(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5654970号
(P5654970)
(45)発行日　平成27年1月14日(2015.1.14)　(24)登録日　平成26年11月28日(2014.11.28)

(51)Int.Cl.　　F I
G06T 7/60 (2006.01)　　G06T 7/60 200Z
G06T 7/00 (2006.01)　　G06T 7/00 250
G06K 9/00 (2006.01)　　G06K 9/00 P

請求項の数 5　(全 29 頁)

(21)出願番号　特願2011-202404 (P2011-202404)
(22)出願日　平成23年9月15日 (2011.9.15)
(65)公開番号　特開2012-64216 (P2012-64216A)
(43)公開日　平成24年3月29日 (2012.3.29)
審査請求日　平成26年9月3日 (2014.9.3)
(31)優先権主張番号　12/883,503
(32)優先日　平成22年9月16日 (2010.9.16)
(33)優先権主張国　米国 (US)

早期審査対象出願

(73)特許権者　502096543
パロ・アルト・リサーチ・センター・インコーポレーテッド
Palo Alto Research Center Incorporated
アメリカ合衆国、カリフォルニア州　９４３０４、パロ・アルト、コヨーテ・ヒル・ロード　３３３３
(74)代理人　100079049
弁理士　中島　淳
(74)代理人　100084995
弁理士　加藤　和詳

最終頁に続く

(54)【発明の名称】画像クラスタリング、分類、および反復構造発見のためのグラフラティス法

(57)【特許請求の範囲】
【請求項１】
原始要素および関係を含む複数のデータグラフをクラスタ化する方法であって、
関連するサブグラフのグラフを含むグラフラティスを生成することを含み、次数１のサブグラフは前記原始要素であり、次数ｉ＞１の各サブグラフは次数ｉ－１のサブグラフと前記原始要素のうちの１つとを含み、前記グラフラティスは、ノードのラティスであり、各ノードは、画像原始要素および関係を表すサブグラフに対応し、かつ各グラフラティスノードは、前記グラフラティスノードが対応する前記サブグラフの記述的情報を提供するように構成され、
前記複数のデータグラフに対する特徴ベクトルを生成するために前記グラフラティスを使用することを含み、
前記生成された特徴ベクトルの間の類似性に従って前記複数のデータグラフをクラスタ化することを含み、
前記方法は、少なくとも１つのデジタルプロセッサを用いて実行され、
ストラットは、レベルＮの親グラフラティスノード、原始要素、及び、前記親グラフラティスノードの境界にリンクされた前記原始要素に加えて前記親グラフラティスノードからなる前記サブグラフであるレベルＮ+１の子グラフラティスノードからなる
複数のデータグラフをクラスタ化する方法。
【請求項２】
前記特徴ベクトルが接合点規格化マッピングカウントを用いて決定される、請求項１に

記載の方法。
【請求項３】
接合点規格化マッピングカウントが、接合点マッピングカウントにより再重み付けされたサブグラフ・マッチング・カウントである、請求項１に記載の方法。
【請求項４】
前記生成された特徴ベクトルの間の前記類似性が共通マイナス差尺度を用いて決定される、請求項１に記載の方法。
【請求項５】
共通マイナス差尺度が絶対値を要素ごとに比較する、請求項１に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、文書認識システムおよび方法に関する。特に、画像が原始的特徴の集まりとして表され、これらの原始的特徴の空間的関係がグラフとして表される文書認識システムおよび方法に関する。
【背景技術】
【０００２】
最近、画像分類へのアプローチが急増している。かかるアプローチでは、オブジェクトやシーンは比較的単純な特徴抽出の大きいベクトルとしてモデル化される。課題は特徴により取得される情報である。従来の特徴は、純粋に外観をベースとした特徴である。しかし最近は、主要ポイントまたは関心のあるポイントでサンプル化された特徴抽出から、情報を空間関係にある情報として抽出しようとする傾向にある。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
空間関係をコード化する一つの方法はグラフである。オブジェクトやシーンは、部分（ノード）や関係（リンク）としてモデル化される。観察された画像は、観察された部分およびかかる部分と他の部分との関係についてのグラフを生成し、認識はサブグラフマッチングによって行われる。サブグラフマッチングにはいくつかの困難が伴う。第一に指数関数的に高価である。この問題は属性グラフの使用によりある程度緩和される。しかし、サブグラフマッチングは、第二の困難な点により比較的小さいサブグラフに限定されている。この第二の困難な点というのは、ノイズと変動性によって、観察されたグラフが理想のモデルから逸脱してしまうことである。その結果、不正確なグラフマッチング技法を使用することになり、ひいては、マッチングコストが急増し、属性グラフマッチングの利点は大半失われる。つまり、画像のノイズおよび変動性は、必要なマッチングを迅速かつ効果的に行うのを困難にしている。かかる問題に対処する方法および／またはシステムの開発が求められている。以下の開示は、かかる方法および／またはシステムについて考察するものである。
【課題を解決するための手段】
【０００４】
本開示のプロセスは、原始要素および関係を含む複数のデータグラフをクラスタ化する方法であって、制御演算装置により、関連するサブグラフのグラフを含むグラフラティスを生成することを含み、次数１のサブグラフは前記原始要素であり、次数ｉ＞１の各サブグラフは次数ｉ－１のサブグラフおよび前記原始要素のうちの１つを含み、制御演算装置により、前記複数のデータグラフに対する特徴ベクトルを生成するために前記グラフラティスを使用することを含み、制御演算装置により、前記生成された特徴ベクトルの間の類似性に従って前記複数のデータグラフをクラスタ化することを含む、複数のデータグラフをクラスタ化する方法を含む。
【０００５】
また、本開示のプロセスは、前記クラスタ化することが、前記特徴ベクトルを用いて前

記複数のデータグラフのそれぞれに対する最良適合クラスタを見つけることと、前記複数のデータグラフの前記それぞれに対する類似性スコアを前記データグラフの前記最良適合クラスタに基づいて決定することと、前記データグラフの前記類似性スコアが第１のしきい値を超えているとき、前記複数のデータグラフの前記それぞれを前記データグラフの前記最良適合クラスタを用いて分類することと、前記データグラフの前記類似性スコアが第２のしきい値よりも小さいとき、前記複数のデータグラフの前記それぞれを新しいクラスタ内に分類することと、前記データグラフの前記類似性スコアが前記第１のしきい値と前記第２のしきい値の間にあるとき、前記複数のデータグラフの前記それぞれを未確定クラスタ内に分類することと、を含む方法を含む。

【０００６】

また、本開示のプロセスは、原始要素および関係を含むデータグラフを分類する方法であって、制御演算装置により、関連するサブグラフのグラフを含むグラフラティスを生成することを含み、次数１のサブグラフは前記原始要素であり、次数ｉ＞１の各サブグラフは次数ｉ－１のサブグラフおよび前記原始要素のうちの１つを含み、制御演算装置により、前記データグラフと第１のカテゴリの見本とに対する特徴ベクトルを生成するために前記グラフラティスを使用することを含み、制御演算装置により、前記データグラフと前記第１のカテゴリの前記見本とに対する前記生成された特徴ベクトルを比較することを含み、　画像と前記見本との前記特徴ベクトルの間の類似性がしきい値を超えているとき、制御演算装置により、前記データグラフを前記第１のカテゴリの要素として分類することを含む、データグラフを分類する方法を含む。

【図面の簡単な説明】

【０００７】

【図１】直線的線画、およびこれらの実施例の中に見られる反復構造の一部を示す図である。

【図２】直線的線画の領域では１３種類の接合点があることを示す図である。

【図３】原始要素の９８個の可能な組み合わせがあることを示す図である。

【図４】グラフラティスを規定する親子関係を示す図である。

【図５】グラフラティスを作るのに使用される可能性がある２つの画像を示す図である。

【図６】ストラットの概念を示す図である。

【図７】実施例データグラフとグラフ・ラティス・ノードのサブグラフの間のマッピング図である。

【図８】グラフラティスを生成するためのアルゴリズムを示す図である。

【図９】拡張の概念を示す図である。

【図１０】マッピングを計算するためのアルゴリズムを示す図である。

【図１１】図１０のアルゴリズムの基礎となる中心概念を示す図である。

【図１２】純粋な（またはきちんとした）マッピングカウントをその要素にもつ特徴ベクトルは、線画クラスタリングおよび分類に対してなぜうまく機能しないのか、その理由を示す図である。

【図１３】サブグラフサイズが最大４までの特徴ベクトルを用いて２００のＮＩＳＴ文書の、対をなす類似性ヒストグラムを示す図である。

【図１４】１１，１８５のＮＩＳＴ納税申告書のクラスタリング結果を示す図である。

【図１５】周期的反復構造および孤立反復構造を示す図である。

【図１６】グラフ・ラティス・システムを示す図である。

【図１７】グラフ・ラティス・システムを用いた文書認識に適用されるコンピュータ・ビジョン・システムを示す図である。

【発明を実施するための形態】

【０００８】

サブグラフの形の複雑な画像特徴の大きな族を、グラフラティスの構造（すなわち、ラティスで結ばれた関連するサブグラフの階層）を通じて、より単純な画像特徴から作ることができる。多数のこれらの特徴サブグラフを支持することにより、正確なグラフマッチ

ングを通じて画像構造を捕捉できる。後で分かるように、グラフラティスは画像雑音およびばらつきがある場合でも効率的なグラフマッチングを促進し、効率的な画像クラスタリング、分類、検索、反復構造発見、および新規性検出を促進することが有利である。下記のシステムおよび方法は、特に文書形状認識の実際的問題に対処するために直線的線画の領域で説明されている。しかしながら、概念は直線的線画以外の、原始要素の集合に分解できる画像に適用できることを理解すべきである。

【０００９】

「グラフラティス」と呼ばれる基本的枠組みはノードのラティスであり、各ノードは画像原始要素および関係を示すサブグラフに対応している。グラフ・ラティス・ノードは、それらの各サブグラフから原始要素を付加（上方に）および除去（下方に）することによりラティス内で互いに関連している。例えば、原始要素が次数１を定義するとき、ラティスが次数１から次数Ｎまで広がっていると仮定すれば、次数ｉ＞１のすべてのサブグラフは次数ｉ－１のサブグラフに原始要素を加えたものを含む。

【００１０】

アイデアは少なくとも２つの理由から直線的線画の領域で説明されている。第１に、直線的線画の領域の基準線は明確な形で交差して接合点および自由端ターミネータを形成するので、直線的線画はグラフとして容易に表すことができる。これらの接合点はグラフのノードとして使用するのによく適しているとともに、接合点を結び付ける基準線部分はグラフの連結棒として使用するのによく適している。第２に、直線的線画は文書において一般的であり、上述のように、本明細書に開示する発明の要旨は、文書といっしょに一般に使用される画像分類、検索、重複検出に関連して特定用途を有している。

【００１１】

図１は直線的線画、およびこれらの実施例の中に見られる反復構造の一部を示す。第１の画像１０２および第２の画像１０４は、それぞれ棒グラフの直線的線画画像を含んでいる。さらに、これらの画像の中には部分構造１０６のような反復部分構造がある。部分構造１０６が示すように、部分構造は単一画像の中で、および／または複数画像にわたって繰り返すことができる。グラフラティス表現は部分構造をサブグラフとして発見して使用することを可能にする。

