(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2014-106961
(P2014-106961A)
(43)公開日　平成26年6月9日(2014.6.9)

(51)Int.Cl.			F I			テーマコード（参考）
G06K	9/62	(2006.01)	G06K	9/62	630Z	5B064
G06T	7/00	(2006.01)	G06T	7/00	350A	5L096

審査請求 未請求　請求項の数 16　ＯＬ　（全 22 頁）

(21)出願番号　特願2013-118680 (P2013-118680)
(22)出願日　平成25年6月5日 (2013.6.5)
(31)優先権主張番号　13/685,088
(32)優先日　平成24年11月26日 (2012.11.26)
(33)優先権主張国　米国 (US)

(71)出願人　511072895
キング・アブドゥルアジズ・シティ・フォー・サイエンス・アンド・テクノロジー（ケイ・エイ・シィ・エス・ティ）
ＫＩＮＧ ＡＢＤＵＬＡＺＩＺ ＣＩＴＹ ＦＯＲ ＳＣＩＥＮＣＥ ＡＮＤ ＴＥＣＨＮＯＬＯＧＹ （ＫＡＣＳＴ）
サウジアラビア、１１４４２ リヤド、ピィ・オゥ・ボックス・６０８６、ザ・ナショナル・センター・フォー・テクノロジー・ディベロップメント
(74)代理人　110001195
特許業務法人深見特許事務所

最終頁に続く

(54)【発明の名称】アラビア語テキストを自動的に認識するためのコンピュータによって実行される方法、およびコンピュータプログラム

(57)【要約】
【課題】アラビア語テキストの認識において、適切にテキスト特徴を抽出する。
【解決手段】アラビア語の文字のラインがデジタル化されることにより、各々が２進数で表現された画素値に関連付けられた二次元の画素の配列が形成される。画素値は２進数で表現される。さらに、アラビア語の文字のラインが複数のライン画像へと分割されて、複数のライン画像の中の１つにおいて複数のセルが規定される。複数のセルの各々は、隣接した画素のグループを有する。さらに、複数のライン画像の中の１つにおいて複数のセルの各々の画素の画素値がシリアル化されることにより、２値セル番号が形成される。また、複数のライン画像の中の１つにおける複数のセルから取得された２値セル番号に従ってテキスト特徴ベクトルが形成される。そして、テキスト特徴ベクトルが隠れマルコフモデルに送られることによりアラビア語の文字のラインが認識される。
【選択図】図１

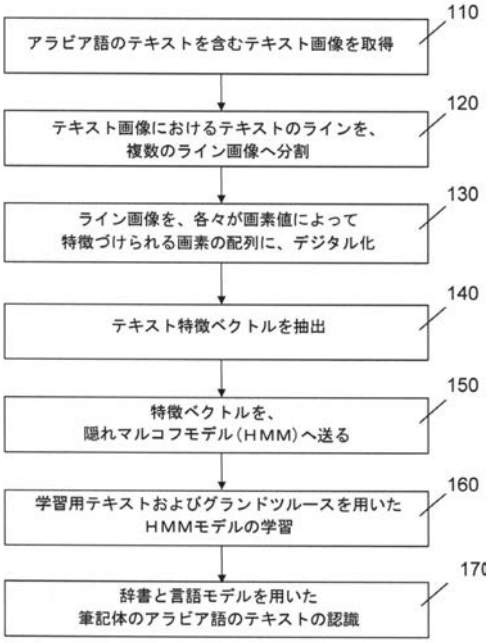

Figure 1

【特許請求の範囲】
【請求項１】
アラビア語テキストを自動的に認識するための、コンピュータによって実行される方法であって、
アラビア語の文字のラインを含むテキスト画像を取得することと、
当該アラビア語の文字のラインをデジタル化することにより、各々が２進数で表現された画素値に関連付けられた二次元の画素の配列を形成することとを備え、前記二次元の画素の配列は、第１の方向における複数の行と第２の方向における複数の列とを含み、
前記方法は、さらに、
画素の列における画素のストリング中の同じ画素値の連続する画素の頻度をカウントすることを備え、各々が異なる画素値を有する隣接した画素のストリングは、それらの間での遷移によって規定され、前記カウントすることは、さらに、
列の遷移数が予め定められた足切遷移番号に達したときに、当該列における同じ画素値の連続する画素の頻度のカウントを停止することと、
前記画素の列におけるストリングから取得される頻度カウントを用いてテキスト特徴ベクトルを形成することと、
当該テキスト特徴ベクトルを隠れマルコフモデルに送ることによりアラビア語の文字のラインを認識することとを備える、方法。
【請求項２】
前記アラビア語の文字のラインは、複数のアラビア語の単語を含む、請求項１に記載のコンピュータによって実行される方法。
【請求項３】
前記テキスト特徴ベクトルは、前記画素の列における連続する画素のストリングから取得された一連の頻度カウントによって形成される、請求項１に記載のコンピュータによって実行される方法。
【請求項４】
前記予め定められた足切遷移番号は、前記アラビア語の文字のラインをデジタル化するステップに先立つ、アラビア語テキストについての統計的解析によって取得される、請求項１に記載のコンピュータによって実行される方法。
【請求項５】
前記予め定められた足切遷移番号は６である、請求項１に記載のコンピュータによって実行される方法。
【請求項６】
前記二次元の配列における画素値は、単一のビットの２進数で表現される、請求項１に記載のコンピュータによって実行される方法。
【請求項７】
前記頻度をカウントすることは、
列における最初の１またはそれ以上の画素の画素値が「０」であるときに、第１の頻度カウントの値に「０」を割り当てることを含み、前記第１の頻度カウントの次に、当該列の初めに画素値「０」を有する連続した画素の数が続く、請求項６に記載のコンピュータによって実行される方法。
【請求項８】
前記頻度をカウントすることは、
列の頂点の１またはそれ以上の画素の画素値が「１」であるときに、第１の頻度カウントの値として「０」を割り当てることを含み、前記第１の頻度カウントの次に、当該列の初めに画素値「１」を有する連続した画素の数が続く、請求項６に記載のコンピュータによって実行される方法。
【請求項９】
コンピュータに以下のことを実行させるためのコンピュータ読取可能なプログラムであって、プログラムコード関数を含み、前記プログラムコード関数は、コンピュータに、

アラビア語の文字のラインを含むテキスト画像を取得させ、
アラビア語の文字のラインをデジタル化させることにより、各々が２進数で表現された画素値に関連付けられた二次元の画素の配列を形成させ、前記二次元の画素の配列は、第１の方向における複数の行と第２の方向における複数の列とを含み、
前記プログラムコード関数は、さらに、前記コンピュータに、画素の列における画素のストリング中の同じ画素値の連続する画素の頻度をカウントさせ、各々が異なる画素値を有する隣接した画素のストリングはそれらの間での遷移によって規定され、前記カウントするステップは、さらに、前記列における遷移の数が予め定められた足切遷移番号に到達したときに、同じ画素値の連続する画素の頻度のカウントを停止することを含み、
前記プログラムコード関数は、前記コンピュータに、
前記画素列におけるストリングから取得される頻度カウントを用いてテキスト特徴ベクトルを形成することと、
前記テキスト特徴ベクトルを隠れマルコフモデルに送ることによりアラビア語の文字のラインを認識することとを実行させる、コンピュータプログラム。
【請求項１０】
前記アラビア語の文字のラインラインは、複数のアラビア語の単語を含む、請求項９に記載のコンピュータプログラム。
【請求項１１】
前記テキスト特徴ベクトルは、前記画素の列における連続する画素のストリングから取得された一連の頻度カウントによって形成される、請求項９に記載のコンピュータプログラム。
【請求項１２】
前記予め定められた足切遷移番号は、前記アラビア語の文字のラインをデジタル化するステップに先立つ、アラビア語テキストについての統計的解析によって取得される、請求項９に記載のコンピュータプログラム。
【請求項１３】
前記予め定められた足切遷移番号は６である、請求項９に記載のコンピュータプログラム。
【請求項１４】
前記二次元の配列における画素値は、単一のビットの２進数で表現される、請求項９に記載のコンピュータプログラム。
【請求項１５】
前記頻度をカウントするステップは、
列における最初の１またはそれ以上の画素の画素値が「０」であるときに、第１の頻度カウントの値に「０」を割り当てることを含み、前記第１の頻度カウントの次に、当該列の初めに画素値「０」を有する連続した画素の数が続く、請求項９に記載のコンピュータプログラム。
【請求項１６】
前記頻度をカウントするステップは、
列の頂点の１またはそれ以上の画素の画素値が「１」であるときに、第１の頻度カウントの値として「０」を割り当てることを含み、前記第１の頻度カウントの次に、当該列の初めに画素値「１」を有する連続した画素の数が続く、請求項９に記載のコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本特許出願は、同一発明者によって２０１１年１２月１４日に出願され、同一譲受人の、係属する米国特許出願第１３／３２５，７８９号、名称「効果的なアラビア語テキスト特徴の抽出に基づく、アラビア語テキスト認識のためのシステムおよび方法」の継続出願であり、当該出願についての優先権を主張する。米国特許出願第１３／３２５，７８９号

