(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2004-7552
(P2004-7552A)
(43)公開日　平成16年1月8日(2004.1.8)

(51)Int.Cl.[7]	F I		テーマコード（参考）
H04N 1/40	H04N 1/40	Z	5B057
G06F 17/60	G06F 17/60	142	5C077
G06T 1/00	G06T 1/00	310Z	

審査請求 未請求　請求項の数 39　ＯＬ　（全 27 頁）

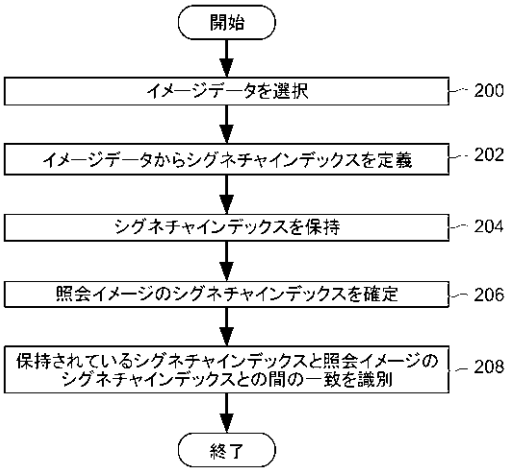
(21)出願番号　特願2003-101866(P2003-101866)
(22)出願日　平成15年4月4日(2003.4.4)
(31)優先権主張番号　10/263423
(32)優先日　平成14年10月1日(2002.10.1)
(33)優先権主張国　米国(US)
(31)優先権主張番号　60/372208
(32)優先日　平成14年4月12日(2002.4.12)
(33)優先権主張国　米国(US)
(31)優先権主張番号　60/384584
(32)優先日　平成14年5月31日(2002.5.31)
(33)優先権主張国　米国(US)

(71)出願人　000002369
セイコーエプソン株式会社
東京都新宿区西新宿２丁目４番１号
(74)代理人　100095728
弁理士　上柳　雅誉
(74)代理人　100107076
弁理士　藤綱　英吉
(74)代理人　100107261
弁理士　須澤　修
(72)発明者　チャンギック　キム
アメリカ合衆国　カリフォルニア州　クパティノ　パークウッド　ドライブ　１０２１５、７番

最終頁に続く

(54)【発明の名称】コンテントベースのイメージコピー検出の方法並びに装置

(57)【要約】
【課題】コンテントベースのイメージコピー検出の方法並びに装置
【解決手段】ディジタルイメージのコピーを検出するための方法は、ディジタルイメージの疑わしいコピーをサブイメージに分割することから始まる。次に、サブイメージそれぞれと関連した平均の濃度を判定する。次に、その平均の濃度を一連の係数に変換する。次に、その一連の係数からランクマトリックスを定義する。次に、疑わしいコピーがディジタルイメージの本当のコピーかどうかを判定するために、一連の係数に基づくランクマトリックスが照会イメージのランクマトリックスと比較する。コンピュータ可読媒体と、コンピュータシステム、及びコンピュータシステムで実行できるように構成されたコンピュータコードも提供する。
【選択図】図１１



【特許請求の範囲】
【請求項１】
ディジタルイメージのコピーを検出するための方法で、
ディジタルイメージの疑わしいコピーをサブイメージに分割するステップと、
前記サブイメージの各々と関連した画素の平均濃度を判定するステップと、
前記平均濃度を一連の係数に変換するステップと、
前記一連の係数からランクマトリックスを定義するステップと、
前記疑わしいイメージが前記ディジタルイメージの本当のコピーかどうかを判定するために前記一連の係数に基づくランクマトリックスを照会イメージのランクマトリックスと比較するステップとを有することを特徴とする方法。
【請求項２】
前記サブイメージは各々が８×８のブロックからなり、前記ディジタルイメージは同じサイズの６４のサブイメージを有することを特徴とする請求項１に記載の方法。
【請求項３】
前記一連の係数は離散コサイン変換（ＤＣＴ）関数のＡＣ係数を含むことを特徴とする請求項１に記載の方法。
【請求項４】
前記方法の前記平均濃度を一連の係数に変換するオペレーションは、
２次元の離散コサイン変換（ＤＣＴ）を行なうことによって２次元の一連の係数を生成するステップ含むことを特徴とする請求項１に記載の方法。
【請求項５】
前記方法の前記一連の係数からランクマトリックスを定義するオペレーションは、
前記一連の係数を１次元のランクマトリックスに配列するステップを含むことを特徴とする請求項１に記載の方法。
【請求項６】
前記１次元ランクマトリックスは、前記一連の係数の大きさを降順に配置していることを特徴とする請求項５に記載の方法。
【請求項７】
コンテントベースのイメージコピー検出方法であって、
イメージと関連したイメージデータを選択するステップと、
前記イメージデータからシグネチャインデックスを定義するステップと、
前記イメージデータからのシグネチャインデックスを保持するステップと、
照会イメージのシグネチャインデックスを判定するステップと、
前記イメージデータからの保持されたシグネチャインデックスと前記照会イメージのシグネチャインデックスとの間の一致を識別するステップとを備えることを特徴とする方法。
【請求項８】
誤った検出を最小限に抑えるために、前記イメージデータからのシグネチャインデックスのサイズと照会イメージのシグネチャインデックスのサイズが最適化されることを特徴とする請求項７に記載の方法。
【請求項９】
前記イメージデータからのシグネチャインデックス及び照会イメージのシグネチャインデックスは、共に２４、３５、４８、６３からなるグループから選択された係数の数を有するシグネチャインデックスによって定義されることを特徴とする請求項７に記載の方法。
【請求項１０】
前記イメージデータのシグネチャインデックスは、データベースに保持されることを特徴とする請求項７に記載の方法。
【請求項１１】
前記方法の前記イメージと関連したイメージデータを選択するオペレーションは、
分散型ネットワークを探索して前記イメージデータを捜すステップと、
複数のシグネチャインデックスからクラスタを定義するステップとを含むことを特徴とす

る請求項７に記載の方法。
【請求項１２】
前記クラスタを定義するためにｋ平均値アルゴリズムが用いることを特徴とする請求項１１に記載の方法。
【請求項１３】
クラスタ数はクラスタ妥当性分析によって判定されることを特徴とする請求項１１に記載の方法。
【請求項１４】
前記クラスタは、クラスタの重心を含むことを特徴とする請求項１１に記載の方法。
【請求項１５】
さらに、前記クラスタが調査されているかどうか判定するために、前記照会イメージのシグネチャインデックスを前記クラスタの重心と比較するステップを有することを特徴とする請求項１４に記載の方法。
【請求項１６】
ディジタルイメージの無許可のコピーを見つけるための方法であって、
テストイメージのランクマトリックスを計算するステップと、
照会イメージのランクマトリックスを計算するステップと、
しきい値を判定するステップとを備え、当該しきい値は前記テストイメージが前記照会イメージのコピーかどうかを示すものであり、さらに
前記テストイメージの前記ランクマトリックスと前記照会イメージの前記ランクマトリックスとの間の距離に関連した距離値を判定するステップを備えると共に、前記距離値を前記しきい値と比較するステップを備え、
もし前記距離値が前記しきい値に満たなければ、前記テストイメージは前記照会イメージのコピーであることを特徴とする方法。
【請求項１７】
前記方法の前記しきい値を判定するオペレーションは、
前記しきい値を算出するために最大事後（ＭＡＰ）基準を適用するステップを含むことを特徴とする請求項１６に記載の方法。
【請求項１８】
前記しきい値は事前確率の比と反比例することを特徴とする請求項１６に記載の方法。
【請求項１９】
さらに、前記テストイメージのランクマトリックスと前記照会イメージのランクマトリックスを２次元離散コサイン変換によって算出するステップを含むことを特徴とする請求項１６に記載の方法。
【請求項２０】
コンピュータシステムで実施されるように構成されたコンピュータコードを含んだコンピュータ可読媒体であって、当該コンピュータコードは、イメージデータを識別するためのプログラム命令と、
識別されたイメージデータから特徴ベクトルを定義するためのプログラム命令と、
前記特徴ベクトルを保持するためのプログラム命令と、
照会イメージの特徴ベクトルと前記識別されたイメージに基づく特徴ベクトルとの一致を判定するためのプログラム命令と、
前記照会イメージの特徴ベクトルと前記識別されたイメージからの特徴ベクトルとの一致を表示するためのプログラム命令とを有することを特徴とする。
【請求項２１】
前記特徴ベクトルはランクマトリックスであることを特徴とする請求項２０に記載のコンピュータ可読媒体。
【請求項２２】
前記特徴ベクトルを定義するプログラム命令は２次元離散コサイン変換（ＤＣＴ）を行なうためのプログラム命令を含むことを特徴とする請求項２０に記載のコンピュータ可読媒

