情報や閲覧属性情報等を持つ文書構造データを採用し、ＯＣＲ付加データを元に印刷活字及び手書文字列からの重要キーワード抽出及び文書検索を行い、更に文書構造データを利用して閲覧者の意図する文書表示機能を構成することで、高度な機能を持つ文書画像検索・閲覧システムについて開示されている。

【特許文献１】特開２００５－１３５０４１公報

【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に開示されている従来技術は、ＯＣＲによる読取精度を高めるための技術であり、ＯＣＲ処理に時間が掛かると共に、そのための構成が複雑になるといった問題がある。

本発明は、かかる課題に鑑み、検索対象となる文書の属性情報として、抽出単語と異表記正規化された単語を保持することにより、検索結果ハイライトが確実にできる文書管理システ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】

【０００５】

本発明はかかる課題を解決するために、請求項１は、画像を電子データとして入力する画像入力手段と、該画像入力手段により入力された画像中から文字列を抽出する文字抽出手段と、該文字抽出手段により抽出された文字列から単語を抽出する単語抽出手段と、該単語抽出手段により抽出された単語をインデックスとして登録し文書検索を行う文書検索手段と、検索のためのキーワードを入力する検索条件入力手段と、前記キーワードによる検索結果から文書表示を行う際に前記キーワードもしくは正規化されたキーワードによるヒット文字列を抽出して表示する検索結果表示手段を有する文書管理システムにおいて、前記抽出したヒット文字列と正規化されたキーワードに基づいて、前記キーワードの位置情報を計算して画像に登録する属性情報を作成する属性情報作成手段を備え、前記属性情報として前記抽出したヒット文字列と正規化されたキーワードを保持することにより、ハイライト表示の抜けを防止することを特徴とする。

本発明は、画像の属性情報に抽出されたキーワードと、キーワードを正規化した文字列の両方を登録し、表示時にこれらの文字列をハイライトすることによって実施される。

請求項２は、前記検索条件入力手段により入力したキーワードと前記正規化されたキーワードとのハイライト表示方法を変えることにより、前記ヒット文字列が前記入力したキーワードと同一か、あるいは正規化されたキーワードかを区別することを特徴とする。

登録処理については、請求項１の発明と同様になる。また表示処理については、請求項１のハイライト箇所で、画像属性情報のうち、正規化によって生成されたキーワードか否かを判断し、正規化されたキーワードと、正規化されていないキーワードとで色を分けて表示するものである。

請求項３は、前記単語抽出手段は、自然言語で書かれた文を意味を持つ最小単位の列に分割し、品詞を見分ける形態素解析により単語を抽出することを特徴とする。

形態素解析による単語抽出は、文字列を言語で意味を持つ最小単位の列に分割して品詞を見分けるので、確実に正確な単語を抽出することができる。

請求項４は、前記文書検索手段は、複数の表記をまとめて一つの表記として扱う表記正規化法により正規化したキーワードに基づいて検索することを特徴とする。

表記正規化法により正規化したキーワードは、逆に一つの単語から複数の正規化単語を抽出する際に、有効に作用して的確な正規化単語を抽出することができる。

【発明の効果】

【０００６】

請求項１の発明によれば、画像の属性情報に抽出されたキーワードと、キーワードを正規化した文字列の両方を登録し、表示時にこれらの文字列をハイライトするので、ハイライト表示時に逆正規化の漏れによって、ハイライトされないケースが防止できる。

また請求項２では、実際に操作者が入力したキーと、ハイライト表示される文字列が異なる場合でも、ハイライト表示色やハイライト表示形式を変えることにより、操作者に対して理解しやすいインターフェースを提供することができる。

また請求項３では、単語抽出手段は、自然言語で書かれた文を言語で意味を持つ最小単位の列に分割し、品詞を見分ける形態素解析により単語を抽出するので、確実に正確な単語を抽出することができる。

また請求項４では、文書検索手段は、複数の表記をまとめて一つの表記として扱う表記正規化法により正規化したキーワードに基づいて検索するので、逆に一つの単語から複数の正規化単語を抽出する際に、有効に作用して的確な正規化単語を抽出することができる。