【００１２】

図２は直線的線画の領域における１３種類の接合点を規定している。これらはグラフラティスの原始要素または第１レベルのサブグラフである。１３個の原始的接合点種別は、原子を組み合わせて分子を形成するのとほぼ同じ方法でそれらの適合する連結方向に従って分類できる。図３はこれらの原始要素の９８個の可能な組み合わせがある、言い換えれば、次数２の９８個のサブグラフがあることを示している。これらの組み合わせのうちの２つは、中ぶらりんの線分を有していないサブグラフである図形（例えば、単一の水平なおよび垂直な線分）を形成している。しかしながら、残りは中ぶらりんの線分を有しており、そのために、それらは図形のサブグラフとしてのみ使用できる。

【００１３】

原始要素とサブグラフの間の親子関係は、ラティスを形成する。より小さいサブグラフは親ノードと呼ばれ、接合点を付加することにより、より小さいサブグラフから生成されたより大きなサブグラフは子ノードと呼ばれる。図４は上述した概念を示している。そのとき、原始要素の例外はあるが、次数ｉの各サブグラフは次数ｉ－１のサブグラフおよび原始要素を含む。例えば、次数３のサブグラフ４０２は、次数２のサブグラフ４０４および原始要素４０６（次数１のサブグラフでもある）を含む。本明細書の議論では子グラフ・ラティス・ノードの次数は常にそれらの親の次数よりも１だけ大きいことを仮定しているが、特定の実施形態では、任意の大きさの２つのサブグラフを結合して、より大きなサブグラフを生成してもよい。

【００１４】

次数Ｎの完全に母集団化されたグラフラティスを作るために、次数のサブグラフのすべての連結適合位置に１３個のすべての原始要素を付加して次数ｉ＋１のすべてのサブグラ

フを定義し、ここで、ｉは１からＮ－１まで広がっている。ラティスの各レベルは次の層のための基礎の役割を果たす。さらに、次数の全グラフラティスを作ることは、各図形の接合点およびすべてのサブグラフを含むすべての可能な図形の空間に対する抽象的表現を提供する。しかしながら、予想通りに、グラフラティスを完全に母集団化すると、次数は３を超えて桁はずれに大きくなる。この問題は下記の段落で検討されるであろう。

【００１５】

ここでは、Ｎ個の接合点を含む単一の図形だけを表すグラフラティス、およびそのサブグラフのすべてについて検討する。この図形は次数Ｎのグラフラティス内の単一ノードを定義することになる。その結果、次数Ｎ－１では、図形はＮ－１ノードを有し、このＮ－１ノードのそれぞれは、その接合点のうちの１つがなくなった状態のサブグラフということになる。次数Ｎ－２のノード数は、図形のトポロジーに依存するであろう。したがって、グラフラティスは平らな基礎を有するひし形を形成し、ここで、基礎は原始要素を表す１３のノードを含む。ひし形は（Ｎ／２）の周囲の層で通常最も広くなり、ここで、存在する接合点およびなくなった接合点のほとんどの組み合わせが生じる傾向があることになる。単一の図形に対するグラフラティス内のノード総数は約2^{N}個である。

【００１６】

グラフラティスの概念は直線的線画との関連で導入されたが、他の定式化も同様に受け入れられることを理解されたい。すなわち、グラフラティスの概念は画像特徴が原始要素の集合に分解できる他の領域にも適用できる。

【００１７】

グラフラティスの生成と関連するいくつかの基本概念を導入した後に、グラフラティスを作るためのアルゴリズムを導入する。アルゴリズムは理論的グラフラティス全体の一部分だけを作ることによりグラフラティスの複雑さを管理し、このアルゴリズムは所与のデータコーパスおよび用途の集合に対して意味があり有用である。

【００１８】

データコーパスはデータグラフの集合であり、各データグラフは画像に対応している。画像は、例えば、文書画像であり得る。上述のように、データグラフは原始要素を用いて画像を表し、データグラフ内のノードは原始要素に対応し、データグラフ内の縁端部は接合点の間の連結棒に対応している。直線的線画の場合、画像のデータグラフは、基準線を抽出して、それにより形成される接合点を決定することにより構成される。その後、これらの接合点はデータグラフのノードを定義し、これらのノードは、基準線により相互接続される。

【００１９】

図５はグラフラティスを作るのに使用される概念を説明する例を示している。図５の２つの画像５０２、５０４のデータグラフに対してグラフラティスを作ることが好ましいことと、各画像のデータグラフがＮ個の接合点を含むこととを仮定する。上述のように、Ｎ＞３のとき、完全に母集団化されたグラフラティスを作るのは一般に非現実的である。この問題に対処するために、グラフラティスは画像５０２、５０４の両方のデータグラフ内に見つけられるサブグラフ５０６、５０８のようなサブグラフとともに母集団化されるだけである。

【００２０】

より低いレベル（次数）のグラフ・ラティス・ノードは、より高いレベルのグラフ・ラティス・ノードのサブグラフであってもよい。複雑さを限定するために、グラフ・ラティス・ノード３個の間の親子関係だけが保持される。これらの３個はレベルＮ（親）のノードと、原始要素（技術的には第２の親）と、親グラフ・ラティス・ノードのサブグラフに、その周辺（子）に結び付けられた原始要素を加えたもので構成されるサブグラフであるレベルＮ＋１のノードと、で構成される。この三方向の関係が、ストラットと呼ばれるデータ構造内で保持される。

【００２１】

ストラットの目的は２つである。第１に、ストラットは親グラフ・ラティス・ノードと

子グラフ・ラティス・ノードの間の接合点インデックスマッピングを保持する。一般に、任意のグラフ・ラティス・ノードは、その構成要素接合点に順序不同にインデックスを付けるであろう。ストラットは、それらを親グラフ・ラティス・ノードと子グラフ・ラティス・ノードの間で系統的に整理された状態に保持する。第２に、ストラットはプリミティブ型、親の配置、および親から子を作る接合点に対する連結棒を示す。

【００２２】

図６はストラットの概念を示している。ストラットは、その構成グラフ・ラティス・ノードへのポインタを保持し、これらのノードは、それらを接続するすべてのストラットへの連結棒を保持する。ストラットは、Ｓ｛Ａ，ｐ，ｉ，Ｂ，Ｍ，Ｌ｝と表され、ここで、Ａは親グラフ・ラティス・ノードであり、ｐは子を生成するために親に付加された原始要素の種類であり、ｉはこの原始要素の子サブグラフ内でのインデックスであり、Ｂは子サブグラフ（グラフラティスノード）であり、Ｍは親接合点インデックスと子接合点インデックスの間のマッピングであり、Ｌは子サブグラフを生成するための原始要素の親との連関である。Ｌは、付加された原始要素に関する方向インデックスから、子のノードインデックスへマッピングする。

【００２３】

グラフ・ラティス・ノードは、それらが関与するストラットのリストを保持する。しかしながら、特定の実施形態では、原始要素は節約のためにこれらのリストを保持していない。このような実施形態では、原始要素はストラットのリストだけを保持しており、ここで、両方の親は原始要素であり、子は次数２のグラフ・ラティス・ノードである。

【００２４】

各グラフ・ラティス・ノードはそれ自体がサブグラフである。したがって、ノードは、グラフ・ラティス・ノードのサブグラフ接合点を、対応するデータグラフ接合点にマッピングすることにより、観察されるデータグラフと照合できる。原始要素は分類されるため、この照合は属性サブグラフマッチングのための任意の公知アルゴリズムを用いて実行できる。図７は実施例データグラフとグラフ・ラティス・ノードのサブグラフの間の、結果として生じるマッピングを示している。一般に、このようなマッピングは一対多数になるであろう（すなわち、グラフ・ラティス・ノードで表される単一のサブグラフは、観察されるデータグラフの複数の部分にマッピングしてもよい）。

【００２５】

マッピングは本明細書でマッピングセット（ＭａｐｐｉｎｇＳｅｔ）と呼ぶデータ構造内に保持される。マッピングセットは｛グラフ・ラティス・ノード、データグラフＩＤ（ＤａｔａＧｒａｐｈＩＤ）、マッピングリスト（ｌｉｓｔ－ｏｆ－Ｍａｐｐｉｎｇｓ）｝の３要素で構成されている。データグラフＩＤは、データグラフおよびデータグラフの関連する原画像（例えば、ファイル名など）へのポインタである。マッピングリストはマッピングのリストであり、このリストのそれぞれは｛順マッピング配列、逆マッピング配列｝の対である。順マッピング配列はグラフ・ラティス・ノードの次数に等しい大きさを有する配列である。この配列はグラフ・ラティス・ノードのサブグラフのノードインデックスから、データグラフ内のノードのインデックス上にマッピングする。逆マッピング配列は、データグラフのノードインデックスから、グラフ・ラティス・ノードにおいて表されるサブグラフのノードインデックスへマッピングする配列またはハッシュ表である。データグラフは非常に大きくなると予想されるため、データグラフの大きさに等しい長さを有する配列の代わりに、ハッシュ表（衝突検出を備えた）として逆マッピングを保存することの方が、より空間効率が良い。

【００２６】

各グラフ・ラティス・ノードはマッピングセットと呼ばれるリストを保持する。これらのマッピングセットは、グラフ・ラティス・ノードのサブグラフがマッピングされているデータグラフ上に身元および位置を記録する。したがって、各グラフ・ラティス・ノードは、そのグラフ・ラティス・ノードのサブグラフがマッピングされている各データグラフに対するマッピングセットを含んでいる。

【００２７】
常に、承認済みグラフ・ラティス・ノードのリスト、および候補グラフ・ラティス・ノードのリストが保持されている。これらのリストの目的についてはさらに詳細に後述するが、簡潔に述べると、承認済みグラフ・ラティス・ノードはグラフラティスに付加されたノードであり、候補グラフ・ラティス・ノードはグラフ・ラティス・ノードへの付加を現在検討しているノードである。開始状態として、原始要素のリストが承認済みグラフ・ラティス・ノードに使用され、空集合が最初の候補グラフ・ラティス・ノードに使用される。

【００２８】
さらに、候補および承認済みグラフ・ラティス・ノードの集合は、次数（原始要素の個数）によりインデックスを付けられた配列で構成されるデータ構造内にそれぞれ保持されている。この配列の各要素は、原始要素の個数の組み合わせによりインデックスを付けられたハッシュ表で構成されている。例えば、ハッシュインデックスは、各原始要素とデータグラフとが一致した回数を数える数字から連結された文字列に関してＪａｖａハッシング関数を用いて計算してもよい。このデータ構造の目的は、重複したグラフ・ラティス・ノードを効率的に検出できるようにすることである。

【００２９】
特定の実施形態では、それぞれの承認済みおよび候補グラフ・ラティス・ノードは、それが作られた原始要素の個数のカウントを保持して、次数によるインデックス付けをより効率的に促進するようにしている。グラフ・ラティス・ノードの原始要素のカウントは、グラフ・ラティス・ノードの次数またはレベルに対応している。