体。
【請求項２３】
前記一致を判定するためのプログラム命令はクラスタベースの検出のためのプログラム命令を含むことを特徴とする請求項２０に記載のコンピュータ可読媒体。
【請求項２４】
前記クラスタベースの検出のためのプログラム命令はｋ平均値クラスタリングプログラム命令であることを特徴とする請求項２３に記載のコンピュータ可読媒体。
【請求項２５】
テストイメージが照会イメージのコピーかどうかを判定するためのコンピュータ可読媒体であって、当該コンピュータコードは、
照会イメージと関連したランクマトリックス及びテストイメージと関連したランクマトリックスとを算出するためのプログラム命令と、
前記テストイメージが前記照会イメージのコピーであるかどうかを示すしきい値を判定するためのプログラム命令と、
前記照会イメージと関連したランクマトリックスを前記テストイメージと関連したランクマトリックスと比較するためのプログラム命令とを備え、当該比較するためのプログラム命令は、
前記テストイメージと前記照会イメージとの間の距離値を判定するためのプログラム命令と、
前記距離値が前記しきい値に満たないかどうかを調べるためのプログラム命令とを含み、
もし前記距離値が前記しきい値に満たなければ前記テストイメージは前記照会イメージのコピーであることを特徴とする。
【請求項２６】
前記テストイメージが照会イメージのコピーかどうかを示すしきい値を判定するためのプログラム命令は、
前記しきい値を算出するために最大事後（ＭＡＰ）基準を適用するためのプログラム命令を含むことを特徴とする請求項２５に記載のコンピュータ可読媒体。
【請求項２７】
前記照会イメージと関連したランクマトリックスとテストイメージと関連したランクマトリックスとを算出するためのプログラム命令は、
２次元離散コサイン変換を適用することによって各ランクマトリックスを算出するためのプログラム命令を含むことを特徴とする請求項２５に記載のコンピュータ可読媒体。
【請求項２８】
前記しきい値は最適適合値及び最適呼出値と関連していることを特徴とする請求項２５に記載のコンピュータ可読媒体。
【請求項２９】
ディジタルイメージのコピーを検出するためのプログラム命令を有するコンピュータ可読メディアであって、
ディジタルイメージの疑わしいコピーをサブイメージに分割するためのプログラム命令と、
前記サブイメージの各々と関連した平均濃度を判定するためのプログラム命令と、
前記平均濃度を一連の係数に変換するためのプログラム命令と、
前記一連の係数からランクマトリックスを定義するためのプログラム命令と、
前記疑わしいコピーが前記ディジタルイメージの本当のコピーかどうかを判定するために前記一連の係数に基づくランクマトリックスを照会イメージのランクマトリックスと比較するためのプログラム命令とを有することを特徴とするコンピュータ可読メディア。
【請求項３０】
前記平均濃度を一連の係数に変換するためのプログラム命令は、２次元の離散コサイン変換（ＤＣＴ）を行なうことによって２次元の一連の係数を生成するためのプログラム命令を含むことを特徴とする請求項２９に記載のコンピュータ可読メディア。

【請求項３１】
前記一連の係数からランクマトリックスを定義するためのプログラム命令は、
前記一連の係数を１次元のランクマトリックスに配列するためのプログラム命令を含むことを特徴とする請求項２９に記載のコンピュータ可読メディア。
【請求項３２】
さらに、複数のランクマトリックスからクラスタを定義するためのプログラム命令を含むことを特徴とする請求項２９に記載のコンピュータ可読メディア。
【請求項３３】
さらに、前記クラスタが探索されているかどうかを判定するために照会イメージのランクマトリックスを前記クラスタの重心のランクマトリックスと比較するためのプログラム命令を含むことを特徴とする請求項３２に記載のコンピュータ可読メディア。
【請求項３４】
コンピュータシステムであって、当該コンピュータシステムは、イメージデータをアセンブルするためのデータベース生成システムを備え、当該データベース生成システムは、
前記イメージデータを識別するためのイメージ収集システムと、
前記イメージデータのシグネチャインデックスを抽出するための特徴抽出コードとを含み、前記コンピュータシステムはさらに、
前記イメージデータを照会データと突き合せるためのデータベース照会システムを備え、当該データベース照会システムは、
前記イメージデータのシグネチャインデックスを保持するように構成されたデータベースと、
前記イメージデータが前記照会データのコピーかどうかを判定するために前記照会データのシグネチャインデックスと前記イメージデータのシグネチャインデックスとの間の一致を識別するように構成された特徴マッチングシステムとを含むことを特徴とする。
【請求項３５】
前記イメージ収集システムはウェブクローラであることを特徴とする請求項３４に記載のコンピュータシステム。
【請求項３６】
前記シグネチャインデックスはランクマトリックスであることを特徴とする請求項３４に記載のコンピュータシステム。
【請求項３７】
前記特徴マッチングシステムは、前記データベースと関連したクラスタが探索されているかどうかを判定するための機能を有することを特徴とする請求項３４に記載のコンピュータシステム。
【請求項３８】
前記シグネチャインデックスは離散コサイン変換（ＤＣＴ）関数を用いて導出されることを特徴とする請求項３４に記載のコンピュータシステム。
【請求項３９】
前記特徴マッチングシステムは、前記イメージデータが前記照会イメージのコピーであるかどうか判定する際に使用するしきい値を定義することを特徴とする請求項３４に記載のコンピュータシステム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、一般にディジタルビデオ技術に関し、より具体的にはイメージのコンテンツに基づいて無許可のディジタルイメージのコピーを検出する方法並びに装置に関する。
【０００２】
【従来の技術】
インターネットの普及とコスト・性能比の高いディジタル記憶装置が幅広く利用できるようになったことで、手軽にディジタルコンテンツを複製、送信、配布できるようになった

。そのために、特にディジタルイメージの著作権に関する知的所有権（ＩＰＲ）の保護がきわめて重要な法的問題になってきている。とりわけ、ディジタルメディア（イメージ、音声、ビデオ）のコピー検出は、著作権違反を調査する者にとって必須の要件である。使用する２つのコピー検出アプリケーションには、使用追跡と著作権違反の取締りがある。

【０００３】

現在、ディジタルイメージの著作権の保護に広く採用されているのは、透かし模様挿入とコンテントベースのコピー検出という２つの方法である。一般に知られているように、透かし模様挿入では配布前にイメージに情報を埋め込む。従って、所有権者を証明するために、マーク入りコンテントのコピーには全て抽出可能な透かし模様が入っている。コンテントベースのコピー検出では、イメージ自体のほかには何も情報は要らない。一般に、イメージにはコピー、それもとりわけ違法に配布されたコピーを検出するために使用可能な独自な情報が十分に含まれている。例えば、もしもあるイメージの所有権者が、そのイメージがインターネットで違法配布されているという疑いをもったら、所有権者はコピー検出システムに照会を求めることができる。なお、コンテントベースのコピー検出は透かし模様挿入を補う相補的方法としても使用できる。コピー検出器が作成者又は配布者に容疑者リストを提供すると、そのメディアの実際の所有権者は、所有権者を証明するために透かし模様やその他の認証技法を用いることができる。

【０００４】

コンテントベースのコピー検出方法では、オリジナルイメージから署名（シグネチャ）が抽出される。テストイメージがオリジナルイメージのコピーであるかどうか判定するために、テストイメージから抽出された署名がオリジナルイメージの署名と比較される。透かし模様挿入と比べた場合のコンテントベースコピー検出の重要な特徴は、イメージを配布する前にシグネチャ抽出を行なわなくてよい点である。しかしながら、オリジナルと同じでないコピー、すなわち、わずかに変更されたコピーは検出できない可能性がある。例えば、コピー検出を回避するために、あるいは画質の向上を図って、第３者が多様な変更版を生成する可能性があるが、コンテントベースのコピー検出では、そうした変更が原因でコピーを検出できなくなる可能性がある。

【０００５】

ヒストグラム交差メソッドなど、カラーヒストログラムをベースにした方法がコンテントベースのイメージ検索システムで使用されている。しかしながら、そうした方法は、カラーヒストグラムがカラーの空間分布に関する情報を保持していないため、コピー検出システムには適さない。画素位置を考慮に入れることができる別の方法として、パーティションをベースにしたアプローチがある。この方法では、イメージがサブイメージに分割される。そうした方法の一つでは、各パーティションのカラー情報が局部カラーヒストグラムによって取得される。２つのイメージの類似性が、その局部カラーヒストグラムを比較し、全てのサブイメージの類似性を考慮することによって測定される。しかしながら、この方法は計算コストが高く、所要探索時間が長い。加えて、この方法では、空間的アウトレイが変更されたイメージを検出することはできない。

【０００６】

コンテントベースのイメージ検索とイメージコピー検出とでは根本的な違いがある。イメージコピー検出器は、照会イメージの全コピーを調べるのに対し、コンテントベースのイメージ検索システムでは一般に色に関する類似イメージを調べる。例えば、図１に示す３つのイメージは、イメージコピー検出器のサーチとコンテントベースのイメージ検索システムの違いを説明するものである。イメージ１０２はオリジナル（又は照会）イメージで、イメージ１０４は色に関してイメージ１０２に類似したもの、イメージ１０６はイメージ１０２の色相が変えられたものである。色をベースにしたイメージ検索システムでは、イメージ１０４は、イメージ１０６よりもイメージ１０２との関連性が高いと見なされるが、イメージコピー検出器では、イメージ１０６がイメージ１０２のコピーの可能性が高いと見なされる。従って、使用するコピー検出方法によって、得られる結果が異なり、ある種のイメージコピーは、使用した検出方法の一部もしくは全てで検出されないことがあ

る。言い換えれば、ある程度の変更を含んだコピーを検出できるイメージコピー検出器でなければならない。変更には、輝度や彩度の変化、色相のずれ、回転や反転（ｆｌｉｐｐｉｎｇ）などを含む空間的歪みがある。そうした変更を保護するための一つの提案は、ウェブでウェーブレットをベースにした複写イメージ検出を取り入れることである。しかしながら、ウェーブレットベースの方法は、反転や回転が入ったコピーを検出することはできない。

【０００７】

それゆえに、上記のような従来技術のかかえる問題を解決する、オリジナルディジタルイメージから空間的アウトレイや色が変更されたコピーも検出できるような、確実（ロバスト）かつ効率的なコンテントベースのイメージコピー検出の方法や装置が求められている。

