(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6082306号
(P6082306)
(45)発行日　平成29年2月15日(2017.2.15)　(24)登録日　平成29年1月27日(2017.1.27)

(51)Int.Cl.　F I
G06K 9/20 (2006.01)　G06K 9/20 340K
G06K 9/20 340L

請求項の数 25 (全 20 頁)

(21)出願番号　特願2013-84694(P2013-84694)
(22)出願日　平成25年4月15日(2013.4.15)
(65)公開番号　特開2013-235574(P2013-235574A)
(43)公開日　平成25年11月21日(2013.11.21)
審査請求日　平成28年2月16日(2016.2.16)
(31)優先権主張番号　13/467,873
(32)優先日　平成24年5月9日(2012.5.9)
(33)優先権主張国　米国(US)

早期審査対象出願

(73)特許権者　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】光学式文字認識用に画像を前処理するための方法およびシステム

(57)【特許請求の範囲】
【請求項１】
光学式文字認識（ＯＣＲ）用に画像を前処理する方法であって、画像は複数の欄を含み、複数の欄のうちの各欄は文字を含み、前記方法は、
文字に関連付けられる複数の構成要素を定めることを備え、構成要素は一連の接続されたピクセルを含み、前記方法は、さらに、
前記複数の構成要素に関連付けられる行高さおよび欄間を計算することと、
行高さおよび欄間の少なくとも一方に基づき、前記複数の構成要素のうちの少なくとも１つの構成要素を、前記複数の欄のうちのある欄に関連付けることと、
前記複数の欄のうちの各欄について第１の組の特性パラメータを計算することと、
前記第１の組の特性パラメータに基づき、前記複数の欄のうちの各欄の中の複数の構成要素を合成して、少なくとも１つの副単語および少なくとも１つの単語の少なくとも一方を形成することとを備え、前記複数の欄は、第１の領域と第２の領域とを含み、
前記方法は、
最も近い文字構成要素が第１の領域に入る構成要素を当該第１の領域に合成することと、
最も近い文字構成要素が第２の領域に入る構成要素を当該第２の領域に合成することとを備え、前記第２の領域の少なくとも部分は前記第１の領域の少なくとも部分と縦方向に重なり、
各欄に関連付けられる語間を計算することをさらに備え、前記語間を計算することは、

各欄に関連付けられた複数の構成要素の連続する構成要素間の間隔のヒストグラムを作成することと、ヒストグラムから頻出間隔を特定することとを含み、前記頻出間隔は行高さによって定められるしきい値範囲内にあり、さらに、前記頻出間隔に基づき語間を計算することを含み、

各欄に関連付けられる行間を計算することをさらに備え、前記行間を計算することは、

各欄に関連付けられる複数の構成要素の複数の水平射影のヒストグラムを作成することを含み、前記複数の水平射影のうちのある水平射影は、ラスタスキャンの各掃引に対応して複数の構成要素に関連付けられるピクセルの数を示し、さらに、２つの連続する最大水平射影間の平均距離を計算することと、前記平均距離に基づき行間を計算することとを含む、方法。

【請求項２】

前記画像は、濃淡画像およびカラー画像の少なくとも一方を２進画像に変換することによって得られる、請求項１に記載の方法。

【請求項３】

前記画像は、ごま塩雑音をふるい落とすことによって得られる、請求項１または請求項２に記載の方法。

【請求項４】

前記画像は、変形ハフ変換を用いて歪みを修正することによって得られ、前記変形ハフ変換はアラビア文字に適合される、請求項１～請求項３のいずれか１項に記載の方法。

【請求項５】

前記複数の構成要素を定めることは、

前記画像に対してラスタスキャンを行なうこと、

前記ラスタスキャンの少なくとも１回の掃引に対応する複数の構成要素のうちの少なくとも１つに関連付けられる複数のピクセルを特定すること、および

複数のピクセル間の相互接続に基づき前記複数のピクセルを統合して、少なくとも１組の接続されたピクセルを形成することを含む、請求項１～請求項４のいずれか１項に記載の方法。