【００３０】
グラフラティスを生成するためのアルゴリズムを図８に示している。アルゴリズムは入力としてデータ見本の集合を取り込み、これらのデータ見本のそれぞれは原始要素と原始要素の間の関係を示す連結棒とを示すノードで構成されるデータグラフである。これらの見本により、データ見本にマッピングできるサブグラフを単に生成するだけで、グラフラティスを手元で使用する形にすることが辛うじて可能になる。アルゴリズムは、候補グラフ・ラティス・ノードを生成すること（動作８０２）、候補グラフ・ラティス・ノードを選択すること（動作８０４）、選択されたグラフ・ラティス・ノードを昇格させること（動作８０６）、および終了条件が満たされるまで繰り返すこと（動作８０８）を含んでいる。

【００３１】
候補グラフ・ラティス・ノードは、承認済みグラフ・ラティス・ノードおよび観察されるデータグラフからラティス生成される（アクション８０２）。概要の方法では、承認済みグラフ・ラティス・ノードの、観察されるデータグラフ上へのマッピングは調査され、新しい候補グラフ・ラティス・ノードを生み出すために使用される。観察されるデータグラフは、以前に見られたデータグラフ、および／または新しく、新規で、以前には観察されていないデータグラフを含んでいてもよい。上述のように、承認済みグラフ・ラティス・ノードは、最初は原始要素のリストを含む。

【００３２】
候補グラフ・ラティス・ノードを生成する第１のステップは、次数Ｎの承認済みグラフ・ラティス・ノードの拡張を生成することである。レベルＮのグラフ・ラティス・ノードの、観察されるデータグラフ上へのあらゆるマッピングは、新しいレベルＮ＋１のグラフ・ラティス・ノードを生み出すための種としての役割を果たすことができ、この新しいレベルＮ＋１のグラフ・ラティス・ノードは、そのレベルＮのグラフ・ラティス・ノードで表されるサブグラフのスーパーグラフである。サブグラフの周辺に結び付けられた各原始要素は、それ自体がサブグラフの大きさをノード１つだけ大きくし、したがって、グラフラティスにおいて次数（レベル）１だけ高くすることができ、これ以降はグラフ・ラティス・ノードの拡張と呼ばれる。

【００３３】

図９を参照すると、この概念を示している。そのとき、種グラフ・ラティス・ノード９０２が使用され４種類の拡張９０４ａ～９０４ｄを生み出される。この実施例では、各拡張の次数は６であり、種グラフ・ラティス・ノードの次数よりも１だけ高い。さらに、各拡張は観察されるデータグラフ９０６の中に見られる。

【００３４】

あらゆるレベルＮの承認済みグラフ・ラティス・ノードにより生成された各拡張は、新しいレベルＮ＋１の候補グラフ・ラティス・ノードとして追加される前に、レベルＮ＋１の既存のグラフラティスと比較されて、各拡張が既存の承認済みまたは候補グラフ・ラティス・ノードと重複していないことが確認される。この重複検査は、上述したグラフ・ラティス・ノードのハッシュ表インデックス付けにより促進される。実際のグラフマッチングにより比較されなければならない、同じ可能性のあるレベルＮ＋１のグラフ・ラティス・ノードの集合は、ハッシュ表を通じて、ほんの少数の候補グラフ・ラティス・ノードに絞られる。

【００３５】

拡張が重複していないことが分かったとき、拡張はレベルＮ＋１の候補グラフ・ラティス・ノードのリストおよびハッシュ表に追加される。その後、それぞれの新しい候補グラフ・ラティス・ノードは、また、それがマッピングするデータグラフに、ストラットを通じて結び付けられる。この新しい候補グラフ・ラティス・ノードを、そのレベルの親および関連する原始要素と結び付けるストラットは明らかである。しかしながら、グラフラティスのラティス特徴のために、他のレベルのグラフ・ラティス・ノードもまた、新しい拡張のサブグラフであってもよい。これらの関係に対するストラットもまた、形成されなければならない。

【００３６】

候補グラフ・ラティス・ノードが生成されると（動作８０２）、承認済みグラフ・ラティス・ノードに昇格させるために候補グラフ・ラティス・ノードの一部を選択する（動作８０４）。一般に、目標は、付加されたノードがクラスタリング、分類、反復構造検出、またはグラフラティスの他の応用の目的を果たすようにグラフラティスを大きく育てることである。

【００３７】

候補グラフ・ラティス・ノードを選択するための１つの方法は、最大ノード種別多様性基準である。この方法のランクは、すべての候補グラフ・ラティス・ノードを、ノードｎに対する原始的ノード種別ｉのエントロピーH_nとして測定されるプリミティブ型の多様性に従って順序付ける。

【００３８】

【数１】

$$H = \sum_i -p_i \log p_i \qquad (1)$$

【００３９】

$$p_i = \frac{c_i}{\sum_i c_i} \qquad (2)$$

【００４０】

ここで、c_iはグラフ・ラティス・ノードｎで使用される種類ｉの原始要素の個数のカウントである。ノード種別多様性基準は多くの異なる種類の接合点を含むノードを含むグラフラティスを大きく育てることに通じ、これらのグラフラティスはクラスタリングおよび分類の観点から見てデータグラフの中で多くの場合最も特徴的なサブグラフである。また

、他の選択基準も可能である。

【００４１】

次に、以前に選択された候補グラフ・ラティス・ノード（動作８０４）を承認済みグラフ・ラティス・ノードに昇格させる（動作８０６）。承認されたステータスを獲得することにより、グラフ・ラティス・ノードは新しい候補グラフ・ラティス・ノードの種としての役割を果たす資格がある。

【００４２】

候補グラフ・ラティス・ノードを昇格させた（動作８０６）後、アルゴリズムは終了条件が満たされるまで繰り返される（動作８０８）。次数Ｎの新しく昇格したグラフ・ラティス・ノードは観察されるデータグラフに対するマッピングを既に参照しているため、繰り返しはわずかである。その後、これらのマッピングは昇格したグラフ・ラティス・ノードの子を捜すために容易に追跡調査され、これらの子は、まだレベルＮ＋１のグラフ・ラティス・ノードで表されていないデータサンプル内で観察されるサブグラフを表している。

【００４３】

可能な終了条件は下記の５項目を含むが、これらに限らない。

【００４４】

（ｉ．）　所与のレベルにおける承認済みグラフ・ラティス・ノードのしきい値個数を含むグラフラティス。

【００４５】

（ｉｉ．）　すべての承認済みグラフ・ラティス・ノードのしきい値個数を含むグラフラティス。

【００４６】

（ｉｉｉ．）　使い尽くされている候補グラフ・ラティス・ノードのリスト。

【００４７】

（ｉｖ．）　候補グラフ・ラティス・ノードに対する品質測定がしきい値を下回る。

【００４８】

（ｖ．）　決定された時間を超える実行時間。

特定の実施形態では、終了条件は動作８０４で検討される採用戦略に依存している。

【００４９】

データグラフを作るための上述のアルゴリズムにかかわらず、他のアルゴリズムも同様に受け入れられることを理解されたい。例えば、１つのアイデアは、ノード種別多様性のエントロピーに基づく尺度を用いて、強く示唆される候補グラフ・ラティス・ノードを選択することである。

【００５０】

グラフラティスを用いて実行する演算は、１つ以上の画像から導出される観察されたデータグラフに対するマッピングを計算することである。グラフラティスは非常に大きくなる可能性がある（数千または数十万のノードを含んでいる）ため、この計算を効率的に行うことは重要である。単純素朴な方法は、観察されるデータグラフと、各グラフ・ラティス・ノードのサブグラフの間で別々にサブグラフマッチングを実行することである。しかしながら、グラフ・ラティス・ノードの間の関係におけるラティス構造を利用する、より効率的なアルゴリズムについて後述する。

【００５１】

図１０は、入力としてデータグラフを取り込み、グラフ・ラティス・ノードで表されるサブグラフから、データグラフ上へのすべてのマッピングを記述するマッピングセットの編集物を出力するマッピングを計算するためのアルゴリズムを示している。アルゴリズムは、原始要素とデータグラフのノードとの一致を計算すること（動作１００２）と、次数２のサブグラフとデータグラフとの一致を計算すること（動作１００４）と、次数３およびより高次のサブグラフの一致を繰り返して計算すること（動作１００６）とを含んでいる。

【００５２】
まず、次数１のサブグラフはデータグラフと照合される（動作１００２）。サブグラフは、グラフ・ラティス・ノードのサブグラフ接合点を、対応するデータグラフ接合点にマッピングすることにより、観察されるデータグラフと照合され得る。
【００５３】
次数１のサブグラフは照合された（動作１００２）後に、次数２のサブグラフがデータグラフにマッピングされる（動作１００４）。原始要素ＡおよびＢの各対に対して、それらがレベル２のグラフ・ラティス・ノードc_1，c_2，．．．，c_Nに対する１つ以上のストラットの親であるかどうかを判断する。それらがc_1，c_2，．．．，c_Nに対する１つ以上のストラットの親であると判断された場合、アルゴリズムはＡの、データグラフ上へのすべてのマッピングを繰り返し適用して、c_iが有効なマッピングであるかどうかを判断する。c_iが有効なマッピングであるかどうかの判断は、そのマッピングに対するＡの適正な近さにおける原始要素Ｂへの連結棒の存在についてデータグラフを試験することにより実行される。
【００５４】
次数２のサブグラフをデータグラフにマッピングした（動作１００４）が、レベル３およびより高レベルのマッピングは、レベル３のグラフ・ラティス・ノードから始めて繰り返し計算される。レベルＮの各グラフラティスノードＲ（Ｎ＝３から始まる）に対して、アルゴリズムは前のレベルの親ノードＡに対する１つのストラットＳを選択する。レベルＮのマッピングを見つけるためには、ノードのレベルＮ－１のサブグラフのすべてをマッピングすることになるため、１つのストラットを考慮することだけが必要である。その後、ストラットＳに関連するレベルＮ－１の親ノードＡに対して、アルゴリズムは、その親ノードＡの、データグラフ上へのマッピングのそれぞれを繰り返すとともに、それぞれのこのようなマッピングに対して、アルゴリズムは、ストラットＳにより示される原始要素ｐもまたデータグラフ上に存在しているかどうか、およびストラットＳにより示される原始要素ｐが、ストラットの連関パラメータＬで示されるように結び付けられているかどうかを調べる。この原始要素が存在しており、適切に結び付けられている場合には、Ｂへのマッピングが確立され得る。
【００５５】
この手順の複雑さはグラフラティス内の親子ストラットの個数の増加とともに直線的に増大し、グラフ・ラティス・ノードとデータグラフの間のマッピングの個数の増加とともに直線的に増大する。重要な点は、各レベルにおいて、すべてのマッピングが前のレベルで見つかるマッピングに対する漸進的拡張であるため、マッピングを計算するには少しの仕事だけが必要であるという点である。
【００５６】
図１１を参照すると、照合アルゴリズムの基礎となる概念を示している。すなわち、レベルＮ＋１のグラフ・ラティス・ノードＢの、データグラフへのマッピングは、そのグラフ・ラティス・ノードＢのレベルＮの親グラフ・ラティス・ノードの、データグラフへのマッピングから、ほとんどの場合引き継がれる。その後、ＡからＢへのストラットは、原始要素ｐ（Ｂ上のインデックス５）の存在についてデータグラフ上のどこで試験すればよいのかを示している。したがって、照合アルゴリズムは漸進的マッピングを経験する。
【００５７】
グラフラティスの枠組みは、文書認識におけるいくつかの重要な用途、およびコンピュータビジョンの他の態様を支持する。しかしながら、これらについて説明する前に、グラフラティスマッピングに基づく特徴ベクトル、および妥当な特徴ベクトル類似性尺度について説明する。
【００５８】
グラフラティス表現のいくつかの使用は、グラフラティスから、観察されるデータグラフへのマッピングに基づく特徴ベクトルを計算することを含んでいる。各グラフ・ラティス・ノードはベクトルの１つの要素を含んでおり、その要素に対するベクトル入力は、そ