【０００８】

【特許文献１】

米国再発行特許発明６，３４７，１４４号明細書

【特許文献２】

米国再発行特許発明６，３３２，０３０号明細書

【特許文献３】

米国再発行特許発明６，２８５，７７４号明細書

【特許文献４】

米国再発行特許発明６，２６６，４２９号明細書

【０００９】

【発明が解決しようとする課題】

おおまかに言えば、本発明は、空間的に変更されたコピーや色が変更されたコピーを検出可能な、コンテントベースのコピー検出方法並びに装置を提供することによって上記のニーズに応えるものである。なお、本発明は、方法、システム、コンピュータコード、あるいはデバイスとしてなど、その実施方法は色々とある。本発明を以下にいくつかの実施例をあげて説明する。

【００１０】

【課題を解決するための手段】

一つの実施例において、ディジタルイメージのコピーを検出するための方法を提供する。この方法は、疑わしいディジタルイメージのコピーをサブイメージに分割することから始まる。その次に、サブイメージの各々の画素と関連した平均濃度が判定される。次に、その平均濃度は一連の係数に変換される。その次に、その一連の係数からランクマトリックスが定義される。次に、疑わしいコピーがディジタルイメージの本当のコピーかどうかを判定するために、その一連の係数に基づくランクマトリックスが、照会イメージのランクマトリックスと比較される。

【００１１】

別の実施例では、コンテントベースのイメージコピー検出方法を提供する。この方法は、イメージと関連したイメージデータを選択することから始まる。次に、イメージごとに、イメージデータからシグネチャインデックスが定義される。次に、そのイメージデータごとのシグネチャインデックスが保持される。その次に、照会イメージのシグネチャインデックスが判定される。次に、保持されたイメージごとのイメージデータのシグネチャインデックスの中から、照会イメージのシグネチャインデックスと一致するものが識別される。

【００１２】

さらに別の実施例では、ディジタルイメージの無許可コピーを見つけ出す方法を提供する。この方法は、テストイメージのランクマトリックスを計算することから始まる。次に、照会イメージのランクマトリックスが計算される。次に、しきい値が判定される。このしきい値は、テストイメージが照会イメージのコピーかどうかを示すものである。次に、テストイメージのランクマトリックスと照会イメージのランクマトリックスとの間の距離と

関連した距離値が判定される。次に、その距離値をしきい値と比較する。もし距離値がしきい値に満たなければ、テストイメージは照会イメージのコピーとなる。

【００１３】

さらに別の実施例では、コンピュータシステムで実行するように構成されたコンピュータコードを提供する。このコンピュータコードは、イメージデータを識別するためのプログラム命令を含んでいる。コンピュータコードは、識別されたイメージデータから特徴ベクトルを定義するためのプログラム命令も含む。その特徴ベクトルをデータベースに保持するためのプログラム命令も含む。照会イメージの特徴ベクトルと、識別されたイメージの特徴ベクトルとの一致を判定するためのプログラム命令も含む。その照会イメージの特徴ベクトルと識別されたイメージの特徴ベクトルとの一致を表示するためのプログラム命令も含む。

【００１４】

また別の実施例では、テストイメージが照会イメージのコピーかどうかを判定するコンピュータコードを提供する。このコンピュータコードは、照会イメージと関連したランクマトリックス並びにテストイメージと関連したランクマトリックスを算出するためのプログラム命令を含む。テストイメージが照会イメージのコピーかどうかを示すしきい値を判定するためのプログラム命令も提供する。照会イメージと関連したランクマトリックスを、テストイメージと関連したランクマトリックスと比較するためのプログラム命令を含む。比較するためのプログラム命令は、テストイメージと照会イメージとの間の距離値を判定するためのプログラム命令と、その距離値がしきい値に満たないかどうかを調べるためのプログラム命令とを含む。距離値がしきい値に満たない場合には、テストイメージは照会イメージのコピーとなる。

【００１５】

また別の実施例では、ディジタルイメージのコピーを検出するためのプログラム命令を有する、コンピュータで判読可能なメディアを提供する。このコンピュータ可読メディアは、疑わしいディジタルイメージのコピーをサブイメージに分割するためのプログラム命令と、そのサブイメージと関連した平均濃度を判定するためのプログラム命令とを含む。平均濃度を一連の係数に変換するためのプログラム命令も含む。その一連の係数から、ランクマトリックスを定義するためのプログラム命令を含み、疑わしいコピーがディジタルイメージの本当のコピーかどうかを判定するために、一連の係数のランクマトリックスと照会イメージのランクマトリックスとを比較するためのプログラム命令を提供する。

【００１６】

また別の実施例では、コンピュータシステムを提供する。このコンピュータシステムはイメージデータをアセンブルするためのデータベース生成システムを含み、このデータベース生成システムは、イメージデータを識別するためのイメージ収集システムと、イメージデータのシグネチャインデックスを抽出するための特徴抽出コードとを含む。イメージデータを照会データと突き合せるためのデータベース照会システムを提供する。このデータベース照会システムは、イメージデータのシグネチャインデックスを保持するように構成されたデータベースと、照会データのシグネチャインデックスとイメージデータのシグネチャインデックスとの間の一致を識別するように構成された特徴マッチングシステムとを含む。

【００１７】

本発明のその他の態様並びに効果は、本発明の原理を例をあげて示した添付の図面とともに以下の詳細な説明から明白になる。

【００１８】

【発明の実施の形態】

ディジタルイメージの無許可コピーを検出するためのシステム、装置、並びに方法として発明を説明する。しかしながら、このような具体的な説明を部分的に或いは全て知らなくても本発明を実施できることは、当業者であれば明白である。また、本発明の説明を不要に曖昧なものにしないために、既によく知られているプロセスオペレーションについては

詳細に説明しない。

【００１９】

本発明の実施例は、ディジタルデータのコピーを効率よく検出するためのアルゴリズムを使用したシステムを提供するものである。以下に詳説する通り、色や目に見える様相だけでなく、空間的構成が変更されたコピーも識別することができる。例えば、インターネットなど分散型ネットワークで見つけた照会イメージの疑わしいコピーを、ウェブクローラ（ｗｅｂ　ｃｒａｗｌｅｒ）によって突き止める。疑わしいコピーは、８×８のサブイメージにサイズ変更される。サブイメージは、２次元離散コサイン変換（ＤＣＴ）され、順序測度（ｏｒｄｉｎａｌ　ｍｅａｓｕｒｅ）とも呼ばれるランクマトリックス、シグネチャインデックス又は特徴ベクトルが生成される。疑わしいコピーのランクマトリックスがデータベースに保持され、その疑わしいコピーが本当のコピーかどうかを判定するために、各ランクマトリックスを照会イメージのランクマトリックスと比較する。一例として、最適な識別能力を提供するために、ランクマトリックスのサイズが指定される。本願で説明する実施例は、以下に詳説するように、効率の高いマッチングプロセスを提供するためにクラスタリングメソッドを活用することもできる。

【００２０】

イメージを突き合せる基本的な方法は、相関に基づく方法で、それは画素の差の和に基づいている。I_1およびI_2が２つのイメージにおける濃度を表わすとする。Ｎの組$(I_1^1, I_2^1), \cdots, (I_1^n, I_2^n), \cdots, (I_1^N, I_2^N)$が存在する。ここで、Ｎは一つのイメージにおける画素数を示している。数量$\frac{\sum_{i=1}^{N} |I_1^i - I_2^i|}{N}$は$(I_1, I_2)$間の距離である。Ｎがブロック数を示している場合、$I_1^n$をｎ番目のブロックの平均濃度と考えることもできる。しかしながら、これらの測度は、中心から離れたところにある一つの画素（又はブロック）によって任意に歪むことがあるので確実でない。さらに、対応する画素に非線形の濃度ばらつきがあるときには適さない。

【００２１】

一実施例として、ステレオイメージマッチングの順序測度を用いて、要求される確実性を実現する。順序測度を用いる一つの例において、イメージはｍ×ｎの同じサイズのブロックに分割されるため、システムは入力イメージサイズに依存しない。各ブロックの平均値をとることにより、ｍ×ｎのサブイメージが算出される。図２ａ～図２ｃは、コピー検出のための順序測度を図形的に表わしたものである。図２ａは、ｍ×ｎ（３×３）の同じサイズのブロックに分割されたイメージを表わしている。各ブロックの画素濃度の平均値が算出され、図２ｂの各ブロックの中に示されている。図２ｃに示すように、平均値の配列がランクマトリックスに変換される。ここで、ランクマトリックスのサイズはｍ×ｎ又は１×ｍｎである。従って、図２ｂに示した濃度値がコピーされたイメージにおいて、そのサブイメージの値が｛｛３０，　６０，　４０｝，　｛７０．　９０，　１１０｝，　｛５０，　１００，　８０｝｝になるように一様に変えても、つまり、ブロックそれぞれの濃度値を２０ずつ増やしても、ランクマトリックスは変わらない。すなわち、画素値のランクマトリックスは図２ｃに示したものと全く同じになり、オリジナルイメージとの完全整合が達成できる。

【００２２】

しかしながら、先に説明したように、階数測度を使用しても、コピーが不規則に変更されていると検出することができない。特に、イメージの空間的アウトレイが変更されている場合は検出が難しい。図３は、図１のイメージ１０２を空間的に変更したイメージである

。図１のランドスケープイメージが、図３のイメージ１０８に示すように、水平方向に反転されている。図３の変更イメージは、上述の濃度値の順序測度を使用しても検出することができない。以下に、１次元のランクマトリックスを定義するために、２次元の離散コサイン変換から導出されたＡＣ係数の順序測度を用いた実施例について説明する。この１次元のランクマトリックスは、コピーを検出するために、照会イメージと比較される特徴ベクトルとして使用される。