【請求項６】

ピクセルは当該ピクセルの８個の隣接するピクセルの少なくとも１つと相互接続される、請求項５に記載の方法。

【請求項７】

前記行高さを計算することは、

前記複数の構成要素の各々の高さに対応する高さのヒストグラムを作成すること、

高さのヒストグラムから頻出高さを特定すること、および

頻出高さに基づき行高さを計算することを含む、請求項１～請求項６のいずれか１項に記載の方法。

【請求項８】

前記欄間は、行高さに基づき計算される、請求項７に記載の方法。

【請求項９】

各欄に関連付けられる前記語間を計算することは、

各欄に関連付けられた前記複数の構成要素の連続する構成要素間の間隔のヒストグラムを作成することと、

前記ヒストグラムから頻出間隔を特定することとを含み、前記頻出間隔は前記行高さによって定められるしきい値範囲内にあり、さらに

前記頻出間隔に基づき前記語間を計算することを含む、請求項１～請求項８のいずれか１項に記載の方法。

【請求項１０】

前記連続する構成要素は、縦方向に重なる構成要素および所定の距離離れている構成要素の少なくとも一方を含み、前記縦方向に重なる構成要素は、縦軸に沿って少なくとも１つの座標を共有する、請求項９に記載の方法。

【請求項１１】
前記複数の構成要素を合成することは、
前記少なくとも１つの副単語および少なくとも１つの単語の少なくとも一方を形成するために、語間に基づき、各欄に関連付けられる連続する構成要素を結合することと、
前記第１の組の特性パラメータに基づき、アラビア文字に関連付けられる複数の構成要素から、非文字項目に関連付けられる複数の構成要素のうちの少なくとも１つの構成要素をふるい落とすこととを含む、請求項９に記載の方法。
【請求項１２】
ある欄に関連付けられる少なくとも１つの座標に基づき、複数の欄を分類することをさらに備え、前記少なくとも１つの座標は、画像における欄の位置に関連付けられる、請求項１１に記載の方法。
【請求項１３】
前記方法はさらに、各欄に関連付けられる、各副単語および各単語の少なくとも一方に関連付けられる第２の組の特性パラメータを計算することを備え、前記第２の組の特性パラメータは、各副単語および各単語の少なくとも一方に関連付けられる行高さ、各副単語および各単語の少なくとも一方に関連付けられる語間、ならびに各副単語および各単語の少なくとも一方に関連付けられる行間の１つであり、さらに
第２の組の特性パラメータに基づき少なくとも２つの副単語をグループ化して、少なくとも１つの副単語および少なくとも１つの単語の一方を形成することを備える、請求項１～請求項１２のいずれか１項に記載の方法。
【請求項１４】
前記方法は、前記少なくとも１つの副単語および前記少なくとも１つの単語を、各副単語および各単語の少なくとも一方に関連付けられる行高さ、ならびに各副単語および各単語の少なくとも一方に関連付けられる行間の少なくとも一方に基づき、少なくとも１本の横行に分割することをさらに備える、請求項１３に記載の方法。
【請求項１５】
光学式文字認識（ＯＣＲ）用に画像を前処理するためのシステムであって、画像は複数の欄を含み、複数の欄の各欄は、アラビア文字および非文字項目の少なくとも一方を含み、前記システムは、
メモリと、
前記メモリに結合されるプロセッサとを備え、前記プロセッサは、
複数の欄の中のアラビア文字および非文字項目の少なくとも一方に関連付けられる複数の構成要素を定めるように構成され、構成要素は一連の接続されたピクセルを含み、前記プロセッサは、さらに、
前記複数の構成要素に関連付けられる行高さおよび欄間を計算することと、
行高さおよび欄間に基づき、複数の構成要素のうちの少なくとも１つの構成要素を、複数の欄のうちのある欄に関連付けることと、
前記複数の欄のうちの各欄について第１の組の特性パラメータを計算することと、
第１の組の特性パラメータに基づき、複数の欄のうちの各欄の中の複数の構成要素を合成して、少なくとも１つの副単語および少なくとも１つの単語の少なくとも一方を形成することとを行なうように構成され、前記複数の欄は、第１の領域と第２の領域とを含み、
前記プロセッサは、
最も近い構成要素が第１の領域に入る構成要素を当該第１の領域に合成することと、
最も近い構成要素が第２の領域に入る構成要素を当該第２の領域に合成することとを行なうように構成され、前記第２の領域の少なくとも部分は前記第１の領域の少なくとも部分と縦方向に重なり、
前記第１の組の特性パラメータは、各欄に関連付けられる行高さ、各欄に関連付けられる語間、各欄に関連付けられる行間、各構成要素に対応するピクセルの数、各構成要素の幅、各構成要素の高さ、各構成要素の座標、各構成要素の密度、または各構成要素のアスペクト比の１つを含み、