は、同一発明者によって２００９年４月２７日に出願された、米国特許出願第１２／４３０，７７３号、名称「効果的なアラビア語テキスト特徴の抽出に基づく、アラビア語テキスト認識のためのシステムおよび方法」の継続出願であり、その開示は、ここに参照により組み込まれる。

【背景技術】

【０００２】

発明の背景

本願は、概して、アラビア語テキストの自動的な認識に関する。

【０００３】

テキスト認識、つまり、テキストの自動読取は、パターン認識の一分野である。テキスト認識の目的は、印刷されたテキストを、人間の精度で、かつ、より速く、読取ることである。多くのテキスト認識の方法は、テキストが個々の文字へと分離できることを前提としている。このような技術では、タイプライタで打たれた、または、活字に組まれたラテン語については首尾よくいくが、アラビア語のような筆記体には、信頼できる程度に適用することはできない。これまでのアラビア語の手描きテキストの認識についての研究によれば、アラビア語の単語を個々の文字へとセグメント化する試みにおける困難性が確認されている。

【０００４】

【数１】

アラビア語は、テキスト認識のアルゴリズムにいくつかの課題を提供している。アラビア語の手描き文字は、本質的に、筆記体であり、分離された文字をブロック体で記述することは容認できない。しかも、アラビア文字の形は、文脈に依存することがあり、つまり、単語における当該文字の位置に依存し得る。たとえば、次の「ه」に示されるような文字は、４つの異なる外形、つまり、「عبده」における独立形「ه」、「هند」における語頭形「هـ」、「نهر」における語中形「ـهـ」、「منه」における語尾形「ـه」を有する。さらに、すべてのアラビア文字が、単語の中で連結されるわけではない。単語内において、空間が特定の文字を分割する場合もあるため、単語間の境界を自動的に決定することは困難な場合がある。

【０００５】

アラビア語テキストの認識には、統計モデルのような異なる分類体系が適用されてきた。しかしながら、適切にテキスト特徴を抽出することは、未だ、正確なアラビア語テキストの認識を達成することにおいての主要な障害のままである。

【発明の概要】

【発明が解決しようとする課題】

【０００６】

発明の概要

概略的な側面において、本願発明は、アラビア語テキストを自動的に認識するための方法に関する。当該方法は、アラビア語の文字のラインを含むテキスト画像を取得することと、アラビア語の文字のラインをデジタル化することにより、各々が画素値に関連付けられた二次元的の画素の配列を形成することとを含み、画素値は２進数で表現され、上記方法は、さらに、アラビア語の文字のラインを複数のライン画像へと分割することと、複数のライン画像の中の１つにおいて複数のセルを規定することとを含み、複数のセルの各々は、隣接した画素のグループを有し、上記方法は、さらに、複数のライン画像の中の１つ

において複数のセルの各々の画素の画素値をシリアル化することにより２値セル番号を形成することと、複数のライン画像の中の１つにおける複数のセルから取得された２値セル番号に従ってテキスト特徴ベクトルを形成することと、テキスト特徴ベクトルを隠れマルコフモデル（Hidden Markov Model）に送ることによりアラビア語の文字のラインを認識することとを含む。

【０００７】

他の概略的な局面において、本願発明は、アラビア語テキストを自動的に認識するための方法に関する。当該方法は、アラビア語の文字のラインを含むテキスト画像を取得すること、アラビア語の文字のラインをデジタル化することにより、各々が２進数で表現された画素値に関連付けられた二次元の画素の配列を形成することを含み、二次元の画素の配列は、第１の方向における複数の行と、第２の方向における複数の列とを含み、上記方法は、さらに、画素の列において同じ画素値を有する連続する画素の頻度をカウントすることと、画素の列から得られた頻度カウントを利用してテキスト特徴ベクトルを形成することと、当該テキスト特徴ベクトルを隠れマルコフモデルへ送ることによりアラビア語の文字のラインを認識することとを含む。

【０００８】

他の概略的な局面において、本願発明は、アラビア語テキストを自動的に認識するための方法に関する。当該方法は、アラビア語の文字のラインを含むテキスト画像を取得することと、アラビア語の文字のラインをデジタル化することにより、各々が画素値に関連付けられた二次元の画素の配列を形成することと、当該アラビア語の文字のラインを複数のライン画像へと分割することと、当該複数のライン画像の少なくとも１つを小型化することにより小型化されたライン画像を生成することと、小型化されたライン画像の各々の列の画素の画素値をシリアル化することにより一連のシリアル化された番号を形成することとを含み、一連のシリアル化された番号はテキスト特徴ベクトルを形成し、上記方法は、さらに、当該テキスト特徴ベクトルを隠れマルコフモデルへ送ることによりアラビア語の文字のラインを認識することとを含む。

【０００９】

他の概略的な局面において、本願発明は、コンピュータ読取可能なプログラムコード関数を含むコンピュータプログラムに関し、当該コード関数は、コンピュータに、アラビア語の文字のラインを含むテキスト画像を取得させ、アラビア語の文字のラインをデジタル化させることにより、各々が画素値に関連付けられた二次元の画素の配列を形成させ、当該画素値は２進数で表現され、上記コード関数は、上記コンピュータに、さらに、アラビア語の文字のラインを複数のライン画像へと分割させ、複数のライン画像の中の１つにおける複数のセルを規定させ、複数のセルの各々は隣接する画像のグループを有し、上記コード関数は、上記コンピュータに、さらに、複数のライン画像の中の１つにおける複数のセルの各々の画素の画素値をシリアル化させ、複数のライン画像の中の１つにおける複数のセルから取得された２進数のセル番号に応じてテキスト特徴ベクトルを形成させ、当該テキスト特徴ベクトルを隠れマルコフモデルに送ることによりアラビア語の文字のラインを認識させる。

【００１０】

システムの実現は、以下に示されたもののうち１またはそれ以上を含む場合がある。上記方法は、さらに、２進数のセル番号を１０進数のセル番号へと変換することと、複数のライン画像の中の１つにおける複数のセルから取得された１０進数のセル番号をシリアル化することにより一連の１０進数のセル番号を形成することと、複数のライン画像の中の１つにおける複数のセルから取得された一連の１０進数のセル番号に従ってテキスト特徴ベクトルを形成することとを含み得る。二次元の画素の配列は、第１の方向における複数の行と、第２の方向における複数の列とを含み得る。アラビア語の文字のラインは、実質的に上記第１の方向に沿って並び得る。複数のライン画像は、上記第１の方向に沿って連続的に並び得る。複数のライン画像の中の少なくとも１つは、第１の方向におけるＭ個の行によって定義される高さと、第２の方向におけるＮ個の列によって規定される幅とを有

し得る。ＭおよびＮは、整数である。二次元の画素の配列は、Ｎ行の画素を含み得る。Ｎは、２とおよそ１００との間の範囲にあり得る。Ｎは、３とおよそ１０との間の範囲にあり得る。二次元の画素の配列における画素値は、単一のビットの２進数で表現され得る。二次元の画素の配列における画素値は、マルチビットの２進数で表現され得る。隠れマルコフモデルは、隠れマルコフモデルツールキットとして実装され得る。

【００１１】

本願において記述されるシステムおよび方法は、アラビア語テキストにおける特徴の抽出のための、包括的な、定量的な、かつ正確な技術を提供する。開示されたアラビア語の文字の認識は、いくつかの従来の技術よりも、より効率的であり、かつ計算時間が短い。開示されたシステムおよび方法は、さらにいくつかの従来の技術よりも、より単純かつ私用しやすい。

【００１２】

発明は複数の実施例を参照することにより具体的に示され記述されているが、形式上の種々の変更や詳細は、発明の精神および範囲を離れることなくなされ得ることが、当業者によって、理解されるであろう。