【００２３】

図４は、本発明の一実施例に係る、コンテントベースのコピー検出システムの概念図である。ウェブクローラ１４２は、インターネットなど分散型ネットワークを探索して照会イメージのコピーを捜す。なお、ウェブクローラ１４２は、イメージのコピーを探索する機能を有する適切なウェブクローラなら何でもよい。ウェブクローラ１４２が照会イメージのコピーを識別したら、そのコピーがランクマトリックスジェネレータ１４４に届けられる。ランクマトリックスジェネレータ１４４は、そのコピーのイメージデータを操作してランクマトリックスを得る。ランクマトリックスは、イメージコピーのディジタル指紋と考えることができる。一例として、イメージデータの８×８表現に２次元離散コサイン変換（ＤＣＴ）を行なうことによってランクマトリックスを導出する。その後、２つのイメージ間の順序測度を用いて、そのイメージが照会イメージのコピーかどうかを判定する。ここで、順序測度とは、２次元離散コサイン変換により生成された全てまたは一部のＡＣ係数である。

【００２４】

図４に戻ると、各コピーのランクマトリックスがデータベース１４６に保持される。データベース１４６は、ランクマトリックス１～ｎと、照会画素のランクマトリックス１４５との間に一致があるかどうか調査できるように構成されている。ランクマトリックスマッチングシステムは、照会イメージランクマトリックス１４５を受け取り、データベース１４６を調査して、その照会イメージランクマトリックスと一致するかどうかを調べる。一例として、より効率的に照会イメージランクマトリックスとの一致を調べるために、データベース１４６のランクマトリックスは、クラスタに分けられている。一致が見つかれば、つまり、照会イメージのコピーが見つかれば、一致を示すデータがモニタ１４８に出力される。例えば、照会イメージのコピーとして識別された一つ以上のランクマトリックスは、汎用リソースロケータ（ＵＲＬ： ｕｎｉｆｏｒｍ ｒｅｓｏｕｒｃｅ ｌｏｃａｔｏｒ）アドレスを示すポインタを有していてもいい。従って、対応する出力ランクマトリックスをクリックすることによって、コピーされたイメージにアクセスすることができる。

【００２５】

図５は、本発明の実施例に基づいて、通信中の２つのサブシステムとして構成されたコピー検出システムのブロック図である。データベース照会システム１５６は、ユーザインタフェース１５８、ランクマトリックスジェネレータ１４４、特徴マッチングサブシステム１５４、データベース１４６を含む。データベース生成サブシステム１５２は、イメージ収集サブシステム１４２、ランクマトリックスジェネレータ１４４、クラスタリング／メタデータ生成１５６を含む。例えば、データベース生成サブシステム１５２への入力はイメージ収集サブシステム１４２によって集められたイメージであり、イメージ収集サブシステム１４２は、手作業又はウェブクローラなどのソフトウェアを用いて操作することができる。収集されたイメージのランクマトリックスは、図８～図１０で詳明するように、先ずイメージから抽出された後、しかるべきグループにまとめられる。そして、データベース照会サブシステム１５０によって使用されるインデックスを生成する。データベース照会サブシステム１５０は、特徴マッチングサブシステム１５４によって、インデックス付き要素の中に照会イメージのコピーがあるかどうかを判定するために、ランクマトリックスの抽出された照会イメージをデータベース１４６のインデックス付きの要素と突合せる。すなわち、特徴マッチングサブシステムは、データベースを調査して、クラスタ環境又はクラスタのない環境のいずれかにおいて照会イメージとの一致を調べる。インデック

ス付き要素と照会イメージの間に何らかの一致があれば、ユーザインタフェース１５８を介して表示される。言うまでもなく、特徴抽出機能及びクラスタリング／メタデータ生成機能を実行するコードを、半導体チップ上にハード的にコーディングすることもできる。当業者であれば、上記の機能を提供できるように構成されたロジックゲートを半導体チップに組み込むことできることが分かるであろう。例えば、フォームウェアと本願で説明する必要な機能を提供するためのロジックゲートファームウェアとを合成するためにハードウェア記述言語（ＨＤＬ）を用いることができる。

【００２６】

一実施例として、ランクマトリックスとして定義されたＡＣ係数の順序測度が、距離の測度として、そして対応するイメージを表わすシグネチャインデックスとして使用される。順序測度を使用したコピー検出方法をさらに詳しく説明するために、次の定義及び数式について以下に説明する。画像データベースをＴ＝｛Ｃ，Ｒ｝と定義する。ここで、Ｃ＝｛Ｑ，　C_1，…，C_m｝は照会（オリジナル）イメージＱ及びその変更された又は変更されていないコピーのセットを表わし、Ｒ＝｛R_1，…，R_n｝はデータベースの中の残りのイメージを表わす。検出器（分類器）は、２つのクラス（例えば、セットＣの「クラスＣ」、セットＲの「クラスＲ」）間の識別能力が高く、同時に、表１および図７Ａ、７Ｂで説明する通り、多様なイメージ変更に対し確実性を有していなければならない。T_iがデータベースからのテストイメージであると仮定すると、コピー検出は仮説テスト項を用いて次のように表現される。

$$H_0 : T_i \in C; H_1 : H_0^c$$

【００２７】

ここでＨ０は帰無仮説、Ｈ１は交代仮説である。例えば、帰無仮説はテストイメージが照会イメージと相関していることを記述し、それ以外を交代仮説が記述する。

【００２８】

本発明の別の実施例では、誤った検出を最小限に抑えながら照会イメージのコピーをできるだけ多く検出するために、イメージを表わす特徴としてＤＣＴ係数の順序測度を用いる。より具体的に言えば、８×８のサブイメージのＡＣ係数の大きさ（ｍａｇｎｉｔｕｄｅ）は降順又は昇順にランク付けされる。なお、ランクが付けられたＡＣ係数の大きさは単純なに整数の順列で、ランクマトリックスと称される。具体的に言えば、S_Nが整数［１，２，…，Ｎ］の全順列の集合を表わすとしたら、任意のランクマトリックスがこの集合の要素である。イメージI_i及びI_jから導出された２つのランクマトリックスr_iとr_j間の相関を測定するには、距離ｄ（r_i，　r_j）を定義しなければならない。当業者であれば、２つのイメージの距離は、そのランクマトリックスr_iとr_j間のミンコフスキー計量（Ｍｉｎｋｏｗｓｋｉ　ｍｅｔｒｉｃ）のL_1ノルムによって表わされることが分かるだろう。

【００２９】

$$d(r_i,r_j)=\left\|\mathrm{r_i}-\mathrm{r_j}\right\|_1=\sum_{l=1}^{N}\left|r_{il}-r_{jl}\right|,\ \forall(r_i,r_j)\in S_N,$$

ここで、Ｎはランクマトリックスのサイズである。L_1ノルム計量は、$\sqrt{\sum_{l=1}^{N}(r_{il}-r_{jl})^2}$に等しい$L_2$ノルムと比べた場合、線外値に対し確実性が高いことが分かっている。さらに、L_1ノルムは計算的に効率性が高い。一例として、任意のフォーマットのイメージをＹＵＶフォーマットに変換する。ただし、色はイメージ検索システムでは決定的な特徴となるが、コピー検出においては重要な役割を果たさないので、Ｙ成分だけが使用される。

【００３０】

図６は、本発明の一つの実施例における、ランクマトリックスジェネレータのモジュールの概略図である。ランクマトリックスジェネレータ１４４にイメージが入力されると、イメージは先ずイメージ分割モジュール１６２に入る。ここで、入力されたイメージは、例えば、同じサイズの６４（８×８）のサブイメージ（又はブロック）に分割される。別の例では、各サブイメージによって定義された、画素に対応する平均濃度が導出される。次に、離散コサイン変換（ＤＣＴ）モジュール１６４で、８×８の２次元（２Ｄ）ＤＣＴを行なうことにより、先に導出された平均濃度が、それぞれ一連の係数に変換される。次に、モジュール１６６で、２ＤのＤＣＴマトリックスは、例えば、１×６３のランクマトリックスなど、降順又は昇順で６３のＡＣの大きさのランクを含んだＡＣ係数の順序測度に変換される。説明のため、オリジナル（照会）イメージＱのランクマトリックスをｑ＝［q_1，q_2，・・・，q_N］とし、テスト画像Ｔのランクマトリックスをｔ＝［t_1，t_2，・・・，t_N］としている。ここではＮ＝６３。すると、照会イメージとテストイメージとの間の順序測度Ｄ（Ｑ，Ｔ）は、２つのランクマトリックス間のL_1ノルムとなる。すなわち、

【００３１】

$$D(Q,T)=d(q,t)=\sum_{l=1}^{N}\left|q_l-t_l\right|.$$

【００３２】

ミスマッチが発生するまでにシステムが許容できるデータ矛盾性の量を決めるのがコピー検出器の確実（ロバスト）性である。様々な変更に対するそれぞれの方法の確実性を検証するため、上に説明した測定方法を他の３つの測定方法と比較する。なお、ＤＣＴ係数から導出されたランクマトリックスを採用する上記の測定方法は、測定方法４として以下に定義する。他の３つの測定方法は測定方法１～３が、全てグリッドに基づく方法である。そこでは、イメージが６４（８×８）の同じサイズのサブイメージ（ブロック）に分割され、その平均濃度が導出される。各測定方法について以下に簡単に説明する。

【００３３】

測定方法１：基本的な相関ベースの方法。ここでは、２つの８×８サブイメージ間のL_1ノルムが算出される。コピーがあるかどうか判定するために８×８イメージの画素値間の差の和が用いられる。

【００３４】

測定方法２：８×８サブイメージから生成された２つのランクマトリックス間のL_1ノルムを算出するサブイメージの順序測度。ここでは、コピーがあるかどうかを判定するのに画素値のランクマトリックスが使用される。