各欄に関連付けられる行間を計算するために、前記プロセッサは、
各欄に関連付けられる前記複数の構成要素の中の複数の水平射影のヒストグラムを作成するように構成され、前記複数の水平射影のうちのある水平射影は、ラスタスキャンの各掃引に対応して、前記複数の構成要素に関連付けられるピクセルの数を示し、さらに前記プロセッサは、
２つの連続する最大水平射影間の平均距離を計算し、
前記平均距離に基づき行間を計算するよう構成されている、システム。
【請求項１６】
前記プロセッサは、
濃淡画像およびカラー画像の少なくとも一方を２進画像に変換すること、
ごま塩雑音をふるい落とすこと、および
変形ハフ変換を用いて歪みを修正すること、のうちの少なくとも一つを行なうようさらに構成されている、請求項１５に記載のシステム。
【請求項１７】
複数の構成要素を定めるために、前記プロセッサは、
画像に対してラスタスキャンを行ない、
ラスタスキャンの少なくとも１回の掃引に対応して前記複数の構成要素の少なくとも１つの構成要素に関連付けられる複数のピクセルを特定し、
複数のピクセル間の相互接続に基づき、前記複数のピクセルを統合して少なくとも１組の接続されたピクセルを形成するようさらに構成されている、請求項１５または請求項１６に記載のシステム。
【請求項１８】
前記行高さを計算するために、前記プロセッサは、
前記複数の構成要素の各々の高さに対応する高さのヒストグラムを作成し、
前記高さのヒストグラムから頻出高さを特定し、
前記頻出高さに基づき行高さを計算するよう構成されている、請求項１５～請求項１７のいずれか１項に記載のシステム。
【請求項１９】
前記プロセッサは、行高さに基づき欄間を計算するようさらに構成されている、請求項１８に記載のシステム。
【請求項２０】
各欄に関連付けられる語間を計算するために、前記プロセッサは、
各欄に関連付けられる前記複数の構成要素のうちの連続する構成要素間の間隔のヒストグラムを作成し、
前記ヒストグラムから頻出間隔を特定するように構成され、前記頻出間隔は行高さによって定められるしきい値範囲内にあり、前記プロセッサは、
前記頻出間隔に基づき語間を計算するようさらに構成されている、請求項１５～請求項１９のいずれか１項に記載のシステム。
【請求項２１】
前記プロセッサは、
語間に基づき各欄に関連付けられる連続する構成要素を結合して、少なくとも副単語および少なくとも１つの単語の少なくとも一方を形成し、
前記第１の組の特性パラメータに基づき、アラビア文字に関連付けられる複数の構成要素から非文字項目に関連付けられる前記複数の構成要素のうちの少なくとも１つの構成要素をふるい落とすようさらに構成されている、請求項２０に記載のシステム。
【請求項２２】
前記プロセッサは、ある欄に関連付けられる少なくとも１つの座標に基づき、複数の欄を分類するようさらに構成されており、前記少なくとも一つの座標は画像における欄の位置に関連付けられる、請求項２１に記載のシステム。
【請求項２３】