のグラフ・ラティス・ノードのサブグラフの、データグラフ上へのマッピングの個数から導出される。

【００５９】

試験は、純粋な（またはきちんとした）マッピングカウントをその要素にもつ特徴ベクトルは、線画クラスタリングおよび分類に対してうまく機能しないことを示している。その理由は、より大きなサブグラフが多過ぎることと関係している。より大きなサブグラフ特徴に対しては、非常に重複の多いサブグラフを非常に多く照合する。データグラフ内の任意のノード（線画接合点）は、低次のサブグラフよりもはるかに多くの高次のサブグラフに関与するであろう。このことは、原始的接合点を検出する際のエラーがたとえ少数であったとしても、結果として多数のマッチカウントの不安定性を引き起こす。

【００６０】

図１２では、丸を付けた接合点１２０２が、図示のサブグラフ１２０４に、より多くの図示していないサブグラフを加えたものを用いてマッピングすることにより重ね合わせる方法で覆われている。上述のように、このような重ね合わせにより、接合点または領域がマッピングカウント特徴ベクトルにおいて不規則に表されるようになる。

【００６１】

これを解決するために、接合点規格化マッピングカウント（ＪＮＭＣ）に基づく特徴ベクトルを使用する。接合点規格化マッピングカウントはレベル当たりで計算される。言い換えれば、ある特定のレベルのノードに対するグラフラティスノードマッピングm_iのすべてを計算し、これらを使用して、そのレベルのすべてのノードに対するマッピングカウントが規格化される。各レベルのマッピングに対するマッピングカウントが互いに別々に規格化される。

【００６２】

レベルＬに対して、観察されるデータグラフ内の各接合点ｊに対して重み付けW_jを計算する。

【００６３】

【数２】

$$w_j = \frac{1}{N(j)} \qquad (3)$$

【００６４】

ここで、Ｎ（ｊ）は、接合点ｊを含むレベルＬのすべてのノードからのマッピングの個数である。その後、グラフ・ラティス・ノードｉに対する接合点規格化マッピングカウント要素Ｃｉは下記の式で表される。

【００６５】

【数３】

$$C_i = \sum_{m_i} \sum_{j \in m_i} w_j \qquad (4)$$

【００６６】

ここで、m_iはグラフ・ラティス・ノードｉによる、観察されるデータグラフ上へのマッピングの集合である。言い換えれば、所与のグラフ・ラティス・ノードに対応する接合点規格化カウントベクトル要素は、そのグラフ・ラティス・ノードによりマッピングされ、そのグラフ・ラティス・ノードの、観察されるデータグラフ上へのすべてのマッピングにわたって合計される、すべての接合点に対する接合点重みの合計とみなされる。

【００６７】

接合点規格化マッピングカウントは、グラフ・ラティス・ノードで表されるサブグラフのマッピングのカウントの特徴ベクトルの構築に向けて、観察されるデータグラフ内の各

接合点に均等な重みを与える働きをする。上述の式を通じて、これらの重みは各接合点を含むマッピングの間に分配される。接合点が１回または数回だけマッピングされる場合、接合点はカウントに対して強く寄与する。他方、接合点が多くの重複マッピングにより覆われている場合、これらのマッピングは、その接合点の寄与重みをすべて共有しなければならない。接合点規格化式は、いくつかのグラフ・ラティス・ノードがたまたま多くの重複マッピングを有するときに、それらが特徴ベクトルを支配するのを防止するが、このような状況は反復構造がある場合に起こる可能性がある。

【００６８】

クラスタリングおよび分類を行うためにデータに対する特徴ベクトル表現を比較することは標準的技法である。類似性／相違点スコアを出すために異なる式が使用されてもよい。当然の選択としてはユークリッド距離およびコサイン距離などがある。これらの選択のどちらも、有効に働くかどうか分かっていない。例えば、コサイン距離は、サブグラフの、観察されるデータグラフ上へのマッピングのカウントから導出された特徴ベクトルを比較する場合には有効に働かない。したがって、共通マイナス差（ＣＭＤ）と呼ばれる下記の類似性尺度が使用される。

【００６９】

【数４】

$$s(v_1, v_2) = \frac{\sum_i (\min(v_{1,i}, v_{2,i}) - |v_{1,i} - v_{2,i}|)}{\max(|G_1|, |G_2|) * N} \qquad (5)$$

【００７０】

ここで、G_kはデータグラフｋの大きさ（接合点の個数）であり、Ｎは接合点規格化特徴ベクトル内で考慮されるサブグラフサイズの個数である。

【００７１】

よく知られているコサイン距離は、ベクトル要素の分布または相対値を比較するように設計されており、他方、ＣＭＤ距離はまた、絶対的な大きさを要素ごとに比較する。コサイン距離は特徴要素の、ともに正のカウントを有する任意の対にクレジットを与え、他方、ＣＭＤの挙動は、より正確である。カウントが似ている限り、正のクレジットが与えられ、カウントが異なる限り、負のクレジットが与えられる。現在比較しているデータグラフの大きさに基づく規格化条件により、ＣＭＤ類似性尺度の範囲は－２（最小、最低の類似性）～１（最大、最高の類似性）である。

【００７２】

上述の議論で開示した発明の要旨の有用な応用は画像クラスタリングである。画像クラスタリングは、文書画像に対する良好な画像クラスタリングを実現するために、グラフラティス表現、サブグラフマッピング、接合点規格化マッピングカウントベクトル、および共通マイナス差の類似性尺度を使用する。接合点規格化マッピングカウントおよびＣＭＤの下で、試験は、より高次のサブグラフ特徴は判別の向上をもたらすことを示している。

【００７３】

画像をクラスタ化するために、単純欲張りクラスタ化アルゴリズムを使用できる。欲張りクラスタ化アルゴリズムの下で、「明確に同一のクラスタのしきい値」および「明確に異なるクラスタのしきい値」の２つのしきい値を設定する。これらのしきい値は手動で設定してもよく、画像の代表的サンプリングに対する、対をなすＣＭＤ距離のヒストグラムから自動的に推定してもよい。アルゴリズムは入力としてクラスタ化される予定の画像のコーパスを取り込む。

【００７４】

コーパス内の各画像に対して、アルゴリズムは最良適合クラスタを見つける。その画像と、既にクラスタの要素である画像との間の最良スコアが、画像に対する最良適合クラスタを決定する。これは、最近傍に基づいて、または要素をカテゴリにサンプリングする最良スコアに基づいて、画像をカテゴリに割り当てる画像分類プロセスに相当する。最良ス

コアは、接合点規格化マッピングカウントを用いて決定される特徴ベクトルに関するＣＭＤを用いて決定される。

【００７５】

画像に対する最良適合クラスタを見つけた後に、画像を分類する。最良適合クラスタと画像との類似性が、明確に同一のクラスタのしきい値よりも大きいとき、画像は最良適合クラスタに加えられる。最良適合クラスタと画像との類似性が、明確に異なるクラスタのしきい値よりも小さいとき、画像は唯一の要素としてその画像を有する新しいクラスタに加えられる。最良適合クラスタの類似性が、明確に同一のクラスタのしきい値と、明確に異なるクラスタのしきい値の間にあるとき、画像は「未確定」のカテゴリに入れられて、すべての画像を検討し終わるまで棚上げにされる。

【００７６】

コーパス内のすべての画像が分類されると、それぞれの未確定画像が再び取り上げられる。特定の実施形態では、未確定画像は、それらの最良適合クラスタに割り当てられる。他の実施形態では、上述のように未確定画像を既存のクラスタに加えることを試みる。このような実施形態の下で、明確に同一のクラスタのしきい値を超えることができない任意の画像は「余り」と呼ばれる新しいクラスタに加えられる。

【００７７】

欲張りクラスタ化アルゴリズムは、スキャナで取り込まれた手書きのおよびタイプされた米国納税申告書を代表する、大きさ２５６０ｘ３３００画素の１１，１８５枚の画像で構成された米国標準技術局（ＮＩＳＴ）納税申告書のデータコーパスで試験された。大きさ１～３または１～４のサブグラフを含む特徴ベクトルを用いて、クラスタ化アルゴリズムは、１つのカテゴリを２つに分けながら、すべての１１，１８５枚のＮＩＳＴ画像を、それらのそれぞれの２０のカテゴリに正確に分類した。図１３はサブグラフサイズが最大４までの特徴ベクトルを用いて２００のＮＩＳＴ文書の、対をなす類似性ヒストグラムを示している。ＮＩＳＴデータに対して、サブグラフ特徴サイズが２を超えると、異なる画像カテゴリが明確に分離される。最終的に、クラスタリング結果を図１４に示している。

【００７８】

クラスタリングの品質は、画像をカテゴリにグラウンドトルースに正しく割り当てるまでの編集距離として記録される。間違って分類されたそれぞれの文書に対して１つの編集操作が記録され、同じグラウンドトルースにカテゴリを表す任意の２つのクラスタを１つにまとめるために１つの編集操作が記録される。唯一のエラーはグラウンドトルースにカテゴリのうちの１つを複製する付加的なクラスタであるため、帳票クラスタリングおよび分類は、大きさ３以上のサブグラフに対してほぼ１００％正確である。

【００７９】

欲張りクラスタ化アルゴリズムについて上述したが、本明細書に開示する概念を基礎とする他のクラスタ化アルゴリズムも同様に受け入れられる。

【００８０】

上述の議論で開示した発明の要旨の他の有用な応用は画像分類である。画像分類は、グラフラティス表現、サブグラフマッピング、接合点規格化マッピングカウントベクトル、および共通マイナス差を使用する。画像分類は、分類に対する各カテゴリの１つ以上の見本の使用を通じてクラスタリングと同じように実行できる。すなわち、分類する予定の各画像に対して、アルゴリズムは見本の最良適合群を見つけ、画像と見本の間の最良スコアが最良適合群を決定する。上述のように、最良スコアは、接合点規格化マッピングカウントを用いて決定される特徴ベクトルに関するＣＭＤを用いて決定される。