【００３５】
測定方法３：ＤＣＴ係数の２つのセット間のＬ$_1$ノルムが算出される。最初に２ＤのＤＣＴを実行することにより８×８サブイメージのＡＣ係数を算出する。次に、照会イメージに基づく対応するＡＣ係数の大きさとテストイメージに基づく対応するＡＣ係数の大きさとの間のＬ$_1$ノルムが算出される。
【００３６】
測定方法４：本明細書に記載の実施例において説明した、ＡＣの大きさの順序測度を用いてＬ$_1$ノルムが算出される。
【００３７】
上に紹介した測定方法をテストするために、同じイメージに基づいた１３の変更イメージを４００００イメージを有するデータベースに入れた。すなわち、オリジナルイメージと１３のイメージ変更版とを４００００イメージのデータベースの中に挿入した。表１Ａは、上記の測定方法１～４として挙げた４つのコピー検出テストに基づく上位１４位までのランクを示している。大文字Ａ～Ｎは表１Ｂに定義したイメージタイプを示す。なお、表１Ｂで説明したイメージタイプはあくまで模式であり、他のタイプのイメージ変更を検出するのに本願で説明している実施例を使用できること限定するものではない。表１Ａに説明を戻すと、順序表現の距離測定方法（測定方法２及び４）は、非順序表現の距離測定方法（測定方法１及び３）よりも確実性が高かった。しかしながら、測定方法２の結果が、画素値の順序測度は、予想される回転や反転（表１ＢのイメージタイプＧ及びＨ）などの変更を検出できないことを示しているのに対し、ＤＣＴ係数のＡＣの大きさの値の順序測度（測定方法４）は、１３全てのコピーとオリジナルを検出できている。表１Ａの「ＦＤ」は検出の誤りを示す。すなわち、コピーでないイメージが対応する測定方法によってランク付けされた。

表１Ａ

RANK	Measure 1	Measure 2	Measure 3	Measure 4
1	A	A	A	A
2	D	D	D	D
3	F	F	F	F
4	N	B	N	N
5	K	J	J	J
6	L	N	H	B
7	I	M	G	H
8	FD	E	E	G
9	E	K	I	M
10	FD	L	L	E
11	FD	I	K	I
12	FD	C	FD	L
13	FD	FD	FD	C
14	FD	FD	FD	K

【００３８】

表１Ｂ

文字指定	イメージタイプ／変更
A	オリジナル
B	ヒストグラム平坦化
C	水彩絵
D	モーションブラ
E	モザイクタイル
F	ガウスノイズ
G	180° 回転
H	水平方向反転
I	色相の変化
J	黒ずみ
K	ねじれ
L	テキスト挿入
M	コントラスト改善
N	リサイジング

【００３９】

上に示した実験結果から、提案した測定方法（測定方法４）は、多種多様な歪みに対し確実性を発揮することが分かる。しかしながら、この確実性は、相関しないイメージを比較する場合、欠点に変わってしまう可能性がある。これは、順序測度に起因するとともに、ＤＣＴ係数を使用することによるものである。従って、システムの識別能力、つまり、識別力も重要になってくる。それは、検出誤りを起こさないように、無関係なイメージを棄却する能力と関係するためである。

【００４０】

本願で説明するコピー検出スキームに係る実施例では、識別能力に影響を及ぼす一因となるのがランクマトリックスのサイズである。ランクマトリックスのサイズによって順序測度に使用されるＡＣ係数の量が決まる。画素値に突然何か変化が起こると、その結果として高周波が変化する可能性があるため、高周波におけるＡＣ係数の作用が重要となる。そして、最適なランクマトリックスサイズにより、識別能力量を測定することができる。テスト手法および結果を以下に説明する。

【００４１】

データベースの中の照会イメージとテストイメージ間の距離を$x = d(q,t)$と仮定する。正規化された距離$\hat{x}$は次のように定義することができる。

$$\hat{x} = x/M, \quad 0 \le \hat{x} \le 1,$$

ここで、Mは$d(r_i,r_j)$, $\forall (r_i,r_j) \in S_N$の最大値である。$M$は２つの順列が互いの逆数のときに得られる。$N$はランクマトリックスのサイズである。また、$w_1$が先行確率$P(w_1)$を有するクラスＣを表わし、$w_2$がクラスＲを表わすとすると、

$$P(w_2) = 1 - P(w_1)$$

システムの自然状態（クラスラベル）とシステムの決定が分かっていると仮定する。従って、そうした情報により識別能力がもっとも高いランクマトリックスサイズを判定できるようになる。２つの可能性を考えると、

- $P\left(\hat{x} > \hat{x}^* \middle| \hat{x} \in w_1\right)$:

$$\text{誤った棄却} = \frac{\text{正規化された距離が}\hat{x}^*\text{よりも大きいコピー数}}{\text{コピー総数}}$$

- $P\left(\hat{x} > \hat{x}^* \middle| \hat{x} \in w_2\right)$:

$$\text{正しい棄却} = \frac{\text{正規化された距離が}\hat{x}^*\text{よりも大きい非コピー数}}{\text{非コピー総数}}$$

【００４２】
ここで、$\hat{x}^*$は、データベース中のテストイメージが、クラスＣに属するかどうかを判定するための正規化されたしきい値である。なお、誤った棄却（false rejection)とは、コピーではないとして棄却されたイメージのコピーを意味する。一方、正しい棄却とは、イメージの非コピーが正しく棄却された場合を意味する。トライアル回数が多ければ、確率を実験的に判定することができる。

【００４３】

図７Ａは、本発明の実施例における、４つの異なるランクマトリックスサイズと、その４つのランクマトリックスサイズに対応した最大識別能力を有するランクマトリックスサイズを判定するために、誤った棄却及び正しい棄却の棄却率をグラフにしたものである。$\hat{x}^*$が０から１まで変化した場合の４つのランクマトリックスサイズ（Ｎ）の棄却率を図７Ａに２次元のグラフで示した。図７Ａのグラフは、レシーバ動作特性（ＲＯＣ）曲線と呼んでもいい。テストしたランクマトリックスサイズは次の通りである。

【００４４】

●Ｎ＝６３：ここでは、８×８のＤＣＴのＡＣ係数の全てが順序測度に使用されている。結果を図７Ａのライン１６７ｄで示す。

●Ｎ＝４８：４８の低周波数ＡＣ係数だけをとり、順序測度に使用する（つまり、ＤＣ係数を除く７×７の左上の係数）。結果を図７Ａの１６７ｂで示す。

●Ｎ＝３５：３５の低周波数だけがとられ、順序測度に使用される（つまり、ＤＣ係数を除く６×６の左上の係数）。結果を図７Ａの１６７ａで示す。

●Ｎ＝２４：２４の低周波数だけがとられ、順序測度に使用される（つまり、ＤＣ係数を除く５×５の左上の係数）。結果を図７Ａの１６７ｃで示す。

【００４５】

ランクマトリックスに関して、説明してきた測定方法の性能を識別するために、スフィライジング（ｓｐｈｅｒｉｚｉｎｇ）、ラジアルブラーリング（ｒａｄｉａｌ　ｂｌｕｒｒｉｎｇ）、ブラシストローク（ｂｒｕｓｈ　ｓｔｒｏｋｅｓ）などもっと多様な変更を追加することによって、セットＣのサイズを１４から３０に増やした。なお、これらのタイプのイメージ変更は、表１Ｂに挙げた変更タイプとは別のものである。最大識別能力を示す曲線は座標（０，１）を通る。ここで、０は誤った棄却に対応する相対的測度で、１は正しい棄却に対応する相対的測度である。図７Ａにライン１６７ａで示すように、ランクマトリックスのサイズＮが３５のとき識別能力が最も高い。これは、高周波おけるＡＣ係数の適切な棄却により、ＡＣ係数の順序測度が向上することを示すものである。なお、ランクマトリックスサイズを６３から３５に小さくすると、インデックス化のためのメモリの節約になる。さらに、高周波のＡＣ係数は、一般に６×６の左上の係数の外にあるため、高周波のＡＣ係数のノイズは全て除去される。

【００４６】

図７Ｂは、本発明の一実施例における、順序測度に用いられる６×６の左上のＡＣ係数を描いた８×８離散コサイン変換の概略図である。図７Ａで説明したように、ＤＣＴ係数１７３を差し引いた６×６のブロック１７２は、ＡＣ係数の最適数と判定された。さらに、領域１７４における高周波ＡＣ係数はノイズを含む可能性が高いため、この実施例では削除されている。なお、最適なランクマトリックスサイズは、特定のアプリケーションによって変わる可能性がある。

【００４７】

イメージデータベースが大きくなると、データベースを効率的に調査するためには、クラスタベースの検出が必要になる。クラスタリングする目的は、類似したデータ項目を一緒にカテゴリー化又はグループ化することによって、データ量を減らすことにある。このようなグループ化は、人間が情報を処理する際の一般的な処理の仕方である。クラスタリングアルゴリズムを使用させるために重要なことの一つは、カテゴリー又は分類（ｔａｘｏｎｏｍｙ）を構築するのに有効な自動化ツールを提供することである。なお、これらの方法は、プロセスにおいて人的要因の影響を最小限に抑えるための方法としても使用できる。

【００４８】

よく使用されるクラスタリングの方法は、その実施し易すさから、ｋ平均値クラスタリングで、Ｔａｂｕサーチ、Ｇｅｎｅｔｉｃアルゴリズム、Ｓｉｍｕｌａｔｅｄアニーリング等の類似の方法と比較した場合、実行時間の観点から見た効率が一番よい。そのため、ｋ平均値クラスタリングは大型のデータセットに適している。よく知られているように、ｋ平均値クラスタリングは、ランダム初期分割から始まり、最も近いクラスタに対してパターンの再割り当てを行ない、その後、収束基準を満たすまで重心を再計算し続ける。