【発明を実施するための最良の形態】

【０００７】

以下、本発明を図に示した実施形態を用いて詳細に説明する。但し、この実施形態に記載される構成要素、種類、組み合わせ、形状、その相対配置などは特定的な記載がない限り、この発明の範囲をそれのみに限定する主旨ではなく単なる説明例に過ぎない。

図１は本発明の文書管理システムのモジュール構成を示す図である。この文書管理システム１００は、画像を電子データとして入力する画像入力モジュール（画像入力手段）１と、画像入力モジュール１により入力された画像中から文字列を抽出する文字抽出モジュール（文字抽出手段）２と、文字抽出モジュール２により抽出された文字列から単語を抽出する単語抽出モジュール（単語抽出手段）３と、単語抽出モジュール３により抽出された単語をインデックスとして登録し文書検索を行う検索モジュール（文書検索手段）４と、検索のためのキーワードを入力する検索条件入力モジュール（検索条件入力手段）５と、キーワードによる検索結果から文書表示を行う際にキーワードもしくは正規化されたキーワードによるヒット文字列を抽出して表示する検索結果表示モジュール（検索結果表示手段）６と、抽出したヒット文字列と正規化されたキーワードに基づいて、キーワードの位置情報を計算して画像に登録する属性情報を作成する画像属性作成モジュール（属性情報作成手段）７と、画像上の単語をハイライト表示する単語ハイライトモジュール８と、検索結果より文書を選択して表示する画像表示モジュール９と、を備えて構成される。

また、ハイライト表示されるのは、操作者が入力した検索語句そのものとは限らない。検索システム登録時に正規化された文字列が表示されることもある。例えば、操作者は、

「メモリー」と検索文字列を入力していても、検索システムの正規化時には「メモリ」と変更されるので、表示画像内に「メモリ」という文字列があった場合もハイライト表示される。

【０００８】

図２は入力した画像を登録するまでの流れを示す図である。同じ構成要素には同じ参照番号を付して説明する。本実施形態は、画像の属性情報に抽出されたキーワードと、キーワードを正規化した文字列の両方を登録し、表示時にこれらの文字列をハイライトすることによって実施される。まず、スキャンされた原稿１１は、文字抽出モジュール２によりＯＣＲ処理が行われる。ＯＣＲ処理では原稿内の文字情報１２を取り出す。文字情報１２には、文字コードの他、文字の位置、大きさが含まれる。例えば、「米ｘ＝０，ｙ＝０，ｗ＝８，ｈ＝８」という情報は、「米」という文字が画像左上から（０．０）の位置にあり、文字幅と高さは８画素であることを表している。次に、単語抽出モジュール３により単語抽出処理が行われ、ＯＣＲによって抽出された文字コードを単語ごとに区切り、その結果を出力する。これらの単語１３は、検索モジュール４に登録される。検索モジュール４では、異表記正規化によって類似した表記の単語をまとめる処理を行った上で、正規化済みの単語から検索用のインデックスを作成する（符号１４）。また、正規化した単語については、正規化情報（変更された単語）を通知する。画像属性作成モジュール７は、抽出した文字情報と、正規化されたキーワードから、キーワードの位置情報を計算し、画像に登録する属性情報１５を作成する。このとき、正規化によって生成された文字については、識別可能な情報を埋め込む（属性情報１５の、ｏ＝Ｔの箇所）。