【００８１】

グラフラティスは高速画像蓄積および検索の基礎としての役割を果たすことができる。グラフラティスの、観察されるデータグラフ上への照合は、グラフ・ラティス・ノードのサブグラフの、データグラフのサブグラフ上へのマッピングのマッピングセットを作ることを含む。これらのマッピングはマッピングの身元および配置を記録する。新しい画像が観察されると、各グラフ・ラティス・ノードによりマッピングされた、したがって、一般

的構造を共有する他の画像は、これらのマッピングから取り出される。雑音およびサンプル変化に起因するような不完全なデータグラフの条件下で、目標と共通して多くのサブグラフを共有する観察されるサンプルから画像を選択するために公知の投票方法を使用できる。

【００８２】

グラフラティスは画像の中の反復構造を検出する基盤としての役割を果たすことができる。グラフラティスの、観察されるデータグラフ上への照合は、グラフ・ラティス・ノードのサブグラフの、データグラフのサブグラフ上へのマッピングのマッピングセットを作ることを含む。同じ画像の異なる領域への複数のマッピングは、その画像内の反復構造を示している。多くの重複サブグラフは、単に反復部分についての人間の直観に対応するサブグラフではなく、繰り返されることが分かっている。

【００８３】

図１５では、反復構造は、１）周期的反復構造、および２）孤立反復構造の２つの大きなカテゴリで生じる。周期的反復構造（図１５の「ａ」）は、反復構造領域がそれ自体とともに境界を共有するときに生じる。これは、反復パターンの境界を定義する際に、エイリアシングまたは位相アンビギュイティの問題を引き起こす。孤立反復構造（図１５の「ｂ」）は、反復領域を囲む材料が、反復領域の１つの例とその隣の例とで全般的に異なっているときに生じる。

【００８４】

重複のない目標ノードで表されるサブグラフに正確にＲ回マッピングするレベル（Ｌ／Ｒ）のグラフ・ラティス・ノードが存在するときには、レベルＬのグラフ・ラティス・ノードはＲ回反復される構造を表している。このような反復ノードは各レベルＬで各ノードを順々に試験することにより検出できる。目標ノードで表されるサブグラフが形成され、その後、サブグラフマッチング用の単純アルゴリズムを使用してレベル（Ｌ／Ｒ）の候補ノードを一度に１つずつ照合する。候補ノードが正確にＲマッピングを有しているとき、目標サブグラフの各接合点がＲマッピングにより正確に１回マッピングされるかどうかを判断する。目標サブグラフの各接合点がＲマッピングにより正確に１回マッピングされると判断された場合、目標レベルノードは反復構造ノードであり、それが含む反復構造は候補レベル（Ｌ／Ｒ）ノードで表される。この方法は、棒グラフで大きさが最大６の接合点までの反復構造を検出することが試験され示されている。

【００８５】

図１６では、グラフ・ラティス・システム１６００を示している。記憶装置と、マイクロプロセッサ、マイクロコントローラ、映像処理装置（ＧＰＵ）などのデジタル／電子プロセッサと、を含むコンピュータ１６０２または他のデジタル／電子制御演算装置がシステム１６００を具現化することがふさわしい。他の実施形態では、システム１６００はデジタルプロセッサを含み、デジタルデータ記憶装置を含むか、もしくはデジタルデータ記憶装置にアクセスできるサーバにより具現化され、このようなサーバはインターネットもしくはローカル・エリア・ネットワークを介してアクセスされることがふさわしく、またはシステム１６００はデジタルプロセッサおよびデジタルデータ記憶装置などを含む携帯情報端末（ＰＤＡ）により具現化される。

【００８６】

コンピュータ１６０２または他のデジタル制御演算装置は、制御システム１６００へのユーザ入力を受信するキーボード１６０４のような１つ以上のユーザ入力装置を含み、またはこのような１つ以上のユーザ入力装置と動作的に接続されていることがふさわしく、コンピュータ１６０２または他のデジタル処理装置は、システム１６００の出力に基づいて生成された出力を表示する表示部１６０６のような１つ以上の表示装置をさらに含み、またはこのような１つ以上のユーザ入力装置と動作的に接続されていることがふさわしい。他の実施形態では、制御システム１６００に対する入力は、コンピュータ１６０２上のシステム１６００に先立って起動している、もしくはシステム１６００と同時に起動している他のプログラムから受信し、またはネットワーク接続などから受信する。同様に、他

の実施形態では、出力はコンピュータ上のシステム１６００の後で起動している、もしくはシステム１６００と同時に起動している他のプログラムへの入力としての役割を果たしてもよく、またはネットワーク接続などを介して伝達されてもよい。

【００８７】

システム１６００は、本願のグラフラティスの１つ以上の態様を実現するグラフ・ラティス・モジュール１６０８と、グラフラティスを用いる方法および／またはアルゴリズムとを含んでいる。特定の実施形態では、グラフ・ラティス・モジュール１６０８はモジュール１６０８の外部のソースから１つ以上の画像のコーパスを受信し、コーパスからグラフラティスを生成する。このような実施形態の一部では、グラフ・ラティス・モジュール１６０８は目標画像をさらに受信し、この目標画像はグラフ・ラティス・モジュール１６０８がコーパスから類似画像を取り出すために使用する。このような実施形態のその他では、グラフ・ラティス・モジュールは画像のコーパス上でクラスタ化を実行し、および／またはコーパスの中の反復サブグラフを特定する。

【００８８】

いくつかの実施形態では、グラフ・ラティス・モジュール１６０８は、実行可能命令を保存する記憶媒体により、例えば、デジタルプロセッサにより具現される。記憶媒体は、例えば、磁気ディスクもしくは他の磁気記憶媒体か、光ディスクもしくは他の光記憶媒体か、ランダム・アクセス・メモリ（ＲＡＭ）、読み出し専用メモリ（ＲＯＭ）、もしくは他の電子メモリ素子もしくはチップもしくは動作的に相互接続したチップセットか、保存された命令をそこからインターネットもしくはローカル・エリア・ネットワークを介して取り出してもよいインターネット・サーバなどを含んでいてもよい。

【００８９】

図１７では、図１６のグラフ・ラティス・システム１６００を用いるコンピュータ・ビジョン・システム１７００を示している。コンピュータ・ビジョン・システム１７００は、撮像装置１７０２と、図１６のグラフ・ラティス・システム１７０４と、を含んでいる。特定の実施形態では、コンピュータ・ビジョン・システム１７００は、例えば、通信ネットワークを経由してコンピュータ・ビジョン・システム１７０４に動作的に接続された文書データベース１７０６をさらに含んでいる。文書データベース１７０６は文書画像のデータベースであり、文書画像は撮像装置１７０２のような装置を介して生成される。

【００９０】

撮像装置１７０２は１つ以上の文書１７０８を受け取り、それらを文書画像１７１０に変換する。撮像装置はカメラ、スキャナ、または他の類似装置であってもよい。さらに、撮像装置１７０２は給紙トレイから延びるコンベヤ経路を介して文書を受け取ってもよい。

【００９１】

その後、グラフ・ラティス・システム１７０４は文書画像１７１０を受信し、それらを用いて１つ以上の作業を実行する。グラフ・ラティス・システム１７０４は通信ネットワークを介して電子的に文書画像１７１０を受信してもよい。さらに、１つ以上の作業は、クラスタ化された文書画像１７１２を生成するために１つ以上のクラスタ化する文書画像１７１０を含み、文書データベース１７０６の中の類似文書画像１７１４を見つけ、文書画像１７１０の中の反復構造１７１６を見つけてもよい。図示の作業にかかわらず、グラフ・ラティス・システム１７０４は、図示していない付加的な作業（例えば、文書分類）を実行できる。

【００９２】

グラフ・ラティス・システム１７０４を使用して、クラスタ化された文書画像１７１２を生成するシナリオの下で、第４．３項（画像分類およびクラスタリング）に関連して説明するように文書画像１７１０をクラスタ化して、クラスタ化された文書画像１７１２を定義する。すなわち、文書画像１７１０を比較するＣＭＤ類似性スコアが、文書画像１７１０のグラフラティスを用いて生成され、文書画像１７１０をクラスタ化するために使用される。

【００９３】
グラフ・ラティス・システム１７０４が文書画像１７１０をクラスタ化すると、必要に応じて文書１７０８および／または文書画像１７１０が処理される。例えば、文書１７０８は、それらのクラスタに基づいて目的地までコンベヤ経路を介して送ってもよい。あるいは、または加えて、文書画像１７１０は、それらのクラスタに従ってデータベース内に保存してもよく、および／またはファイルシステムの中に保存してもよい。
【００９４】
グラフ・ラティス・システム１７０４を使用して類似文書画像１７１４を見つけるシナリオの下で、文書画像１７１０を使用して文書データベース１７０６から類似文書画像１７１４を取り出す。これは画像インデキシング、記憶装置、および検索に関連して説明するように実行される。すなわち、文書データベース１７０６内の文書画像のグラフラティスを文書画像１７１０にマッピングする。その後、単純投票方法を使用して、文書画像１７１０と共通して最も多くの構造を有する、文書データベース１７０６内の文書画像を見つける。
【００９５】
グラフ・ラティス・システム１７０４が類似文書画像１７１４を見つけると、必要に応じてそれらの類似文書画像１７１４を処理してもよい。例えば、類似文書画像１７１４はデータベース内に保存してもよく、および／またはファイルシステムの中に保存してもよい。あるいは、または加えて、類似文書画像１７１４は表示部および／またはプリンタを介してコンピュータ・ビジョン・システム１７００のオペレータに提供してもよい。
【００９６】
グラフ・ラティス・システム１７０４を使用して反復構造１７１６を見つけるシナリオの下で、反復構造を探して文書画像１７１０を検索する。これは共通構造および反復構造の検出に関連して説明するように実行される。文書画像１７１０を使用してグラフラティスを生成し、その後、重複のない目標ノードで表されるサブグラフに正確にＲ回マッピングするレベル（Ｌ／Ｒ）のグラフ・ラティス・ノードが存在するときには、レベルＬのグラフ・ラティス・ノードはＲ回反復される構造を表しているという了解の下で、各レベルＬで各ノードを順々に試験することにより反復ノードを検出する。グラフ・ラティス・システム１７０４が反復構造１７１６を見つけると、必要に応じて反復構造を処理してもよい。例えば、反復構造は表示部および／またはプリンタを介してコンピュータ・ビジョン・システム１７００のオペレータに提供してもよい。
【００９７】
グラフラティスおよびグラフラティスに適用するアルゴリズムは、画像クラスタリングの効率、精度、およびスケーラビリティと、分類と、類似の、および重複した画像インデキシングおよび検索と、反復構造検出とを促進する。効率は、グラフラティスの単純なパターンから、より複雑なパターンまでを作るサブグラフへのマッピングを計算するアルゴリズムから導出する。精度は、グラフラティス内のサブグラフの非常に大きな集合の保存によるグラフラティスの冗長性から導出する。スケーラビリティは、観察されるデータに合わせた大きなグラフラティス（このグラフラティスは、すべてのサブグラフの空間に比べれば、まだはるかに小さい）を育てるように我々が開示するアルゴリズムから導出する。
【００９８】
下記の変形は予測できると考えられる。
【００９９】
（ｖｉ．）　観察されるデータからグラフラティスを適応的に大きく育てる方法。特に重要な問題はグラフラティスの中に経路を深く延ばすことであり、その結果、より小さいサブグラフの急増にノードをささげることなく、大きなサブグラフを示す。これは、さらに耐雑音性を獲得するほど十分な冗長性を備えた上で、高レベルノードまで経路を選択的に延ばすことを意味している。
【０１００】