【００４９】

しかし、クラスタ数の選択が分析に大きく影響する。すなわち、ｋが変わるとまったく異なる種類のクラスタが出現する。ｋ平均値アルゴリズムでは、クラスタの数、つまり、ｋがはっきりと指定されていなければならない。最適なクラスタ数は前もってはっきりとは分かっていないので、最適クラスタ数を判定するために、クラスタの妥当性分析を行う。例えば、最適クラスタ数は、クラスタ間の距離を最大にしながら、同時にクラスタ内の距離を最小にするクラスタの数である。つまり、クラスタが異なればそのランクマトリックスは大きく変わるが、一つのクラスタ内でのランクマトリックスはコンパクトであるべきである。この特徴を数式で表すと次のようになる。

$$v(k)=\frac{1}{k}\sum_{i=1}^{k}\max_{1\le j\le k,i\neq j}\left\{\frac{\eta_i+\eta_j}{\varepsilon_{ij}}\right\}$$

ここで、

$$\eta_j=\frac{1}{n_j}\sum_{i=1}^{n_j}d\left(r_i^{(j)},\mu_j\right)\ \text{and}\ \ \varepsilon_{ij}=d(\mu_i,\mu_j).$$

η_{ij}はクラスタｊのクラスタ内距離で、ε_{ij}はクラスタｉとｊのクラスタ間距離である。$r_i^{(j)}$はｊ番目のクラスタに属するｉ番目の特徴ベクトルで、n_jはｊ番目のクラスタにおける特徴ベクトルの数である。従って、上記方程式は、各クラスタのコンパクトさと、クラスタの離れ具合を測るものである。最適なクラスタ数$\hat{k}$は$\hat{k}=\min_{2\le k\le 10}v(k)$として選択される。言い換えれば、$k$平均値アルゴリズムは$k=\{2,\cdots,10\}$がテストされ、$v(k)$の最小値を与えるものが選ばれる。

【００５０】

図５に関して、先に説明したコピー検出システムを、性能の観点から評価した。性能の評価では、イメージデータの５つのコピーセットをデータベースに挿入した。ここで、各セットはオリジナルイメージの１０の変更コピーを有する。従って、イメージに対応する計５５（１１×５）のランクマトリックスがデータベースに挿入された。オリジナルの変更コピーには、表１Ｂに関して説明した変更のほかに、次のようなタイプの変更：ひび割れ（ｃｒａｑｕｅｌｕｒｅ）、ウォーターペーパー（ｗａｔｅｒ　ｐａｐｅｒ）、モザイク、リップル、オーシャンリップル、スポンジ、アスペクト比変更、ダークストローク（ｄａｒｋ　ｓｔｒｏｋｅｓ）、彩度増などがある。ここでもまた、本実施例では、ここに挙げた変更や表１Ｂに挙げた変更を有するコピーに限らず、オリジナルが変更されているどんなコピーでも検出するように構成されている。言うまでもなく、色ベースのスキームでは、カラーやコントラストが変更されているイメージは場合によっては検出できないこと

があると予想されるが、その一方、相関ベースのスキームではイメージが回転していたり反転していたりすると検出できないことが予想される。また、相関ベースのスキームは、対応する画素の非線形の濃度のばらつきに対して弱いと予想される。また、テクスチャベースのアプローチは、一般にひび割れ、オーシャンリップル、モザイクタイルと呼ばれる変更など、エッジロケーションが変更されたり、新たにエッジポイントが追加されたコピーに対処するのは容易ではない。

【００５１】

以上、説明してきたコピー検出システム並びにアルゴリズムの性能を評価するために、適合率（ｐｒｅｃｉｓｉｏｎ　ｒａｔｅ）及び呼出率（ｒｅｃａｌｌ　ｒａｔｅ）を次のように定義する：

$$\text{適合}(\hat{x}^*)=\frac{\hat{x}<\hat{x}^*\text{を有する検出コピー数}}{\hat{x}<\hat{x}^*\text{を有する検出数}}$$

$$\text{呼出}(\hat{x}^*)=\frac{\hat{x}<\hat{x}^*\text{を有する検出コピー数}}{\text{コピー総数}}$$

ここで、$\hat{x}^*$は正規化されたしきい値である。各イメージをデータベースから１１のコピー（１０の変更版と１のオリジナル）を検出するために照会として提起した。

【００５２】

図８は、本発明の一実施例における、適合率及び呼出率対正規化しきい値（正規化しきい値に対する適合率及び呼出率）のグラフである。グラフから分かるように、適合率及び呼出率がともに比較的に高いので、望ましい結果は正規化しきい値０．３あたりで得られる。なお、検出し損なったコピー数を最低限に抑えるには、高い呼出率であることが重要である。データベースからのテストイメージが、コピーのクラスに属するかどうかはしきい値によって決まる。例えば、しきい値の決定は以下に説明するアルゴリズムによって計算される。

【００５３】

例えば、コピー検出器、つまり分類器は、データベースからのあるテストイメージがクラスＣ、つまり、コピーのクラスに属するかどうかを判定するために、あるしきい値を用いる。コピー検出システムの最適なしきい値θ_xは次のような数式で表わすことができる：

$$C=\bigcup_i T_i \quad \text{ここで} \quad x=d(q,t_i)<\theta_x$$

【００５４】

ここで、Ｃはクラスに属するイメージのセットで、T_iはイメージデータベースからのテストイメージである。照会（オリジナル）イメージとテストイメージとの間の距離$d(q,t_i)$は、x値がクラスＣに属するイメージの平均（mean）μ_1と、データベースにある残りのイメージの平均μ_2とを有する確率変数ｘとしてモデル化することができる。分布が正規、すなわち、$p(x|w_i) \sim N(\mu_i, \sigma_i^2)$で、ここで、$w_i$は可能性のあるクラスを示し、クラスＣでは$i=1$で、クラスＲでは$i=2$であるとすれば、最適しきい値を見出すための合理的なスタート点は次のような尤度比テストを用いることである。

$P(w_1|x) > P(w_2|x)$ならばw_1を決定し、それ以外はw_2を決定する

【００５５】
ここで、$P(w_1|x)$及び$P(w_2|x)$は、H_0及びH_1にそれぞれ対応する事後条件付き確率を表わす。ベイズの定理が式の両辺に適用され、各項が下記ように整理される、

$$\frac{p(x|w_1)}{p(x|w_2)} \underset{H_1}{\overset{H_0}{\gtrless}} \frac{P(w_2)}{P(w_1)}, \qquad (1)$$

【００５６】
左辺の比は尤度比として知られており、方程式全体は尤度比テストと呼ばれることが多い。尤度比テストは、最大事後確率を用いてそのイメージクラスを選択するもので、その決定基準は最大事後基準（MAP）と呼ばれている。なお、平均的に、MAP基準は間違った決定を最小数とするので、この基準は最小誤差基準とも呼ぶことができる。Kが事前の$P(w_2)/P(w_1)$の比を示すとすれば、方程式（１）を次のように書ける：

$$\frac{p(x|w_1)}{p(x|w_2)} \underset{H_1}{\overset{H_0}{\gtrless}} K. \qquad (2)$$

【００５７】

クラス条件付き確率密度関数をガウス分布によりモデル化すると、次の式を得る：

$$p(x|w_1)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right], \tag{3}$$

$$p(x|w_2)=\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]. \tag{4}$$

【００５８】

分布は実際のデータをほぼモデル化している。方程式（３）及び（４）を（２）に代入することにより、次の式を得る。

$$\exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]\underset{H_1}{\overset{H_0}{\gtrless}}K\frac{\sigma_1}{\sigma_2}. \tag{5}$$

【００５９】

さらに、両辺の対数をとると、

$$-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\underset{H_1}{\overset{H_0}{\gtrless}}\ln\left(K\frac{\sigma_1}{\sigma_2}\right) \tag{6}$$

【００６０】

両クラスとも同じ分散（variance）を有すると仮定すると、歪みが多すぎるコピーは、オリジナルイメージのＩＤを失っているとみなすようにクラスＣを統計的に束縛することができる。$\sigma=\sigma_1=\sigma_2$だから、方程式（６）は次のように変形される。

$$2x(\mu_1-\mu_2)+\mu_2^2-\mu_1^2\underset{H_1}{\overset{H_0}{\gtrless}}2\sigma^2\ln(K) \tag{7}$$

【００６１】

最終的に、方程式（７）を下記のように変形することによりxの最適しきい値θ_xを算出する解は、

$$x\underset{H_1}{\overset{H_0}{\lessgtr}}\frac{\mu_1+\mu_2}{2}-\frac{\sigma^2\ln(K)}{\mu_2-\mu_1}=\theta_x \tag{8}$$

【００６２】

方程式（８）を分析する際、Ｋ＝１のとき、又は両クラスの事前確率が同じ場合は、最適しきい値は２つの平均値（mean value）の平均値、つまり、$(\mu_1+\mu_2)/2$になる。Kが大きくなるにつれて、$\frac{\sigma^2\ln(K)}{\mu_2-\mu_1}>0$だから最適しきい値は小さくなる。

実際には、Ｋはずっと大きな値である可能性が大きい。方程式（８）を用いて最適しきい値を算出するために、表１で先に説明した同じテスト材料を使用した。４００００のイメージを有するデータベースに、３０のコピーと一つのオリジナルを入れた。μ_1、μ_2、σ_2^2は、正規化された値としてそれぞれ 0.1464、0.5842、0.0043 という値が算出された。この実施例において、方程式（８）を用いて判定された最適しきい値は$K=40{,}000/30$のとき 0.30 である。なお、しきい値は$K=1$のとき 0.37 である。ここで、最適しきい値は、Kが大きくなるにつれて両クラスとも２つの平均値の単なる平均よりも低く設定される。なお、データベースからのテストイメージは、Kが大きくなればなるほどコピーのクラスに属する確率が低くなる。