【００１３】

図面の簡単な説明

以下の図面は、出願書類に組込まれかつその一部を形成し、本願発明の実施例を説明し、かつ、明細書とともに、発明の本質を説明するために供される。

【図面の簡単な説明】

【００１４】

【図１】本開示におけるアラビア語のテキスト認識の工程を説明するためのフロー図である。

【図２】アラビア語テキストを含むテキスト画像を説明する図である。

【図３Ａ】テキスト画像を、各々が複数の画素を含む複数のライン画像へと分割することを説明する図である。

【図３Ｂ】図３Ａに示されたライン画像の一部分における、画素および画素値を説明する図である。

【図３Ｃ】図３Ａに示されたライン画像の一部分における、画素および画素値を説明する図である。

【図４】本願に従ったテキスト特徴抽出の方法を説明する図である。

【図５】図４に示されたテキスト特徴抽出の工程を説明するフロー図である。

【図６】本願に従ったテキスト特徴抽出の他の方法を説明する図である。

【図７Ａ】本開示に従った他のテキスト特徴抽出方法を説明する図である。

【図７Ｂ】本開示に従った他のテキスト特徴抽出方法を説明する図である。

【図７Ｃ】本開示に従った他のテキスト特徴抽出方法を説明する図である。

【図７Ｄ】本開示に従った他のテキスト特徴抽出方法を説明する図である。

【図８】図７Ａ～図７Ｄに示されたテキスト特徴抽出の工程を説明するフロー図である。

【発明を実施するための形態】

【００１５】

発明の詳細な説明

図１は、本発明に従ったアラビア語のテキスト認識の概略的な流れを説明する。図１～図３Ｃを参照して、アラビア語のテキスト文書から、テキスト画像２００が取得される（図１のステップ１１０）。テキスト画像２００におけるアラビア語テキストは、複数のテキストライン２１１－２１４に配置され得、その各々は、筆記体のアラビア語の文字のストリングを含む。テキストライン２１１－２１４は、複数のライン画像３１１－３１３へと分割される（図１のステップ１２０）。ライン画像３１１，３１２，または３１３は、それから、各々が画素値を割り当てられた画素３２１－３２３へと分割される（図１のステップ１３０）。ライン画像３１１，３１２，または３１３の幅は、２画素と１００画素との間の範囲にあり得、または、３画素と１０画素との間の範囲にあり得る。ライン画像

３１１，３１２，または３１３は、完全な文字、部分的な文字、または結合した文字を含み得る。

【００１６】

画素値は、特定の画素の位置でのテキスト画像２００の明度値を表わす。ある実装では、明度値が高いことは、白色背景に位置し得る画素における明るい画像の色（または、低密度）を表す。明度値が低いことは、一筆のアラビア語の文字（a stroke of an Arabic character）内に位置し得る暗い画像の色（または、高密度）を表わす。画素値は、２進数、１０進数、および１６進数のような、異なる計数法で表現されてもよい。

【００１７】

図３Ａ～図３Ｃを参照して、ライン画像３１１は、複数の画素３２１－３２３を含む画像部分３２０を含む。画素３２１－３２３の各々は、２進数の画素値「０」または「１」を割り当てられている。画素値「１」は、白色の背景を表わす。画素値「０」は、一筆のアラビア語の文字内にある、暗画像色（つまり、低い明度値）を表わす。開示されたシステムおよび方法は、２進数で表わされたマルチビットの画素値にも適合可能であり、当該２進数で表わされたマルチビットの画素値は、多段階のトーンレベル（たとえば、グレースケール）で、画像濃度を表わし得る。

【００１８】

本開示に従うと、テキスト特徴ベクトルは、テキストライン２１１またはライン画像３１１－３１３から抽出され得る（図１のステップ１４０）。テキスト特徴抽出のさまざまな実装の詳細については、以下に、図４～図８に関連付けられて、議論される。テキスト特徴ベクトルの厳密な形態は、以下に記載されるように、抽出方法によって変化し得る。

【００１９】

ステップ１４０において取得された特徴ベクトルは、次に、隠れマルコフモデル（ＨＭＭ）に送られる（図１のステップ１５０）。本開示では、ＨＭＭは、隠れマルコフモデルツールキット（ＨＴＫ）によって実装される場合があり、それは、隠れマルコフモデルを構築し操作するための移植可能なツールキットである。ＨＴＫは、語彙集がなく、学習用サンプル文字からのモデルおよび文法に依存する。ＨＭＭは、確率解釈を提供し、特徴ベクトルにおいて見い出されたパターンにおける変化を許容し得る。ＨＴＫの機能性の大部分は、Ｃソースコードで利用可能なライブラリモジュールに組込まれ得る。これらのモジュールは、従来のコマンドライン形式のインターフェイスで動作するように設計されているため、ＨＴＫツールの実行を制御するためのスクリプトの記述がシンプルになる。

【００２０】

ＨＭＭは、既知のアラビア語の単語を含むテキスト画像から取得された特徴ベクトルを用いることによって、学習させることができる（データ転記）（図１のステップ１６０）。ＨＴＫは、学習用サンプルのための文字モデルとグランドツルース（ground truth）とともに提供される。文字のモデル化のためのコンポーネントは、特徴ベクトルとそれに対応するグランドツルースとを利用し、文字モデルを評価する。学習用サンプルによって生成された観察結果は、モデルパラメータを調整するのに用いられるが、テスト用のサンプルによって生成された観察結果は、システムの性能を調査するのに利用される。モデルの各状態は、アルファベットの組における字を表わし、各特徴ベクトルは、１つの観察結果に相当する。ＨＴＫ学習ツールは、準備された学習用データを利用して文字モデルパラメータを調整し、既知のデータ転記を予測することができる。

【００２１】

ＨＭＭパラメータは、学習用画像セグメントのためのグランドツルースから推定された。このセグメント化は輪郭にも適用されて、セグメント化のポイントを発見し、これらのセグメントから特徴を抽出し、そして、特徴ベクトルを観察シーケンスに伝達し得る。セグメント化を基礎とした技術は、単語の画像と文字列とを一致させるためのダイナミックプログラミングに利用される。学習段階では、テキスト画像に相当するテキストであるグランドツルースと一体となった、走査されたテキストのラインが、入力として取得される。そして、各ラインは、狭い縦割りの窓へと分割され、そこから特徴ベクトルが抽出され

る。
【００２２】
学習したＨＭＭは、辞書および言語モデルを利用して、特徴ベクトルにおけるアラビア語テキストを認識するために用いられる（図１のステップ１７０）。認識段階は、最も高い尤度の文字シーケンスを見つけるための学習段階において推定された異なる知識源とともに用いられる特徴ベクトルを抽出するのと同じ工程に引き続く。認識ツールは、あるモデルから他のモデルへの遷移確率を記述するために、ネットワークを必要とする。辞書および言語モデルが当該ツールに入力され、認識装置が正しい状態シーケンスを出力するのに役立つことができる。
【００２３】
いくつかの実施形態では、図３Ａ～図５を参照して、ライン画像３１１－３１３は、各々が画素値によって特徴付けられる画素３２１－３２３の配列へとデジタル化される（図５のステップ５１０）。ライン画像３１１は、図４に示されるように、複数のセル４１０－４６０へと分割される（図５のステップ５２０）。セル４１０－４６０の各々は、３×３画素の配列のような、隣接する画素のグループを含む。たとえば、セル４２０は、画素４２２，４２３および他の画素を含む。
【００２４】
次に、各々のセルの画素値が、２進数のセル番号で表わされる（図５のステップ５３０）。各セルにおける画素値は、まず、シリアル化される。たとえば、セル４２０における９つの画素３２２－３２３は、連続する３行の順に、次のようにシリアル化される：１，１，１，１，０，０，１，０，０。一連の２進数の画素値は、その後、９ビットの２進数のセル番号へとマップされる。画素３２２の画素値は、最上位ビットにマップされ、画素３２３の画素値は、最下位ビットにマップされる。結果として、セル４２０における画素値は、２進数で表わされる９ビットのセル番号１１１１００１００で表わされる。同様に、セル４１０－４６０における画素値が、それぞれが０と５１１との間の範囲にある、２進数で表わされるセル番号４８０へと変換される。
【００２５】
ライン画像３１１のセルにおける、２進数のセル番号は、次に、１０進数のセル番号４９０へと変換される（図５のステップ５４０）。１０進数のセル番号４９０は、その後、ライン画像３１１のための特徴ベクトルを形成するためにシリアル化される（図５のステップ５５０）。ステップ５２０－５５０は、別のライン画像のために繰返される。別のライン画像３１１－３１３からの特徴ベクトルは、その後、隠れマルコフモデルへと送られ、テキストラインにおけるアラビア語の文字を認識する（図５のステップ５６０）。
【００２６】
図４～図５と併せて記述された上記の抽出方法は、図１において説明された処理のためのテキスト特徴抽出の実装を表す。上記のテキスト特徴抽出方法は、データストリングにおけるマルチビットの画素値および他の数値表現に適合することが理解されるべきである。たとえば、画素値は、テキスト画像におけるグレースケール情報（または、マルチトーン）を取り込むことのできる、３ビットまたは５ビットの２進数によって表わされ得る。マルチビットの画素値は、ストロークのエッジに沿ったテキスト特徴の記述の精度を改善し得る。
【００２７】
さらに、２進数の代わりに、画素値は、最小値と最大値との間のいかなる数値範囲によっても表わされ得る。いくつかの実装においては、画素値は、［０，１］または［－１，１］のような、所定の範囲に比例した（または、正規化された）値となり得る。そして、画素値は、量子化され得る。特徴ベクトルは、ステップ５３０－５５０と同様に取得され得る。
【００２８】
いくつかの実施形態では、図６を参照して、ライン画像６１０は、分解能において縮小され（つまり、小型化され）、これにより、小型化されたライン画像６２０が形成される