図３は登録までの流れを示すフローチャートである。まず、原稿１１がスキャンされて入力される（Ｓ１）。次に文字抽出モジュール２によりＯＣＲ処理が行われる（Ｓ２）。ＯＣＲ処理では原稿内の文字情報１２を取り出す。文字情報１２には、文字コードの他、文字の位置、大きさが含まれる。例えば、「米ｘ＝０，ｙ＝０，ｗ＝８，ｈ＝８」という情報は、「米」という文字が画像左上から（０．０）の位置にあり、文字幅と高さは８画素であることを表している。次に、単語抽出モジュール３により単語抽出処理が行われ、ＯＣＲによって抽出された文字コードを単語ごとに区切り、その結果を出力する（Ｓ３）。これらの単語１３は、検索モジュール４に登録される（Ｓ４）。検索モジュール４では、異表記正規化によって類似した表記の単語をまとめる処理を行った上で、正規化済みの単語から検索用のインデックスを作成する（符号１４）。また、正規化した単語については、正規化情報（変更された単語）を通知する。画像属性作成モジュール７は、抽出した文字情報と正規化されたキーワードからキーワードの位置情報を計算し、画像に登録する属性情報１５を作成する（Ｓ５）。このとき、正規化によって生成された文字については、識別可能な情報を埋め込む（属性情報１５の、ｏ＝Ｔの箇所）。属性情報１５の画像を登録して終了する（Ｓ６）。

【０００９】

図４は検索からハイライト表示までの処理を示す図である。操作者が「メモリ」という検索条件を検索条件入力モジュール５により入力すると、検索条件入力モジュールから検索キーが検索モジュール４に渡され、ヒットした文書の一覧２１が検索結果一覧画面に表示される。操作者が見たい画像を指定すると、画像、画像属性情報、検索キーワードがハイライト表示モジュール２３に渡される。ハイライト表示モジュール２３は、画像属性情報２２の中に、検索キーワードが含まれるかを走査する。この場合、必ず画像属性情報２２の中に、検索条件の文字列が存在することになる。見つかった文字列の座標に対応する範囲をハイライト指定（２３ａ）を行い、画像表示モジュール９が実際に画像を表示する。

図５は画像表示までの流れを示すフローチャートである。操作者が「メモリ」という検索条件を検索条件入力モジュール５により入力すると（Ｓ１１）、検索条件入力モジュールから検索キーが検索モジュール４に渡され（Ｓ１２）、ヒットした文書の一覧２１が検索結果一覧画面に表示される（Ｓ１３）。操作者が見たい画像を指定すると（Ｓ１４）、画像、画像属性情報、検索キーワードがハイライト表示モジュール２３に渡される（Ｓ１

５、１６）。ハイライト表示モジュール２３は、画像属性情報２２の中に、検索キーワードが含まれるかを走査する。この場合、必ず画像属性情報２２の中に、検索条件の文字列が存在することになる。見つかった文字列の座標に対応する範囲をハイライト指定（２３ａ）を行い（Ｓ１７）、画像表示モジュール９が実際に画像を表示する（Ｓ１８）。

【００１０】

図６は本発明の画像処理のフローチャートである。本発明はハイライト箇所で、画像属性情報のうち、正規化によって生成されたキーか否かを判断し、正規化されたキーと、正規化されていないキーとで色を分けて表示する。まず画像、画像属性情報、検索キーワードがハイライト表示モジュール２３に渡される（Ｓ２１）。その結果、正規化されたキーワードか否かを判断し（Ｓ２２）、正規化されたキーワードであると画像内の文字列を反転表示する（Ｓ２３）。一方ステップＳ２２で正規化されたキーワードでないと画像内の文字列をマーキング表示する（Ｓ２５）。

以上の通り本発明によれば、画像の属性情報に抽出されたキーワードと、キーワードを正規化した文字列の両方を登録し、表示時にこれらの文字列をハイライトするので、ハイライト表示時に逆正規化の漏れによって、ハイライトされないケースが防止できる。

また、実際に操作者が入力したキーと、ハイライト表示される文字列が異なる場合でも、ハイライト表示色やハイライト表示形式を変えることにより、操作者に対して理解しやすいインターフェースを提供することができる。

また、単語抽出モジュール３は、自然言語で書かれた文を言語で意味を持つ最小単位の列に分割し、品詞を見分ける形態素解析により単語を抽出するので、確実に正確な単語を抽出することができる。