【００６３】

上記性能テストは、クラスタのないコピー検出として実施したが、ｋ平均値クラスタリングに対して、先に説明したクラスタベースの検出をこの性能テストに統合してもいい。上述の通り、データベースの中の特徴ベクトル、つまり、ランクマトリックスは、ｋ平均値アルゴリズムによりクラスタ化されており、最適なクラスタ数を判定するために、クラスタの妥当性分析を行なうことができる。

【００６４】

図９は、本発明の一実施例における、最適クラスタ数を判定するためのクラスタ妥当性分析の結果を示すグラフである。この図から、クラスタ妥当性測度ｖ（ｋ）は、ｋ＝７のとき最低であることが分かる。そこで、システムのクラスタ数を例えば７に設定する。ここでは、先ずその７つのクラスタの重心が、照会イメージの特徴ベクトルと比較される。例えば、ランクマトリックスのような、その重心がＮ個の最も近接した重心に属するようなクラスタの特徴ベクトルだけが選ばれ、コピー検出をより効率的なものにする。図９のコピー検出システムの場合、最適クラスタ数は図１０から４と判定された。従って、７つのクラスタのうちの４つが、照会イメージのランクマトリックスとアライメントしていると識別され、コピーが捜される。もちろん、当業者であれば、クラスタ妥当性分析によって示されるクラスタ数が、アプリケーションによって７以下になったり７以上なったりすることは明らかである。また、調査対象として識別されるクラスタ数は、クラスタ総数の一部分であることもあるし、クラスタ全てを含むこともある。上に述べた通り、コピー検出は、クラスタのない環境で行なうこともできる。

【００６５】

図１０は、本発明の一実施例において、クラスタベースの検出を用いて最適しきい値での適合率と呼出率を比較したグラフである。クラスタのない場合と、上に述べた７つのクラスタのうちの１つ、２つ、３つ、４つを考察した場合の、最適しきい値での適合率、呼出率を表わす。図１０に示すように、コピー検出に、照会イメージに最も近いクラスタだけを用いた時に適合率は最も高く、呼出率は最も低かった。使用する近接クラスタ数が増えると、つまり、Ｎが大きくなると、適合率はわずかに低下するが、呼出率は増加する。

【００６６】

最も近接した使用クラスタ数が４の時の、クラスタベースのコピー検出結果の詳細な値を表２にまとめた。この場合、検出結果はクラスタなしの検出結果と同じであったが、図１０に示した通り、検出のための処理時間はクラスタなしの検出のときと比べて６５％短かった。照会イメージの特徴ベクトル抽出及びデータベースに入っている４００００のイメージとの距離マッチングを含めたクラスタなし検出の処理時間は、９３３ＭＨｚのペンティアムIIIで約１．８秒であった。表２に示したように、４００００のイメージのデータベースに挿入された５５のコピーの中から、最適しきい値で５３のコピーが検出されたが、イメージＣの照会では、２つのコピーだけを検出できなかった。検出できなかったコピーはダークストロークで変更されていたコピーと水彩絵具で変更されていたコピーであった。イメージＢの照会では、１２の間違った検出があったが、照会イメージのコピーは、全て検出された上位１１のイメージの中に入っていた。従って、一例として、照会イメージの特性いかんにより適応型しきい値化スキームが役に立つかもしれない。なお、図８で説明したように、正規化された最適しきい値$\hat{\theta}_x$を 0.3と指定した。さらに、図１０の適合値及び呼出値はクラスタなし検出と４クラスタ検出とでは同じである。言うまでもなく、先に述べた通り、４クラスタ検出の時間はクラスタなしの検出と比べ著しく短い。

【００６７】

表２

照会イメージ	X＝コピー数$<\hat{\theta}_x$	Y＝検出数$<\hat{\theta}_x$	精度($\hat{\theta}_x$)＝X/Y	呼出($\hat{\theta}_x$)＝X/11	上位１１のコピー数	$\hat{\theta}_x$で検出できなかったコピー数
A	11	12	0.92	1	11	0
B	11	23	0.48	1	11	0
C	9	9	1	0.82	10	2
D	11	11	1	1	11	0
E	11	15	0.73	1	10	0
平均			0.83	0,96	10.6	0.4

【００６８】

表２のデータは、先に説明したコピー検出方法の効率の良さを示している。照会イメージＡ，Ｂ，Ｄ、Ｅに関して、全て１１のイメージ（１０の変更されたコピーと１つのオリジナル）が検出された。照会イメージＣでは、１１のうちの９つのイメージが検出された。

さらに、１１以上のイメージが検出された場合は、この１１のイメージは見つかった上位１１のイメージの中に常時入っていた。なお、イメージはランクマトリックスとしてデータベースに保持され、先に説明したように、照会イメージのランクマトリックスに基づいてデータベースの探索が行なわれた。言うまでもなく、データベース内のデータはクラスタになっていてもクラスタなしでも可能である。

【００６９】

図１１は、本発明の一実施例における、ディジタルイメージのコピーを検出する方法のフローチャートである。このフローチャートは、イメージが選択されるオペレーション２００から始まる。ここで選択されたイメージは、インターネットなど分散型ネットワーク上にあるイメージで、例えば、ウェブクローラなどのサーチソフトウェアの支援によって見つけられる。次に、オペレーション２０２に進み、イメージからシグネチャインデックスが定義される。特徴ベクトルとも呼ばれるこのシグネチャインデックスは、先に説明したように、１次元のランクマトリックスを含んでいる。例えば、選択されたイメージを同じサイズの６４のサブイメージ又はブロックに細分することによって、シグネチャインデックス、つまり、ランクマトリックスが導出される。この６４の同じサイズのサブイメージは、８×８のグリッドに配列される。サブイメージと関連した画素値から算出された平均濃度の平均がとられる。この平均濃度は２次元ＤＣＴオペレーションにより、一連の係数に変換される。このＤＣＴに基づく一連の係数は、次にランクマトリックスに変換される。図７Ａ及び７Ｂで述べたように、ランクマトリックスは６３のＡＣ係数を全て或いは部分的に含んでいてよい。

【００７０】

図１１の方法では、次にオペレーション２０４に進み、定義されたシグネチャインデックスを保持する。シグネチャインデックスは、一例として、データベースに保持される。図８～１０で述べた通り、データベースに入っているイメージのコピーの検出をより効率的に行うため、シグネチャインデックスはクラスタになっている。前述の通り、ｋ平均値クラスタリングなどのクラスタリング方法によってデータがデータベースでクラスタ化される。次にオペレーション２０６に進み、照会イメージのシグネチャインデックスが判定される。なお、照会イメージとは、コピー検出中のオリジナルイメージのことである。例えば、この照会イメージのシグネチャインデックスの判定は、オペレーション２０２で説明したイメージデータのシグネチャインデックスを定義する方法で行われる。次にオペレーション２０８に進み、シグネチャインデックスが保持されていれば、それと照会イメージのシグネチャインデックスとの一致が識別される。一例として、ランクマトリックス間の一致を識別するために、ランクマトリックスマッチングシステムが用いられる。別の例では、クラスタ妥当性分析により、最適シグネチャインデックス数を判定することによって、一致を検出するタイミングが最適化される。

【００７１】

要約すれば、上に説明した発明は、ディジタルイメージの無許可のコピーを検出することができる。一実施例では、コピー検出器はイメージデータベースを有し、イメージデータベースには数百万のウェブページから集めたイメージが入っている。もしくは、システムがイメージのシグネチャインデックス、つまり、特徴ベクトルやＵＲＬを維持するようにすることもできる。なお、保持されているイメージは、その特徴ベクトルが抽出されたら削除されるから、メモリスペースを節約するだけでなく著作権問題を起こさないためにもシグネチャのデータベースが望ましい。ウェブを周期的にクロールする自動イメージ探索ツールによって、新たに発見されたイメージのプロファイルが作成された後、そうしたプロファイルは特徴に基づいてインデックス化される。

【００７２】

【発明の効果】

本明細書で説明した特定の例では、高周波ノイズを避けるために８×８のサブイメージの３５ＤＣＴ係数の順序測度を採用しているが、任意の数のＤＣＴ係数を使用して構わない。結果は、色の変更だけでなく反転や回転を含み、多様な変更に対して確実（ロバスト）

である。さらに、本書で説明した実施例は誤った検出の発生を最小限度に抑える識別力が高い。ランクマトリックスをインデックス化するには、ランクマトリックス毎に６ビット×３５／イメージ、つまり、約２７バイト／イメージが必要である。その場合、ランクマトリックスのサイズは３５となる。当業者であれば、特定のアプリケーションによっては、数百万のイメージを保有しているデータベースに有効に対処するためにより効率的なクラスタ化及びインデックス化を考慮してもいい。なお、ＤＣＴ係数の最適数やデータベースでのシグネチャインデックスの最適クラスタ数など、本書で提供した具体的な実施例は、具体的な例との関係において示したものである。従って、最適数は、シグネチャインデックスのタイプ、データベースのサイズ、望ましい適合及び呼出、最適しきい値など、によって変わってくる。しかしながら、最適レベルに到達する上述の方法は、コンテントベースのイメージコピー検出のどんなアプリケーションにも応用することができる。

【００７３】

上記に鑑み、本発明は、コンピュータシステムに保持されたデータを用いて様々なコンピュータで実施するオペレーションを採用してもいいことが分かる。こうしたオペレーションには、物理的数量を物理的に操作しなければならないオペレーションも含まれる。必ずしもそうとは限らないが、こうした数量は、保持、転送、結合、比較、及びその他の操作が可能な電気信号又は磁気信号の形をとるのが一般的である。さらに、実行される操作は、生成、識別、判定、或いは比較といった表現で呼ばれることが多い。