。たとえば、ライン画像６１０は、６０画素の高さを有し得る。小型化されたライン画像６２０は、１／３倍の寸法で、２０画素の高さを有し得る。小型化されたライン画像６２０は、各々が画素値によって表わされる画素の配列６３０を形成するために、デジタル化される。配列６３０における各列の画素値は、２進数を形成するために、シリアル化される。異なる列からの２進数は特徴ベクトルを形成するデータストリング６４０を形成する。テキストラインのライン画像から取得された特徴ベクトルは、隠れマルコフモデルへ送られ、これにより当該テキストラインにおけるアラビア語の文字を認識することができる（図５のステップ５６０）。

【００２９】

図７Ａ，図７Ｂ，および図８を参照して、ライン画像７００は、ステップ５１０（図５）と同様に、画素の配列へとデジタル化される（図８のステップ８１０）。画素は、複数の列に配置される。画素値は、値「１」または値「０」を有する、単一のビットの２進数によって表わされる。各列の画素値がシリアル化されることにより、単一のビットの２進数の列が形成される（図８のステップ８３０）。

【００３０】

次に、図７Ｃおよび図７Ｄに示されるように、値「０」および値「１」の、同じ２進数の画素値を有する連続した画素の頻度が、計算される（図８のステップ８４０）。当該頻度は、足切遷移番号（cut off transition number）まで、カウントされる。当該頻度を表形式化して、頻度カウント７５０および７６０を形成する（図８のステップ８５０）。コンプリメンタリ画素値、たとえば、

【００３１】

【数２】

0	1
0	1
0	1
1	0
1	0
0	1
0	1
1	0

【００３２】

以外同じ数の遷移を有する２つの画素の列を区別するために、列の最上部の画素から値「１」の数のカウントを開始することによって、頻度カウントが実行される。左側の列では、初めは、画素値「１」のカウントは「０」であり、「３」カウントの画素値「０」が続く。当該２つの列におけるコンプリメンタリ画素値は、結果として、次の頻度カウントのようになる：

【００３３】

【数３】

0	3
3	2
2	2
2	1
1	0
0	0

【００３４】

各列の初めにおける、当初の画素カウントが、本発明の精神から逸脱することなく、画素値「０」について行なうこともできることが、理解されるべきである。

【００３５】

表形式の頻度カウント７５０，７６０（図７Ｃ，図７Ｄ）における各行は、白色の背景（画素値「１」を有する）から暗テキスト領域（画素値「０」を有する）への、またはその逆の、画素値における遷移を表わしている。データを圧縮するために、頻度カウントが、最大遷移番号で切り捨てられている。

【００３６】

表形式の頻度カウント７５０，７６０の各列における頻度カウントは、特徴ベクトルを形成している（図８のステップ８６０）。したがって、本実施の形態では、各列は、ベクトルと称することもできる。ライン画像におけるさまざまな列からの特徴ベクトルが、隠れマルコフモデルへ送られる（図８のステップ８７０）。

【００３７】

最大遷移番号は、アラビア語テキストの大標本についての統計的解析によって決定される。表１に示されるように、およそ９９．３１％の列が、６以下の遷移を有している。換言すれば、テキスト画像の大多数が、足切遷移番号として６を選択することにより適切に特徴付けられ得る。

【００３８】

【表１】

表 1:コーパスにおける遷移カウント

列内の遷移番号	列のカウント	パーセント
0	3003663	18.44%
1	95418	0.59%
2	7694625	47.24%
3	74196	0.46%
4	4231776	25.98%
5	45013	0.28%
6	1028765	6.32%
<= 6		**99.31%**
7	7403	0.04%
8	94771	0.57%
9	900	0.01%
10	9543	0.05%
12	1367	0.01%
12 を超える遷移		0.01%

【００３９】

ＨＭＭをベースとしたシステムを構築するときには、このシステムの学習および検査において用いられる特徴ベクトルのタイプが最初に規定される。特徴ベクトルは、継続タイプと分離タイプとに分類されることができる。継続タイプの特徴ベクトルを利用するシステムでは、上記モデルに送られる係数の配列が、またある場合はマトリクスが、利用される。分離タイプの特徴ベクトルが利用されるシステムでは、単一の係数が、上記モデルに送られる。ベクトル量子化手段が、継続タイプのベクトルを分離タイプのベクトルに変換し、これは、ＨＴＫに伴うHQuantツールとHCopyツールとが用いられることによってなされる。HQuantは、後に分離タイプのベクトルを生成するHCopyとともに用いられる学習用データからコードブックを構築するために用いられる。コードブックの構築は、システムのサイズに応じて当該システムの性能に影響を及ぼし、また、その構築に利用されたデータの量に影響を受ける。HQuantは、コードブックの構築に、線形ベクトル量子化アルゴリ

ズム（Liner Vector Quantization Algorithm）を利用し、これは、計算するのには計算コストが高いアルゴリズムである。本開示では、ユニークベクトル量子化（Unique Vector Quantization（ＵＶＱ））という名前の新しい方法が導入され、これにより、演算時間が削減され、そして、システムの性能が改善される。この方法は、特徴ベクトルの繰返しを削除することによって、線形ベクトル量子化アルゴリズム（Liner Vector Quantization Algorithm）を利用するコードブックの構築に利用される特徴ベクトルの数を減らすことおよび、各特徴ベクトルのたった一つのコピーを保持するために用いられる特徴ベクトルの数を減らすことに焦点を当てている。表２に示されるように、コーパス内の特徴ベクトルの数は、大幅に削減されている。

【００４０】

【表２】

表 2: コーパス内のユニークベクトルカウント

コーパス ラインカウント	ベクトルカウント	ユニークベクトル カウント	縮小率
10,000 ライン	12,285,426	413,410	96.64%
15,000 ライン	16,288,252	591,673	96.37%

【００４１】

２０００個の異なるライン画像の特徴ベクトルのすべてを用いてコードブックを構築しようとしたとき、このコードブックについて構築できる最大のサイズが７２８であることを発見した。ユニーク特徴ベクトルのみから１０２４サイズのコードブックを構築するのに１時間３０分を要したのに対し、このコードブックの構築にはおよそ９時間を要した。モノラルモデル（mono models）を用いたこれらの実験からの認識速度は、表３に示される。ユニークな特徴ベクトルが線形ベクトル量子化アルゴリズムとともに用いられると、コードブックのサイズは増大する。計算速度は６倍に上昇し、認識速度は上昇した。

【００４２】

【表３】

表 3: ユニークベクトルカウントの認識率

コードブックタイプ	コードブック サイズ	構築時間	認証率
UVQ 利用無し	728	9 時間	83.59%
UVQ 利用	1024	1 時間 30 分	85.22%

【００４３】

上述の方法は、言及された特定の例に限定されるものではないことが、理解されるべきである。設定は、発明の精神から逸脱することなく変更され得る。たとえば、足切遷移番号は、６以外にも選択され得る。ライン画像の高さおよび幅は、当該ライン画像内のセルのサイズと同様に、上述の例とは異なるものにされ得る。テキスト特徴ベクトルの形態は、抽出方法に応じて変更され得る。たとえば、特徴ベクトルは、２進数、１０進数、または他の記数法で記述された数値の形態を取り得る。

【００４４】

今回開示された実施の形態およびその変形例はすべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は上記した説明ではなくて特許請求の範囲によって示され、特許請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。実施の形態およびその変形例において開示された技術は、可能な限り単独でも組み合わせても実施され得ることが意図される。