また、検索モジュール４は、複数の表記をまとめて一つの表記として扱う表記正規化法により正規化したキーワードに基づいて検索するので、逆に一つの単語から複数の正規化単語を抽出する際に、有効に作用して的確な正規化単語を抽出することができる。

【図面の簡単な説明】

【００１１】

【図１】本発明の文書管理システムのモジュール構成を示す図。

【図２】入力した画像を登録するまでの流れを示す図。

【図３】登録までの流れを示すフローチャート。

【図４】検索からハイライト表示までの処理を示す図。

【図５】画像表示までの流れを示すフローチャート。

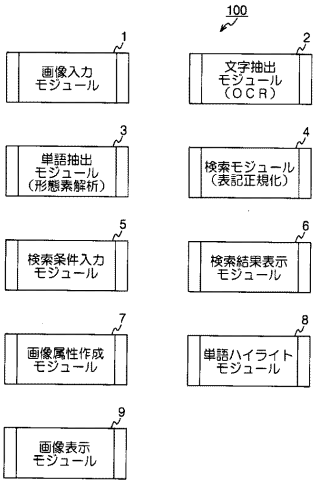
【図６】本発明の画像処理のフローチャート。

【符号の説明】

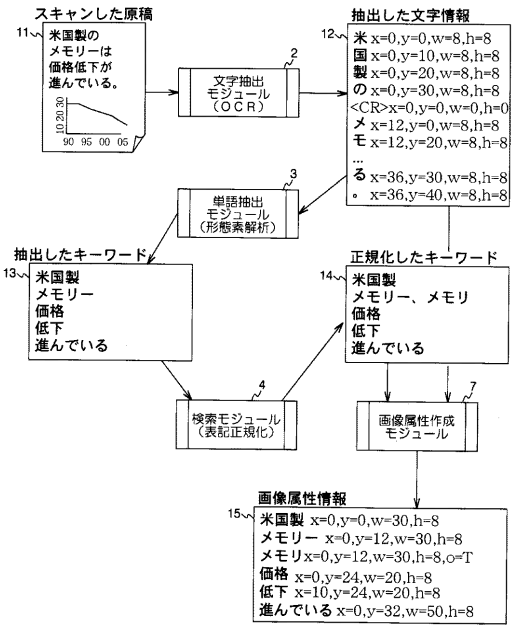
【００１２】

１　画像入力モジュール、２　文字抽出モジュール、３　単語抽出モジュール、４　検索モジュール、５　検索条件入力モジュール、６　検索結果表示モジュール、７　画像属性作成モジュール、８　単語ハイライトモジュール、９　画像表示モジュール、１００　文書管理システム

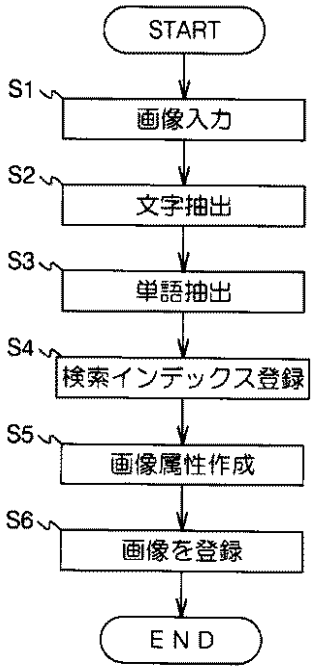
【図１】


100
1
画像入力モジュール
2
文字抽出モジュール（ＯＣＲ）
3
単語抽出モジュール（形態素解析）
4
検索モジュール（表記正規化）
5
検索条件入力モジュール
6
検索結果表示モジュール
7
画像属性作成モジュール
8
単語ハイライトモジュール
9
画像表示モジュール