（ｖｉｉ．）　作業またはデータに依存する偶然性に従って、グラフラティスの、データ上へのマッピングを選択的に計算する方法。

【０１０１】

（ｖｉｉｉ．）　クラスタ、反復構造、および例外的パターンを発見するためにデータサンプル上のグラフラティスのマッピングカウントを分析する方法。

【０１０２】

（ｉｘ．）　埋め込み、カーネル法、および他の統計的パターン認識法に対して特徴ベクトルに関与するためにグラフ・ラティス・ノードを選択する方法。

【図１】

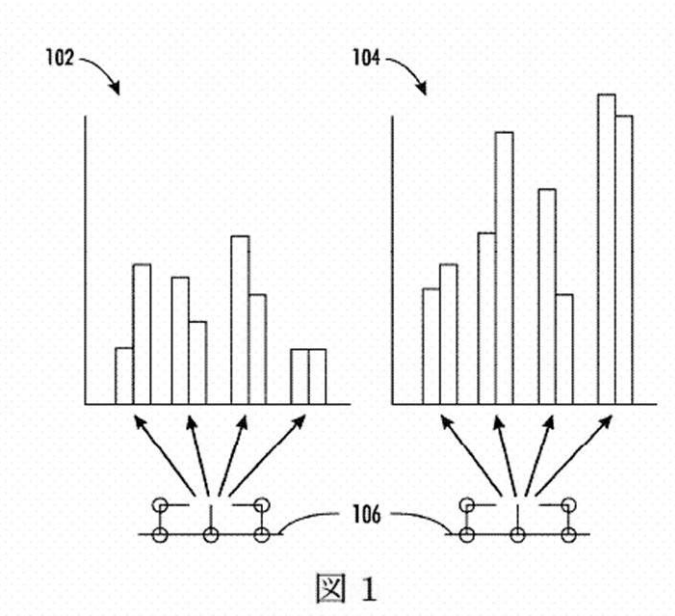

図１

【図２】

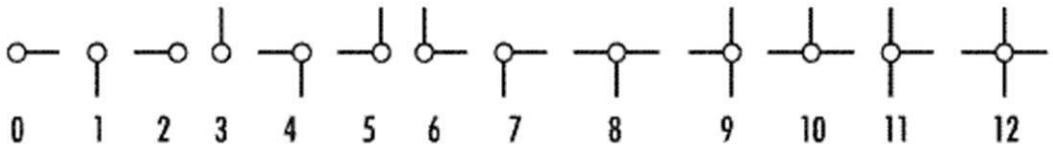

図２

【図３】

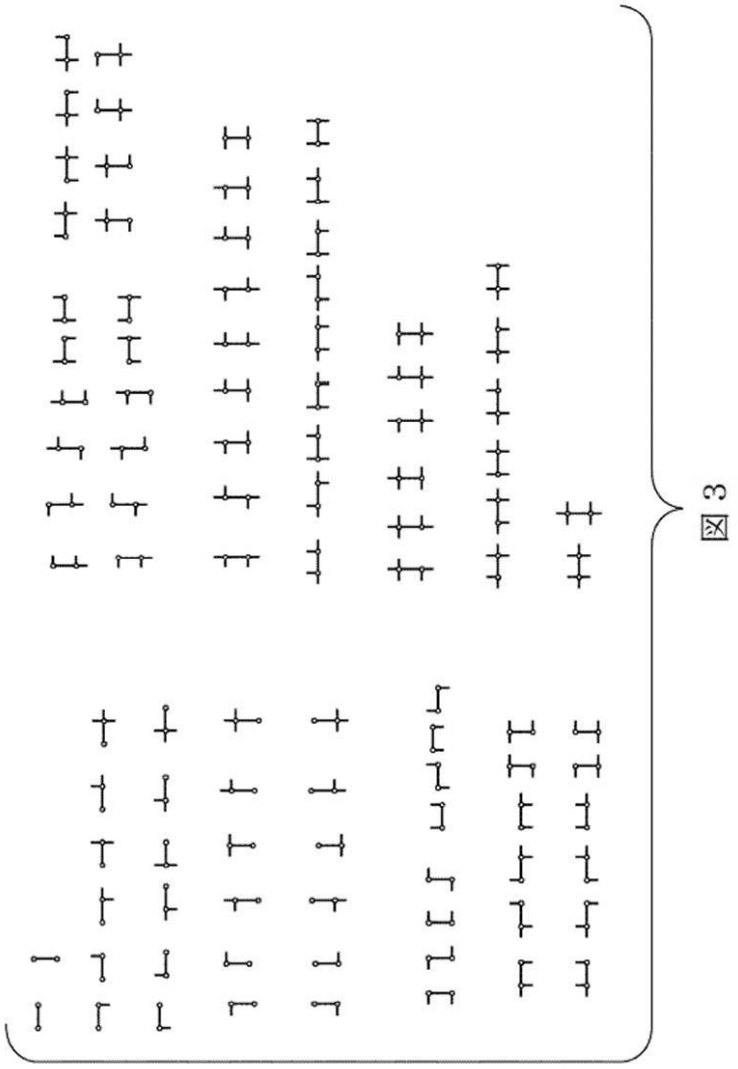

図３

【図４】

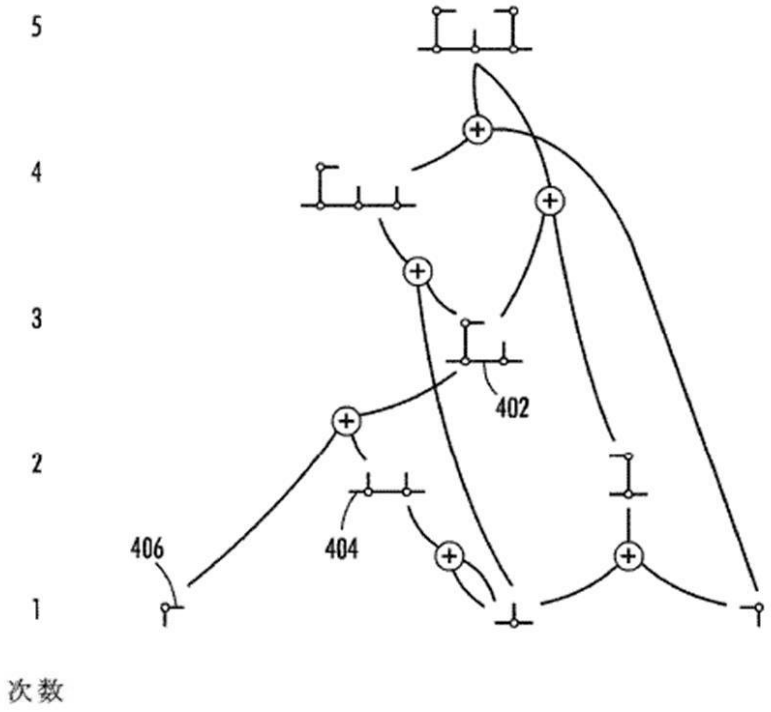

図４

【図５】

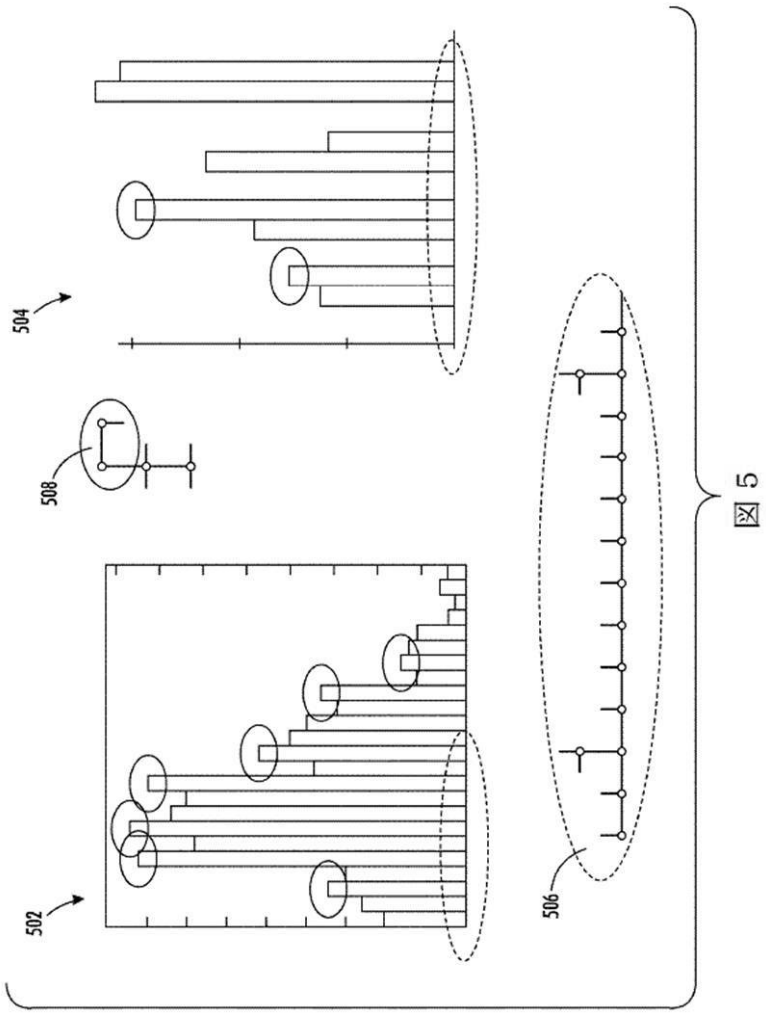

図５

【図６】

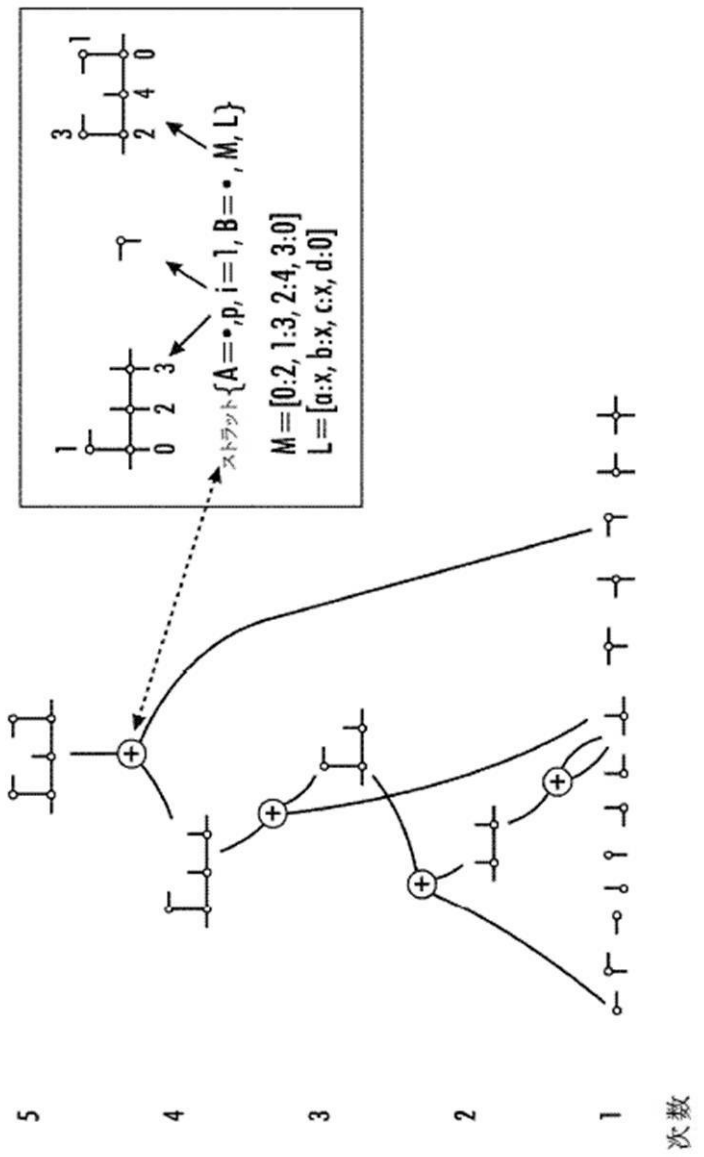

図６

【図７】

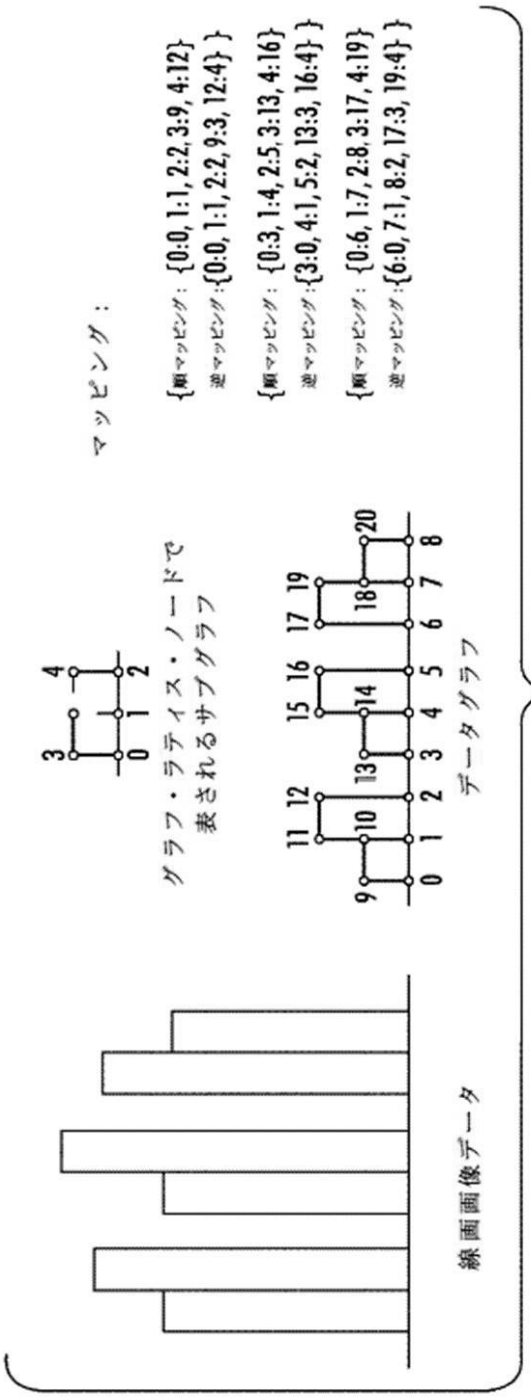
線画画像データ
データグラフ
0 1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16 17 18 19 20