【図１】

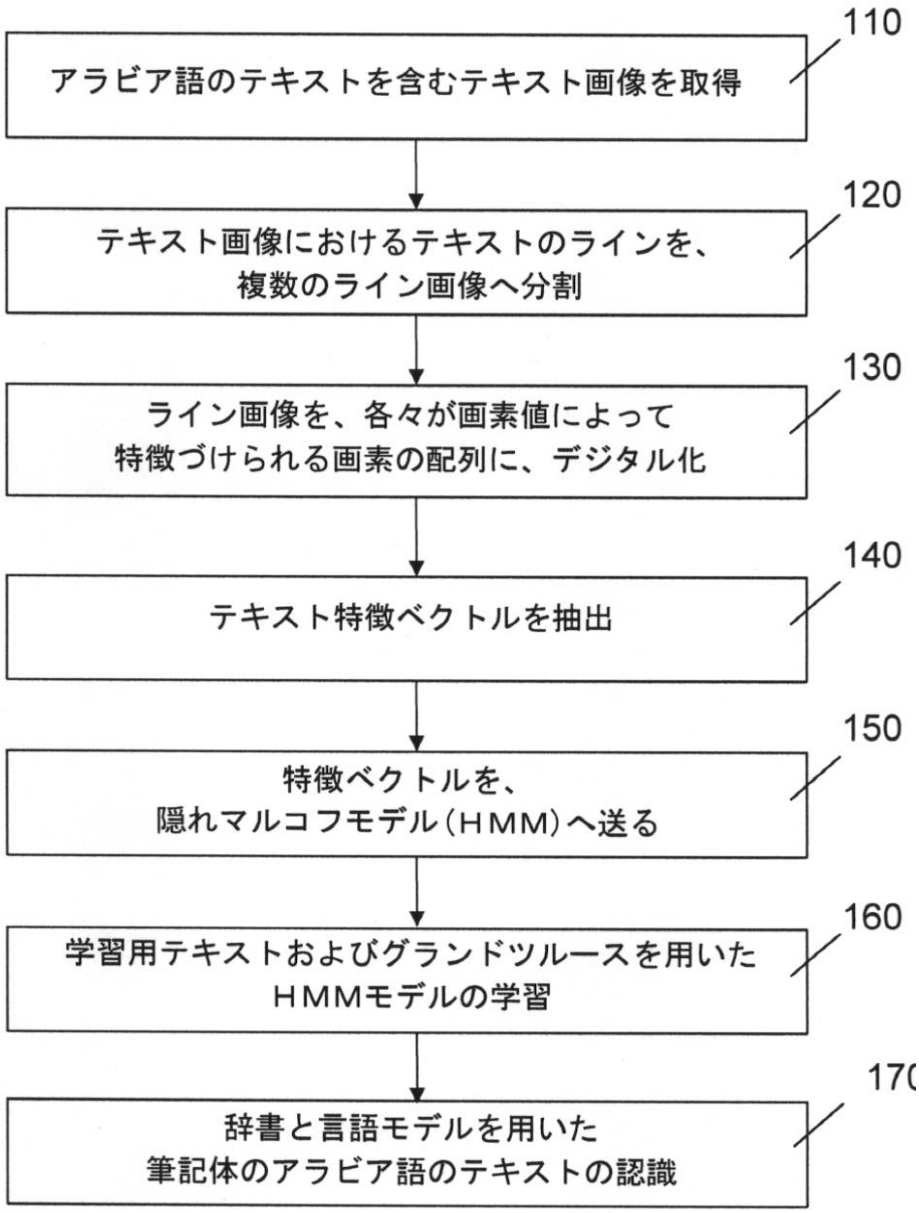
110
アラビア語のテキストを含むテキスト画像を取得
120
テキスト画像におけるテキストのラインを、
複数のライン画像へ分割
130
ライン画像を、各々が画素値によって
特徴づけられる画素の配列に、デジタル化
140
テキスト特徴ベクトルを抽出
150
特徴ベクトルを、
隠れマルコフモデル（ＨＭＭ）へ送る
160
学習用テキストおよびグランドツルースを用いた
ＨＭＭモデルの学習
170
辞書と言語モデルを用いた
筆記体のアラビア語のテキストの認識