【００７４】

上に説明した発明は、ハンドヘルドデバイス、マイクロプロセッサシステム、マイクロプロセッサベースの或いはプログラム可能な消費者用エレクトロニクス、ミニコンピュータ、メインフレームコンピュータなどを含み、他のコンピュータシステム構成で実施してもいい。本発明はまた、通信ネットワークを介してリンクされた遠隔処理装置によってタスクが実行される分散コンピューティング環境で実施することもできる。

【００７５】

本発明はまた、コンピュータ可読媒体にコンピュータ可読コードとして実装することもできる。コンピュータ可読媒体とは、データを保持し、その後そのデータをコンピュータシステムで判読できる任意のデータ記憶装置である。コンピュータ可読媒体の例としては、ハードディスクドライブ、ネットワーク接続型ストレージ（ＮＡＳ）、リードオンリーメモリ、ランダムアクセスメモリ、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－ＲＷ、磁気テープや、その他の光データ記憶装置及びオプティカルでないデータ記憶装置などが挙げられる。コンピュータ可読媒体は、コンピュータ可読コードを分散して保持、実行できるようにネットワークに接続されたコンピュータシステムで配布することもできる。

【００７６】

上記の発明を理解しやすいように幾分詳細に説明してきたが、特許請求の範囲から逸脱しない限りにおいて何らかの変更、修正が可能なことは明白である。従って、本明細書で明らかにした実施例は説明のためのもので限定を意図したものでないと考えるべきであり、本発明は本明細書で述べた詳細に限定されるものではなく、特許請求の範囲内において修正変更することができる。特許請求において、各要素及び／又はステップは、請求項にはっきりと記載されている場合を除き、オペレーションの特定の順番を暗に示すものではない。

【図面の簡単な説明】

【図１】イメージコピー検出器のサーチ間の相違を示す３つのイメージ及びコンテントベースのイメージ検索システム。

【図２】コピー検出のための順序測度の使用を表現した図。

【図３】図１のイメージ１０２を空間的に変更したイメージ。

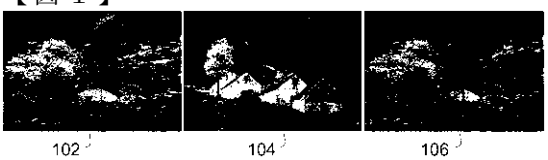
【図４】コンテントベースのコピー検出システムの概念図。

【図５】通信している２つのサブシステムとして構成されたコピー検出システムのブロック図。

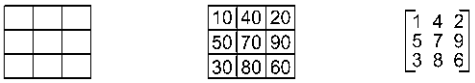
【図６】ランクマトリックスジェネレータのモジュールの概略図。

【図７】Ａ４つの異なるランクマトリックスサイズと、それに対応する、高い識別能力を有するランクマトリックスサイズを判定するための正しい棄却率及び誤った棄却率のグラフ。
Ｂ順序測度として使用される（６×６）－１の左上ＡＣ係数を描いた８×８離散コサイン変換の概略図。
【図８】適合率及び呼出率対正規化しきい値のグラフ。
【図９】最適クラスタ数を判定するためのクラスタ妥当性分析の結果
【図１０】クラスタベースの検出を用いた最適しきい値での適合率と呼出率を比較したグラフ。
【図１１】ディジタルイメージのコピーを検出する方法のオペレーションのフローチャート。
【符号の説明】
１４０　インターネット
１４２　ウェブクローラ
１４４　ランクマトリックスジェネレータ
１４５　照会イメージランクマトリックス
１４６　データベース
１４８　モニタ
１５０　データベース照会サブシステム
１５２　データベース生成サブシステム
１７２　６×６ブロック
１７３　ＤＣＴ係数
１７４　領域

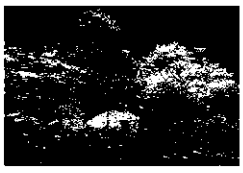
【図１】



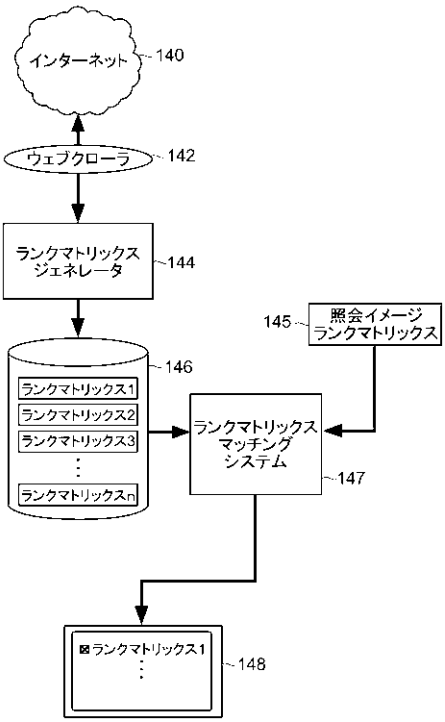
【図２】



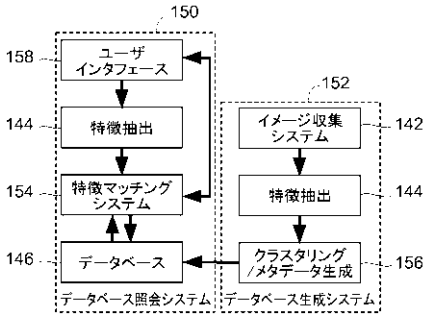
【図３】



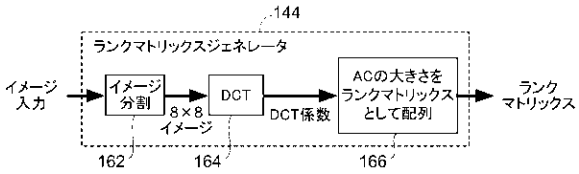
【図４】



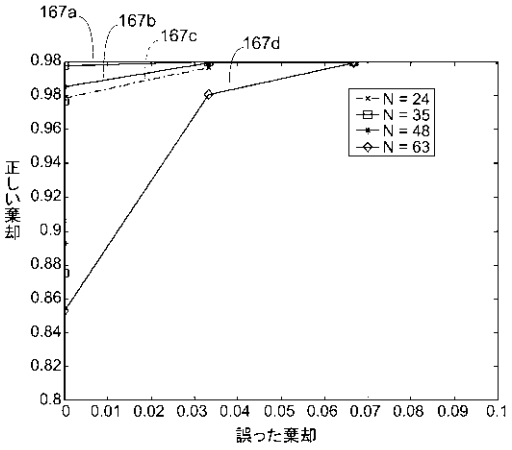
【図５】

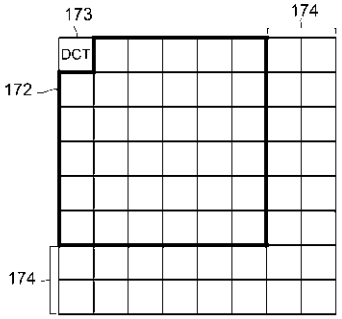


【図６】

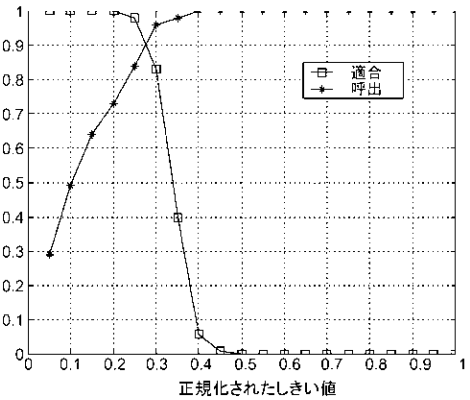


【図７】

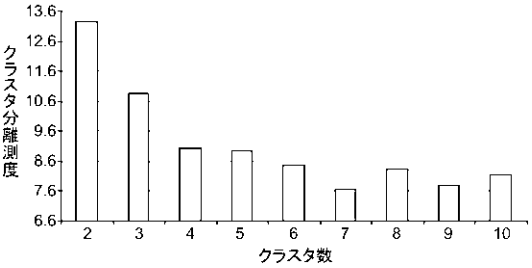




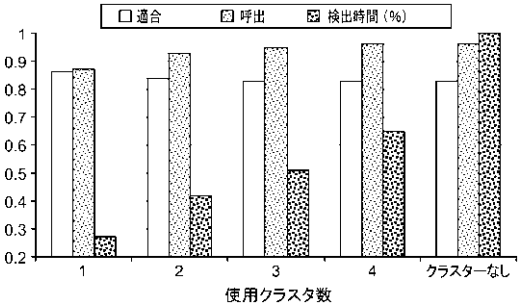
【図８】



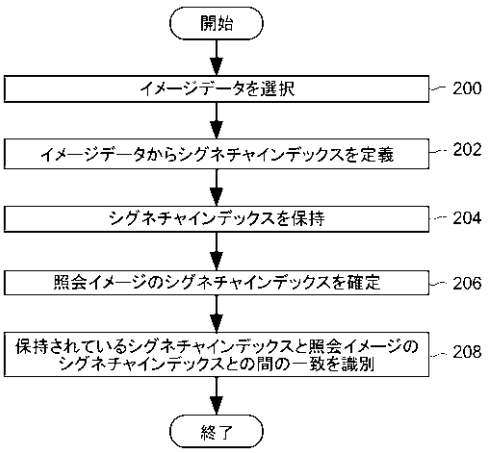
【図９】



【図１０】



【図１１】



フロントページの続き

Ｆターム(参考) 5B057 AA11 BA02 CA08 CA12 CA16 CB08 CB12 CB16 CG02 CG05
DA12 DB02 DB09 DC06 DC33
5C077 LL14 MP01 PP46 PP49 PP68 PQ12 PQ22 TT06