グラフ・ラティス・ノードで
表されるサブグラフ
0 1 2 3 4
マッピング：
{順マッピング：{0:0, 1:1, 2:2, 3:9, 4:12}
逆マッピング：{0:0, 1:1, 2:2, 9:3, 12:4}}
{順マッピング：{0:3, 1:4, 2:5, 3:13, 4:16}
逆マッピング：{3:0, 4:1, 5:2, 13:3, 16:4}}
{順マッピング：{0:6, 1:7, 2:8, 3:17, 4:19}
逆マッピング：{6:0, 7:1, 8:2, 17:3, 19:4}}

図７

【図８】

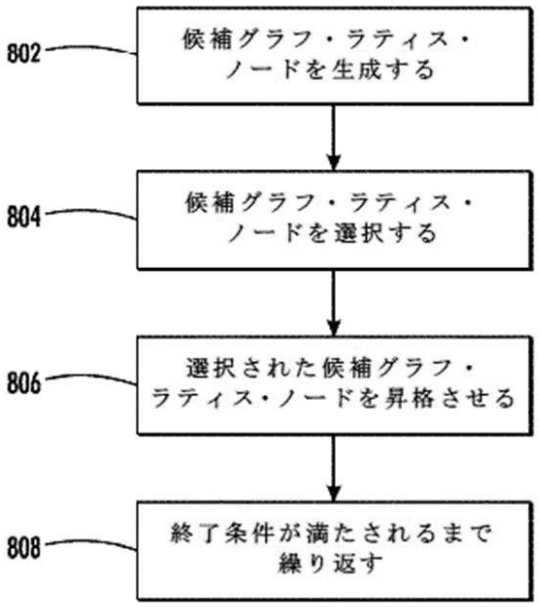
802
候補グラフ・ラティス・
ノードを生成する
804
候補グラフ・ラティス・
ノードを選択する
806
選択された候補グラフ・
ラティス・ノードを昇格させる
808
終了条件が満たされるまで
繰り返す

図８

【図９】

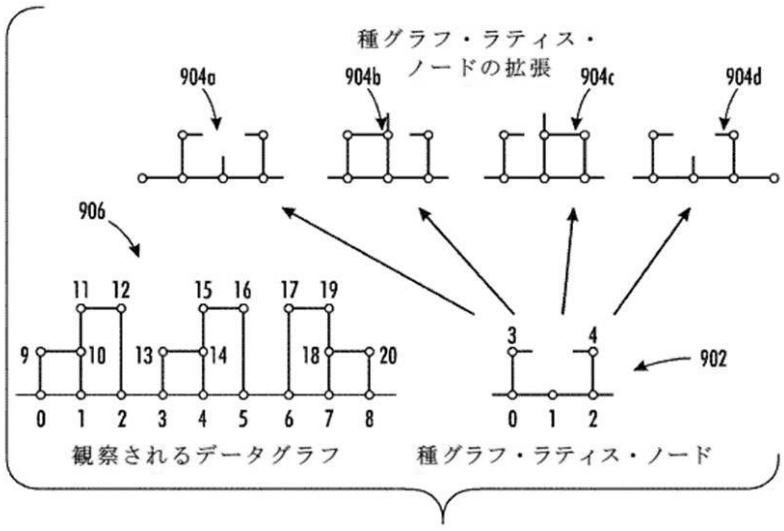
種グラフ・ラティス・
ノードの拡張
904a
904b
904c
904d
906
11 12
15 16
17 19
9 10
13 14
18 20
0 1 2 3 4 5 6 7 8
観察されるデータグラフ
3 4
902
0 1 2
種グラフ・ラティス・ノード