Figure 1

【図２】

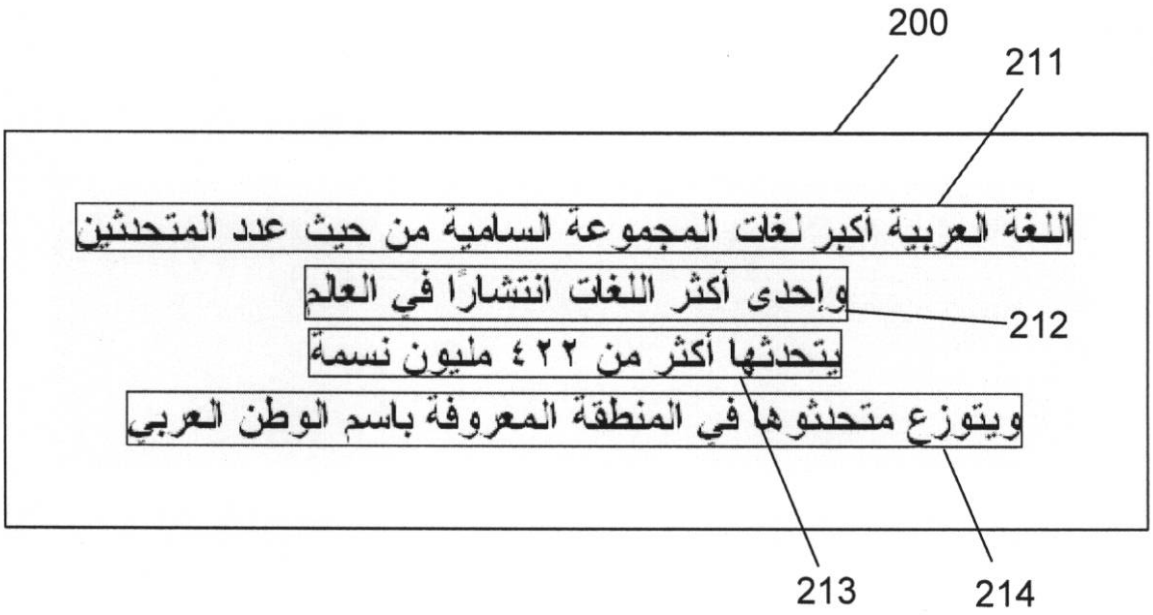

Figure 2

【図３Ａ】

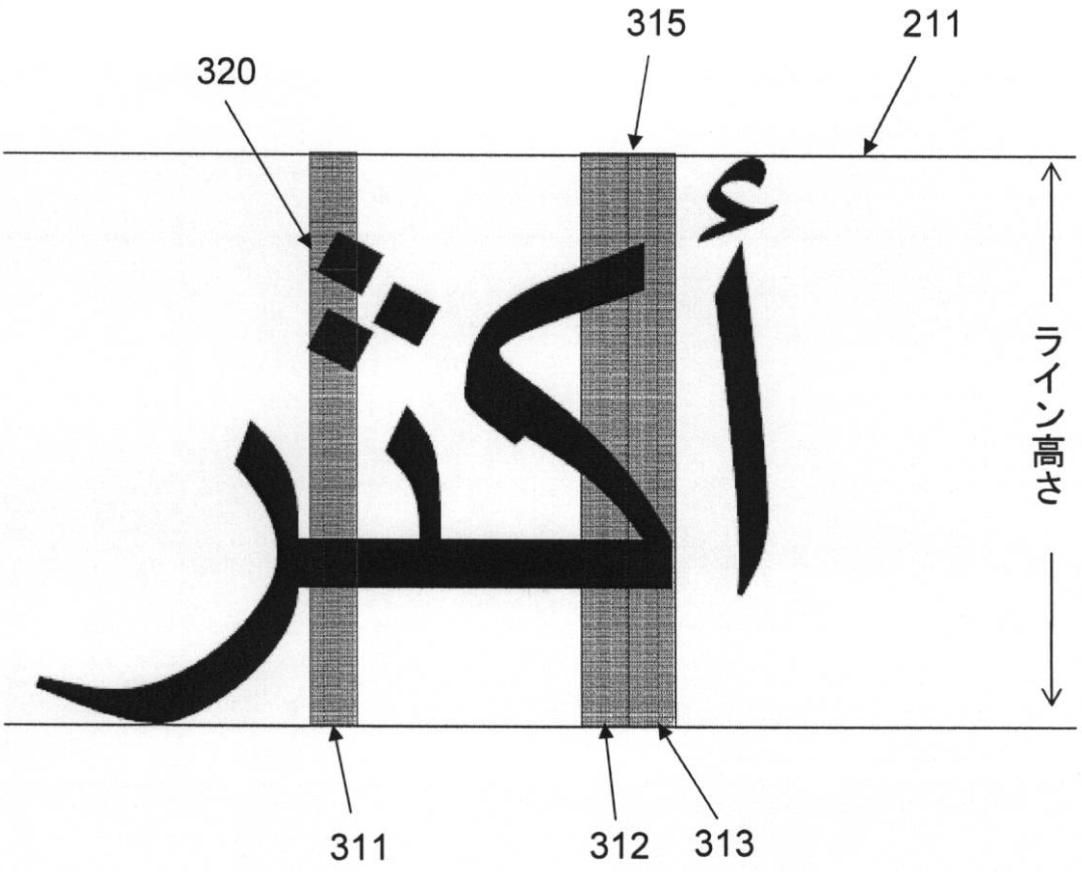

Figure 3A

【図３Ｂ】

Figure 3B

【図３Ｃ】

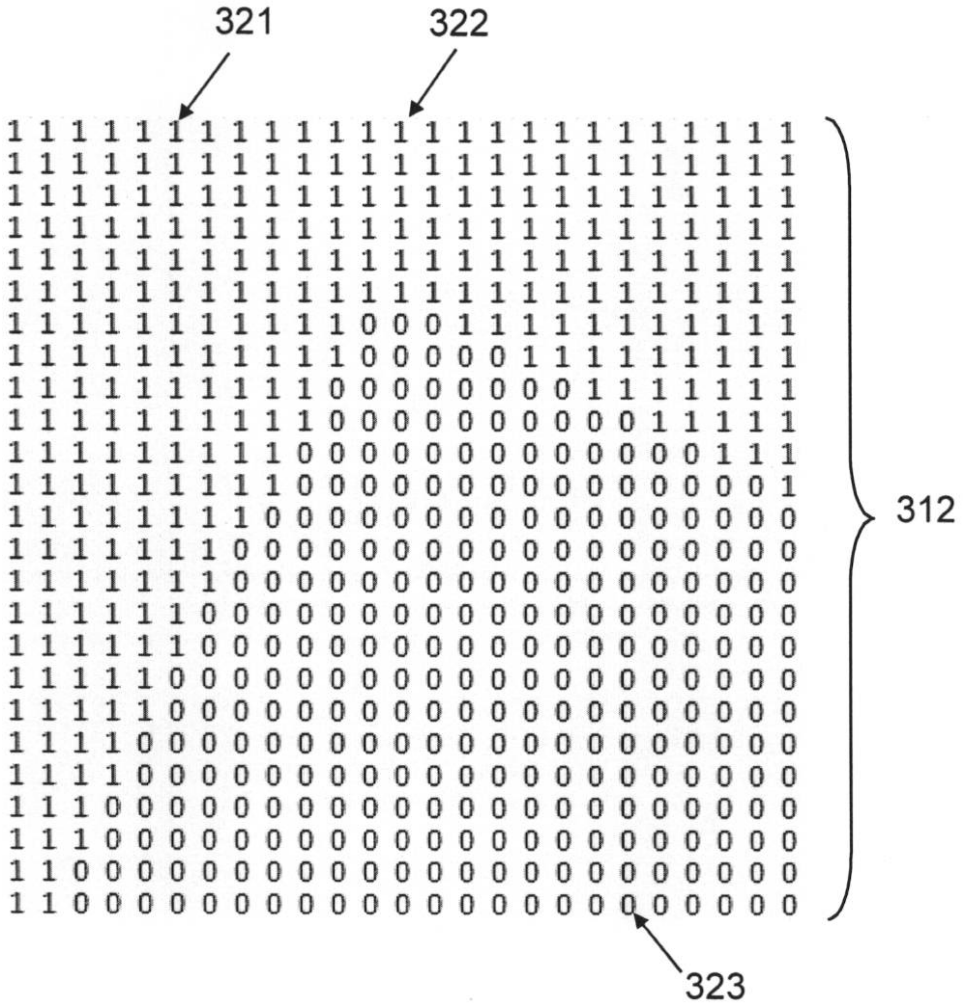

Figure 3C

【図４】

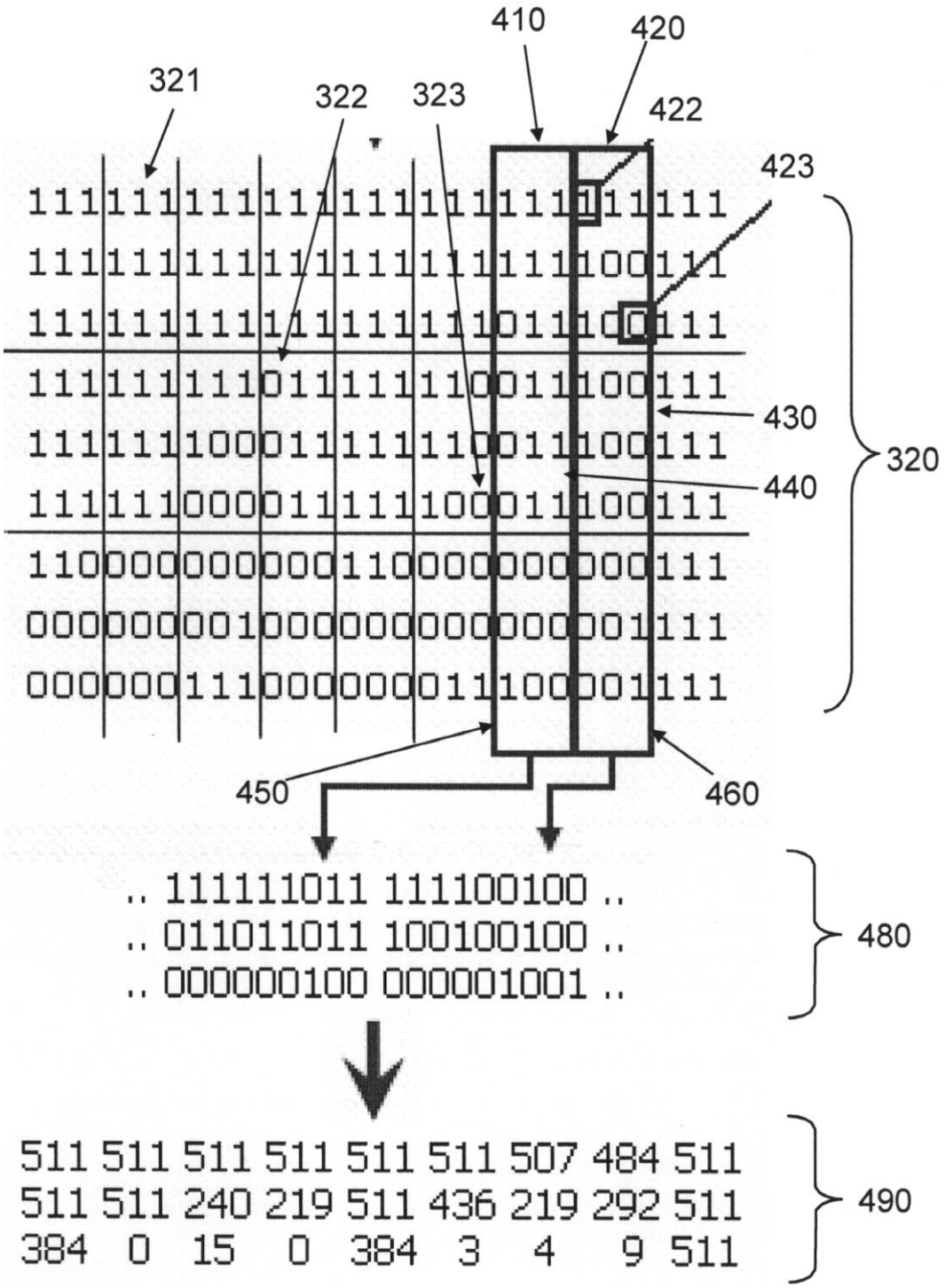
410
420
321
322
323
422
423
430
440
320
450
460
.. 111111011 111100100 ..
.. 011011011 100100100 ..
.. 000000100 000001001 ..
480
511 511 511 511 511 511 507 484 511
511 511 240 219 511 436 219 292 511
384 0 15 0 384 3 4 9 511
490

Figure 4

【図５】

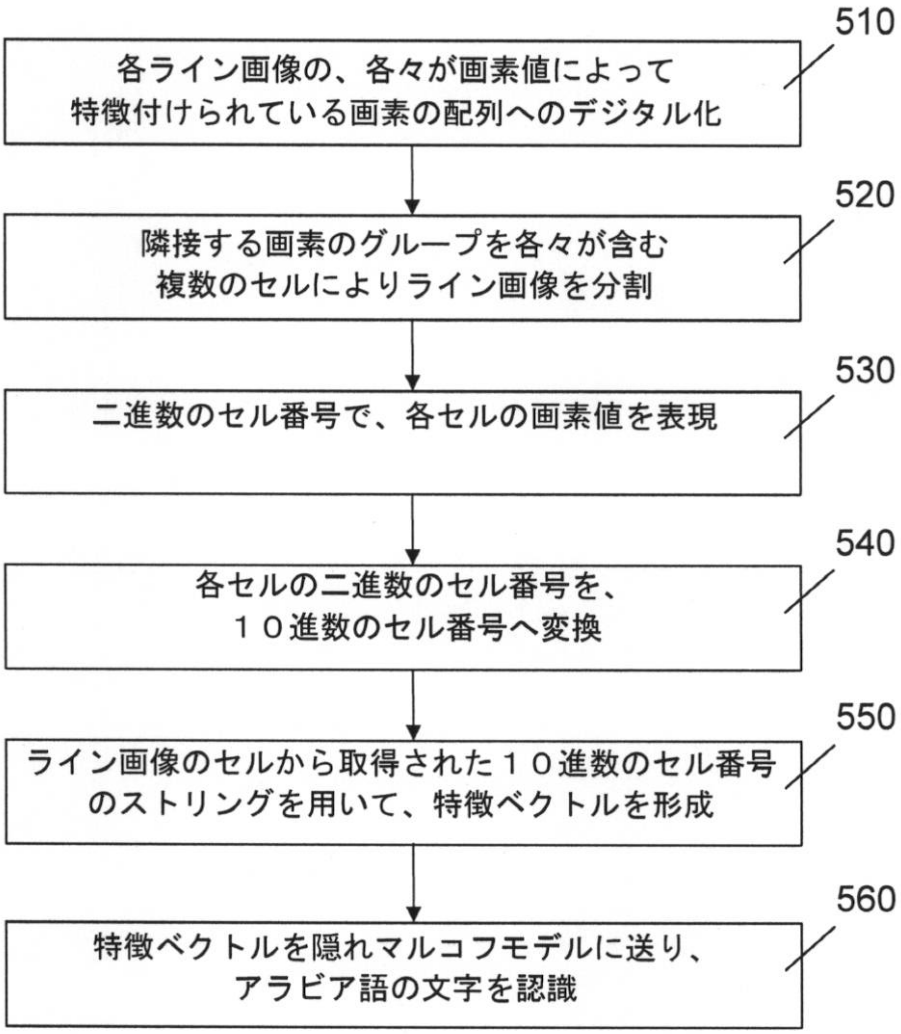
510
各ライン画像の、各々が画素値によって
特徴付けられている画素の配列へのデジタル化
520
隣接する画素のグループを各々が含む
複数のセルによりライン画像を分割
530
二進数のセル番号で、各セルの画素値を表現
540
各セルの二進数のセル番号を、
１０進数のセル番号へ変換
550
ライン画像のセルから取得された１０進数のセル番号
のストリングを用いて、特徴ベクトルを形成
560
特徴ベクトルを隠れマルコフモデルに送り、
アラビア語の文字を認識