前記プロセッサは、
各欄に関連付けられる、各副単語および各単語の少なくとも一方に関連付けられる第２の組の特性パラメータを計算するようにさらに構成され、第２の組の特性パラメータは、各副単語および各単語の少なくとも一方に関連付けられる行高さ、各副単語および各単語の少なくとも一方に関連付けられる語間、ならびに各副単語および各単語の少なくとも一方に関連付けられる行間のうちの１つであり、さらに前記プロセッサは、
第２の組の特性パラメータに基づき少なくとも２つの副単語をグループ化して、少なくとも１つの副単語および少なくとも１つの単語の一方を形成するようさらに構成されている、請求項１５～請求項２２のいずれか１項に記載のシステム。
【請求項２４】
前記プロセッサは、少なくとも１つの副単語および少なくとも１つの単語を、各副単語および各単語の少なくとも一方に関連付けられる行高さ、ならびに各副単語および各単語の少なくとも一方に関連付けられる行間の少なくとも一方に基づき、少なくとも１つの横行に分割するようさらに構成されている、請求項２３に記載のシステム。
【請求項２５】
コンピュータによって実行されるプログラムであって、請求項１～請求項１４のいずれか１項に記載の方法を前記コンピュータに実行させるための、プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願への相互参照
これは、米国特許出願連続番号第１２／８１４，４４８号の継続出願であり、その全体を引用により援用する。
【０００２】
発明の分野
本発明は一般に光学式文字認識（OCR:Optical Character Recognition）用に画像を前処理することに関し、画像はアラビア文字および／または非文字項目を含む。より具体的には、本発明は複数の欄を含む画像の前処理に関し、各欄はアラビア文字および／または非文字項目を含む。
【背景技術】
【０００３】
背景
スキャンされた文書を編集可能および検索可能な文章に自動的に変換するには、正確かつ安定した光学式文字認識（ＯＣＲ）システムを使用する必要がある。英文に対するＯＣＲシステムは、さまざまな理由により、高いレベルの精度に達している。その主な理由の１つは、ＯＣＲシステムへの入力として、英文を分離された文字に前処理する機能にある。英文の各文字は、印刷された英字が繋がっていない性質により、分離することができる。しかし、スキャンされた繋がっている文字はＯＣＲシステムへの課題であり、ピッチが変化している場合にその精度を落とす。
【発明の概要】
【発明が解決しようとする課題】
【０００４】
アラビア語のスキャンされた文章は一連の繋がっている文字を含むので、文字に分割するのが難しい。アラビア文字での別の問題は、文字や後に続く母音の発音を示すために、多くの点やアクセント記号が文字の上下に入っていることである。これにより、英語向けに設計された前処理技術が正確にアラビア文字を処理することに用いられるのを妨げることとなる。
【０００５】
アラビア文字の別の特徴は、アラビア語の文章は母音を示すアクセント記号を付けてもまたは付けなくても書くことができる点にある。さらに、英文は大文字または小文字の表

示を有するが、多くのアラビア語の文字は、その位置が単語の初め、単語の間、もしくは単語の終わりにあるのか、または単独の単語としてあるのかに応じて、３つまたは４つの形を含む。したがって、アクセント記号および単語内での文字の位置により、アラビア文字にはさまざまな組合せが可能であるので、現在のＯＣＲ前処理システムでアラビア文字を前処理することは不的確となる。

【０００６】

さらに、アラビア文字および非文字項目の欄を複数有する画像では、各欄に関連付けられるアラビア文字はフォントのサイズ、スタイル、色などが変わり得る。フォントサイズが変わることにより、隣接する欄は行が揃わず、正確に分割できないかもしれない。

【０００７】

したがって、アラビア文字および／または非文字項目を含む複数の欄を有する画像を前処理するための方法およびシステムが必要である。

【０００８】

添付の図面であって、同じ参照符号は、それぞれの図面において同じまたは機能的に類似した要素を指し、以下の詳細な説明とともに明細書の中に組込まれてその一部をなす図面は、さまざまな実施例を示し、かつ本発明に従うさまざまな原理および利点を説明する役割を果たす。