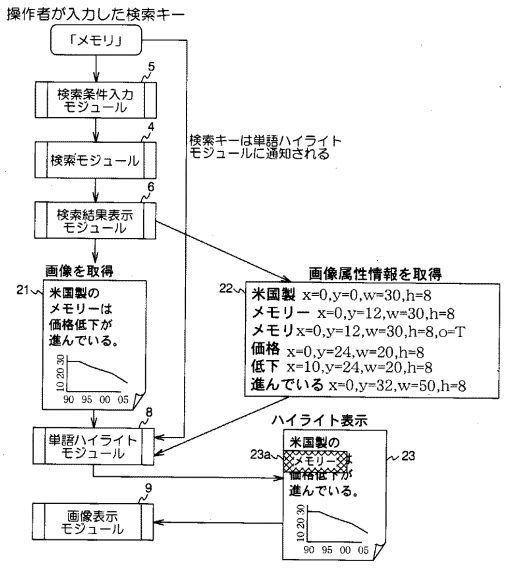
【図２】


スキャンした原稿
11
米国製のメモリーは価格低下が進んでいる。
文字抽出モジュール（ＯＣＲ）
2
抽出した文字情報
12
米 x=0,y=0,w=8,h=8
国 x=0,y=10,w=8,h=8
製 x=0,y=20,w=8,h=8
の x=0,y=30,w=8,h=8
<CR>x=0,y=0,w=0,h=0
メ x=12,y=0,w=8,h=8
モ x=12,y=20,w=8,h=8
...
る x=36,y=30,w=8,h=8
。 x=36,y=40,w=8,h=8
3
単語抽出モジュール（形態素解析）
抽出したキーワード
13
米国製
メモリー
価格
低下
進んでいる
正規化したキーワード
14
米国製
メモリー、メモリ
価格
低下
進んでいる
4
検索モジュール（表記正規化）
7
画像属性作成モジュール
画像属性情報
15
米国製 x=0,y=0,w=30,h=8
メモリー x=0,y=12,w=30,h=8
メモリ x=0,y=12,w=30,h=8,o=T
価格 x=0,y=24,w=20,h=8
低下 x=10,y=24,w=20,h=8
進んでいる x=0,y=32,w=50,h=8

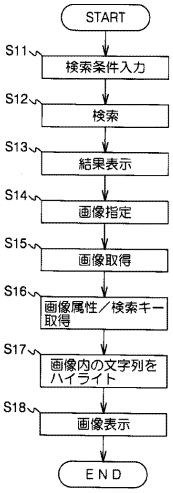
【図３】


START
S1
画像入力
S2
文字抽出
S3
単語抽出
S4
検索インデックス登録
S5
画像属性作成
S6
画像を登録
END

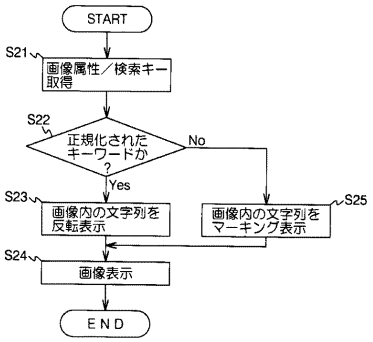
【図４】


操作者が入力した検索キー
「メモリ」
5
検索条件入力モジュール
4
検索モジュール
検索キーは単語ハイライトモジュールに通知される
6
検索結果表示モジュール
画像を取得
21
米国製のメモリーは価格低下が進んでいる。
画像属性情報を取得
22
米国製 x=0,y=0,w=30,h=8
メモリー x=0,y=12,w=30,h=8
メモリ x=0,y=12,w=30,h=8,o=T
価格 x=0,y=24,w=20,h=8
低下 x=10,y=24,w=20,h=8
進んでいる x=0,y=32,w=50,h=8
8
単語ハイライトモジュール
ハイライト表示
23a
米国製のメモリーは価格低下が進んでいる。
23
9
画像表示モジュール

【図５】


START
S11
検索条件入力
S12
検索
S13
結果表示
S14
画像指定
S15
画像取得
S16
画像属性／検索キー取得
S17
画像内の文字列をハイライト
S18
画像表示
END

【図６】


START
S21
画像属性／検索キー取得
S22
正規化されたキーワードか？
No
Yes
S23
画像内の文字列を反転表示
S25
画像内の文字列をマーキング表示
S24
画像表示
END