Figure 5

【図６】

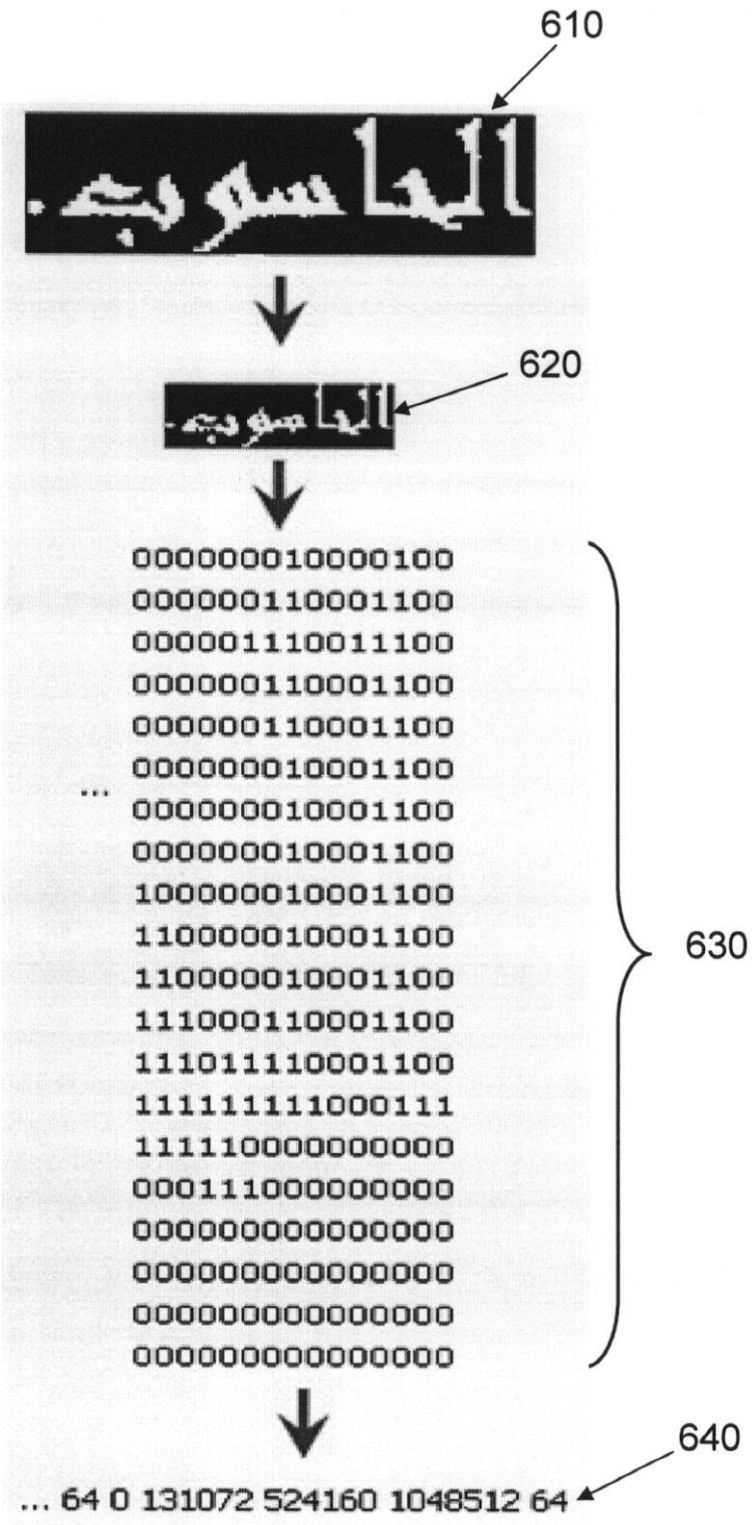
610
620
0000000010000100
0000001100011100
0000011100111100
0000001100011100
0000001100011100
0000000100011100
...
0000000100011100
0000000100011100
1000000100011100
1100000100011100
1100000100011100
1110001100011100
1110111100011100
1111111111000111
1111100000000000
0001110000000000
0000000000000000
0000000000000000
0000000000000000
0000000000000000
630
640
... 64 0 131072 524160 1048512 64