図９

【図１０】

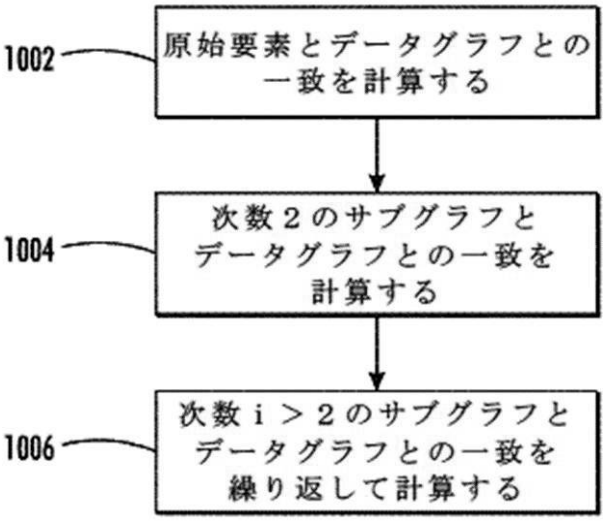

図１０

【図１１】

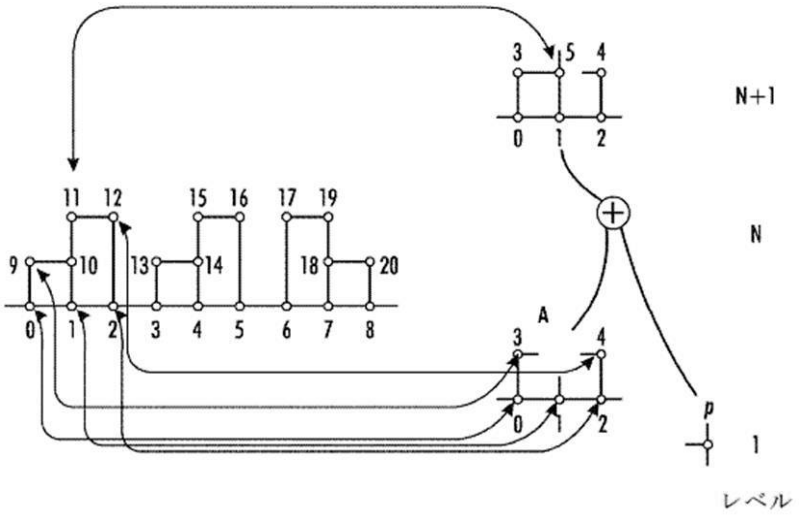

図１１

【図１２】

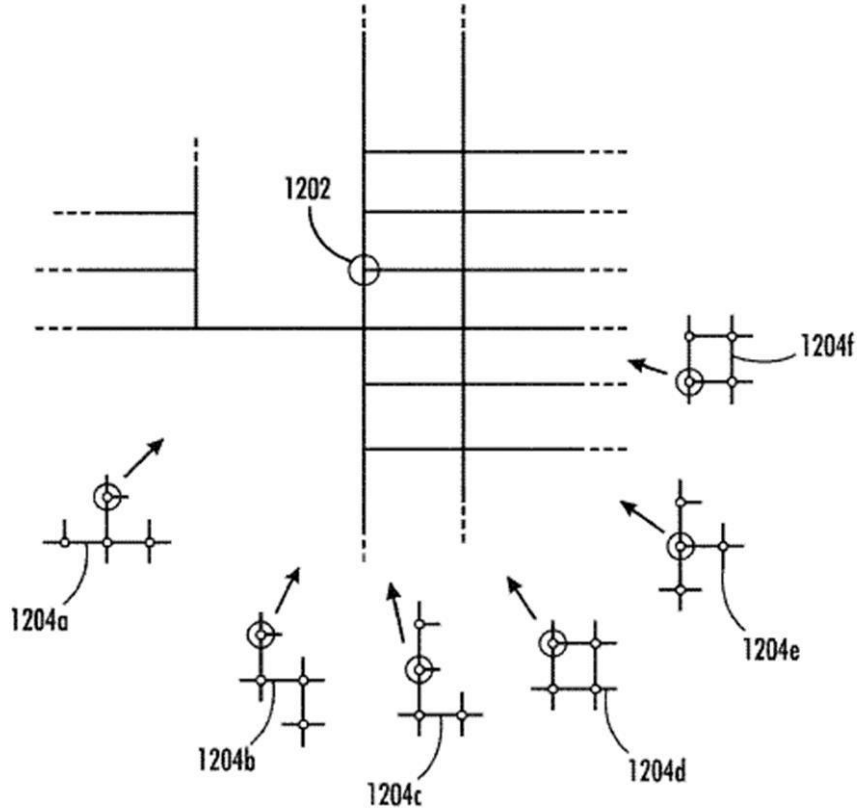

図１２

【図１３】

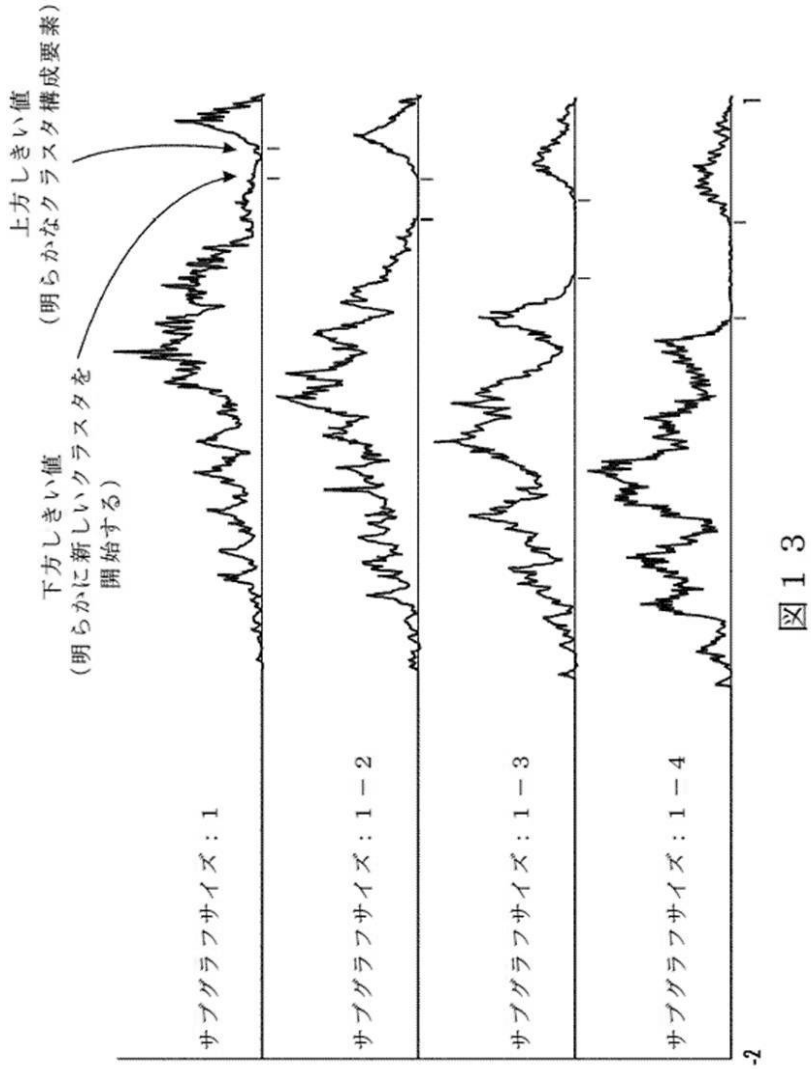

図１３

【図１４】

サブグラフサイズ	特徴ベクトルサイズ	エラー	偽クラスタ	クラスタ品質（グラウンドトルースまでの編集距離）
1	13	6	18	24
1-2	87	3	5	8
1-3	534	0	1	1
1-4	2953	0	1	1

図１４

【図１５】

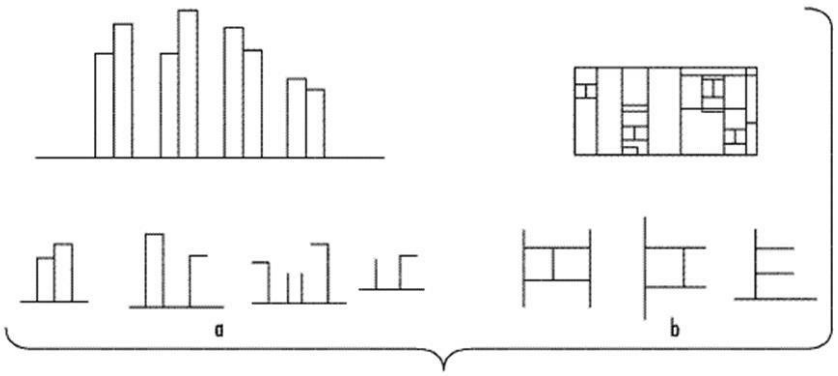

図１５

【図１６】

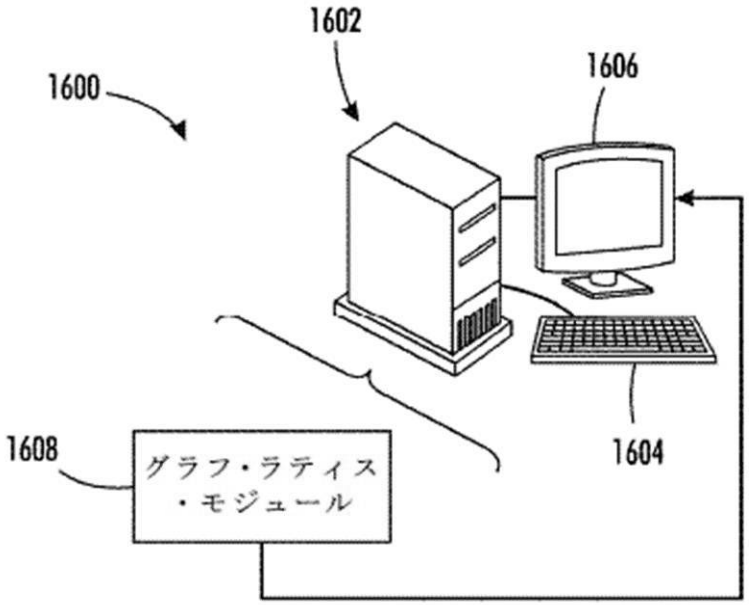

図１６

【図１７】

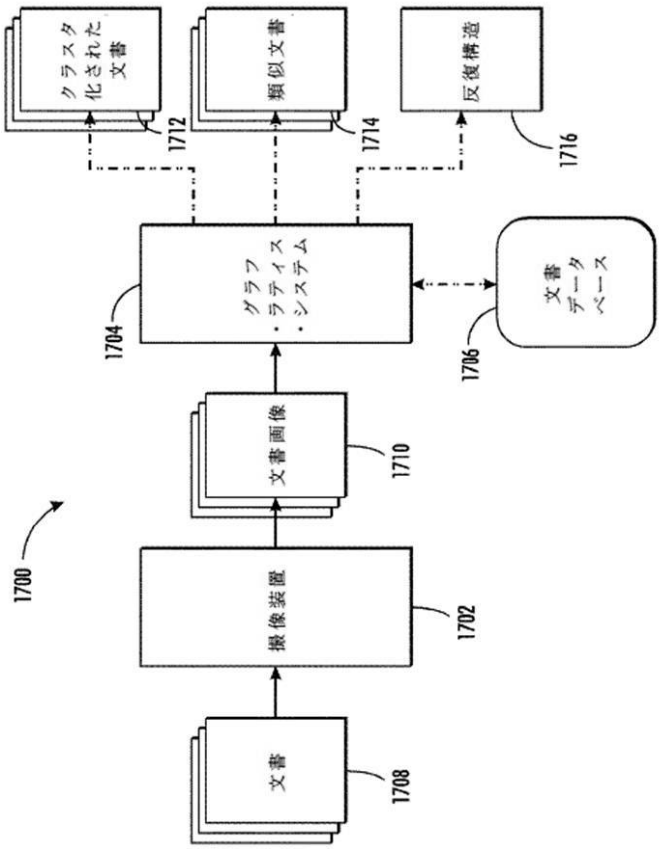
クラスタ化された文書
1712
類似文書
1714
反復構造
1716
グラフ・ラティス・システム
1704
文書データベース
1706
文書画像
1710
撮像装置
1702
文書
1708
1700

図１７

フロントページの続き

(72)発明者　エリック・サウンド
アメリカ合衆国　カリフォルニア州　９４０７０　サン・カーロス　クリフトン・アヴェニュー　４６９

審査官　佐田　宏史

(56)参考文献　米国特許出願公開第２００６／０１８２３１７（ＵＳ，Ａ１）
米国特許出願公開第２００２／０１６８０９３（ＵＳ，Ａ１）
B. Ozdemir, S. Aksoy, "Image Classification Using Subgraph Histogram Representation", 2010 20th International Conference on Pattern Recognition (ICPR), 米国, IEEE, ２０１０年　8月２３日, p.1112-1115
Monika Akbar, Rafal A. Angryk, "Frequent Pattern-Growth Approach for Document Organization", ONISW '08 Proceedings of the 2nd international workshop on Ontologies and information systems for the semantic web, 米国, ACM, ２００８年１０月３０日, p.77-82
E. Barbu et al, "Clustering document images using a bag of symbols representation", Proceedings of the 2005 Eighth International Conference on Document Analysis and Recognition, 米国, IEEE, ２００５年　8月２９日, Vol.2, p.1216-1220
Chuntao Jiang, Frans Coenen, "Graph-based Image Classification by Weighting Scheme", Applications and Innovations in Intelligent Systems XVI, 英国, Springer London, ２００９年１２月３１日, p.63-76, ＵＲＬ, https://cgi.csc.liv.ac.uk/~frans/PostScriptFiles/gicw-ai08.pdf
I. Fischer, T. Meinl, "Graph Based Molecular Data Mining - An Overview", 2004 IEEE International Conference on Systems, Man and Cybernetics, 米国, IEEE, ２００４年１０月１０日, Vol.5, p.4578-4582

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｔ　　１／００，７／００－７／６０
Ｇ０６Ｋ　　９／００
Ｈ０４Ｎ　　１／３８７，１／４１