Figure 6

【図７Ａ】

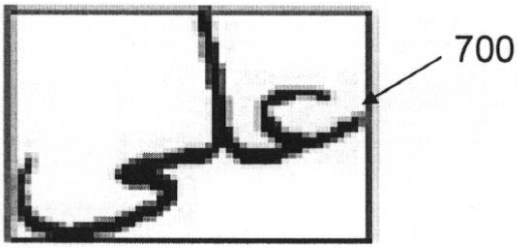

Figure 7A

【図７Ｂ】

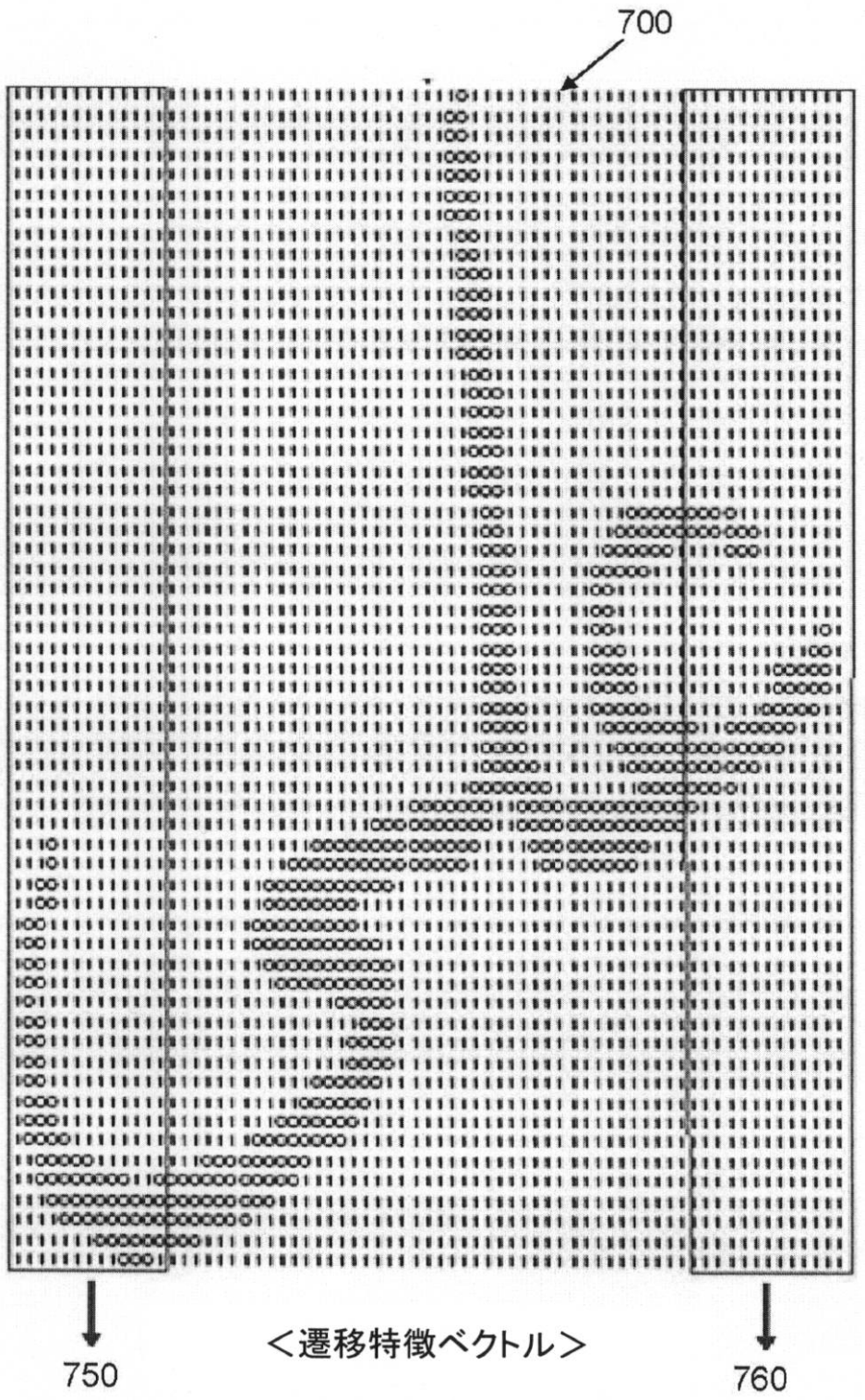

Figure 7B

【図７Ｃ】

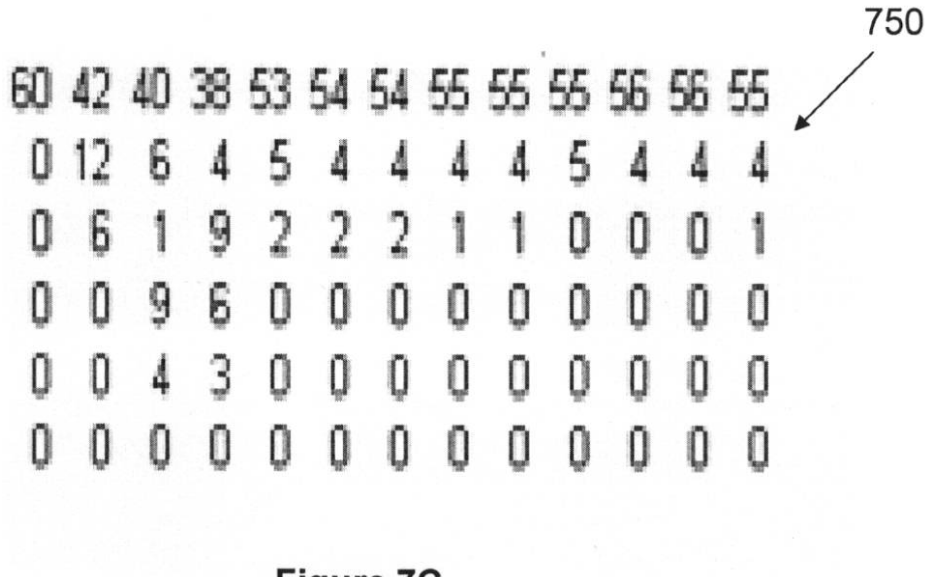

Figure 7C

【図７Ｄ】

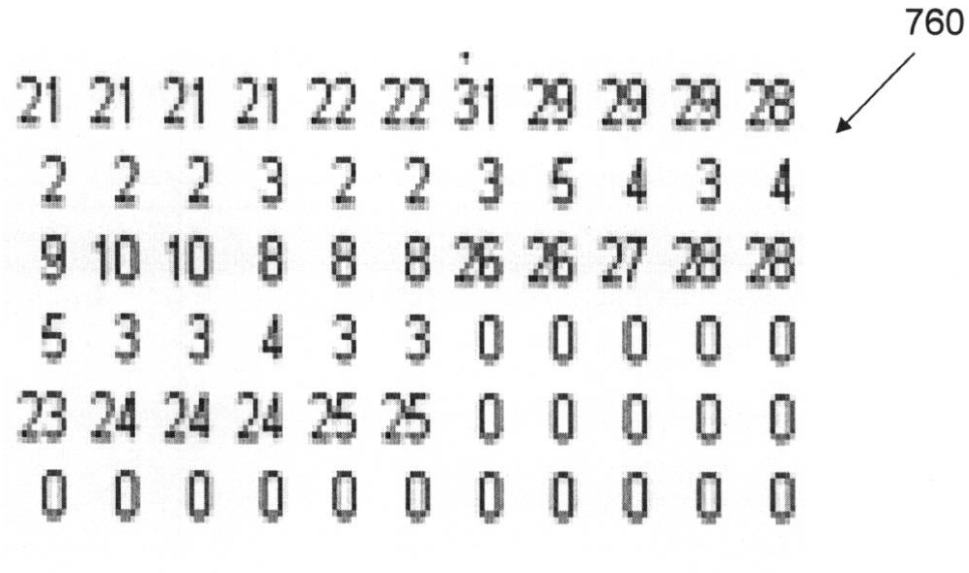

Figure 7D

【図８】

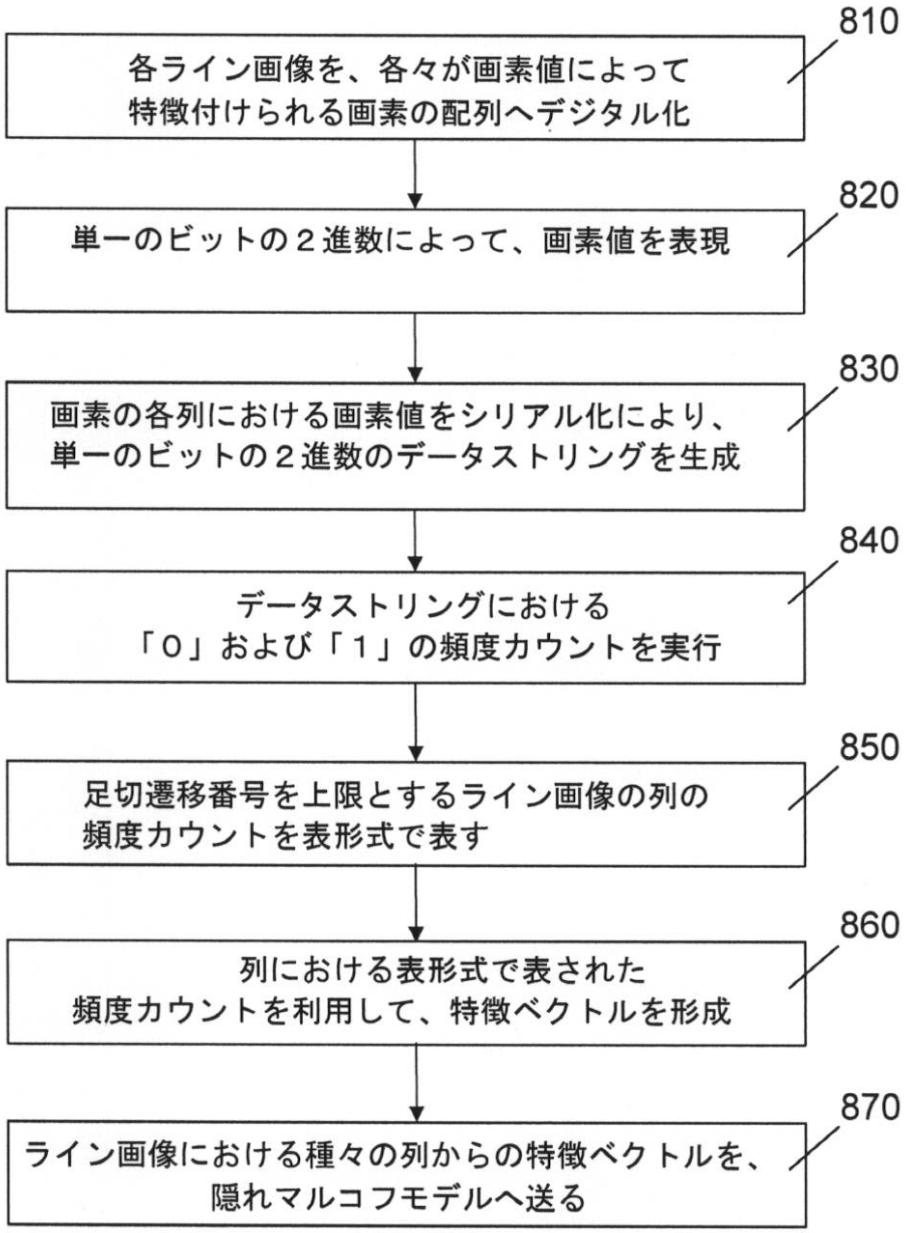
各ライン画像を、各々が画素値によって
特徴付けられる画素の配列へデジタル化
810
単一のビットの２進数によって、画素値を表現
820
画素の各列における画素値をシリアル化により、
単一のビットの２進数のデータストリングを生成
830
データストリングにおける
「０」および「１」の頻度カウントを実行
840
足切遷移番号を上限とするライン画像の列の
頻度カウントを表形式で表す
850
列における表形式で表された
頻度カウントを利用して、特徴ベクトルを形成
860
ライン画像における種々の列からの特徴ベクトルを、
隠れマルコフモデルへ送る
870

Figure 8

フロントページの続き

(72)発明者　モハメド・エス・ホルシード
サウジアラビア、１１４４２　リヤド、ピィ・オゥ・ボックス・６０８６、キング・アブドゥルアジズ・シティ・フォー・サイエンス・アンド・テクノロジー
(72)発明者　フセイン・ケィ・アルーオマリ
サウジアラビア、１１４４２　リヤド、ピィ・オゥ・ボックス・６０８６、キング・アブドゥルアジズ・シティ・フォー・サイエンス・アンド・テクノロジー

Ｆターム(参考) 5B064 AB03 AB19 DA27 DC24 EA07
5L096 BA17 DA02 JA28 KA04