【図面の簡単な説明】

【０００９】

【図１】本発明のさまざまな実施例に従い、アラビア文字および／または非文字項目を有する複数の欄を含む画像の一例を示す図である。

【図２】本発明の一実施例に従い、画像に関連付けられる歪みを修正する際の画像の一例を示す図である。

【図３】本発明の一実施例に従い、２進画像に変換された画像の一例を示す図である。

【図４】本発明の一実施例に従い、光学式文字認識のために画像を前処理する方法のフロー図である。

【図５】本発明の一実施例に従い、複数の構成要素を定める方法のフロー図である。

【図６】本発明の一実施例に従い、行の高さを計算する方法のフロー図である。

【図７】本発明の一実施例に従い、複数の欄のうちのある欄に関連付けられる１つ以上の構成要素を有する画像を示す図である。

【図８】本発明の一実施例に従い、語間を計算する方法のフロー図である。

【図９】本発明の一実施例に従い、行間を計算する方法のフロー図である。

【図１０】本発明の一実施例に従い、１つ以上の副単語および／または１つ以上の単語を形成するために、複数の構成要素を合成する方法を示す図である。

【図１１】本発明の一実施例に従い、非文字項目が取除かれた画像の一例を示す図である。

【図１２】本発明の一実施例に従い、行高さおよび行間に応じて、１つ以上の単語および１つ以上の副単語を１つ以上の横行に分割する一例を示す図である。

【図１３】本発明のさまざまな実施例に従い、アラビア文字および／または非文字項目を含む複数の欄を含む画像を前処理するためのシステムのブロック図である。

【図１４】発明の一実施例に従い、１つ以上の別の欄または領域に縦方向および／または横方向に重なる欄または領域と関連付けられる複数の構成要素のうち１つ以上の概略図である。

【図１５】発明の一実施例に従い、１つ以上の別の欄または領域に縦方向および／または横方向に重なる欄または領域と関連付けられる複数の構成要素のうち１つ以上の概略図である。

【発明を実施するための形態】

【００１０】

図面の要素は簡潔におよび明瞭にするために示されており、必ずしも尺度通りに描かれていないことは、当業者なら理解するであろう。たとえば、図面の一部の要素の寸法は、

本発明の実施例をわかりやすくするために、他の要素に対して拡大されて示されているかもしれない。

【００１１】

詳細な説明

本発明に従う実施例を詳細に説明する前に、実施例は主に光学式文字認識（ＯＣＲ）用に画像を前処理するための方法およびシステムに関する方法の工程および装置の構成要素の組合せに基づいていることに注意しなければならない。画像は複数の欄を含み、各欄はアラビア文字および／または非文字項目を含む。したがって、装置の構成要素および方法の工程は、図面において適する場合は従来の記号によって示され、ここでの記載が当業者にとって容易に明らかとなる詳細でもって開示を曖昧にしないよう、本発明の実施例の理解に関連する具体的詳細のみが示されている。

【００１２】

本明細書では、第１および第２、上および下などのような相関的用語は、あるエンティティまたは動作を別のエンティティまたは動作と区別するためにのみ用いられており、これらのエンティティまたは動作間において実際にこのような関係または順序を必ずしも必要としないまたは意味しない。「含む」、「有する」またはその他のこのような用語の変形は、限定されない含有を網羅するために意図されており、一連の要素を含むプロセス、方法、物品または装置は、これらの要素のみを含むのではなく、プロセス、方法、物品もしくは装置に明記されていない要素、または固有の他の要素をも含み得る。「…を含む」の用語が付いている要素は、それ以外の制限がなければ、その要素を含むプロセス、方法、物品または装置において付加的同一要素の存在を排除するものではない。

【００１３】

ここに記載される発明の実施例は、ＯＣＲ用に画像を前処理する方法の機能の一部、大部分、またはすべてを実施するために、特定の非トランザクション－クライアント回路と併せて、１つ以上の従来のトランザクション－クライアントと、その１つ以上のトランザクション－クライアントを制御する固有の記憶されているプログラム命令とを含み得ることが理解されるであろう。画像は複数の欄を含み、各欄はアラビア文字および／または非文字項目を含む。非トランザクション－クライアント回路は、無線受信装置、無線送信装置、信号ドライバ、クロック回路、電源回路、およびユーザ入力装置を含むことができるが、これらに限定されない。したがって、これらの機能は、ＯＣＲ用に、アラビア文字および非文字項目を含む画像を前処理する方法の工程として解釈することができる。代替的に、機能の一部またはすべては、プログラム命令が記憶されていない状態マシンによって、または１つ以上の特定用途向け集積回路（ＡＳＩＣ）において実施することができ、各機能または特定の機能の一部の組合せは、カスタム論理として実施される。当然、これら２つのアプローチを組合せて用いることもできる。これらの機能の方法および手段がここに記載される。さらに、当業者なら、ここに開示されている概念および原理により、たとえば利用可能な時間、現行の技術および経済的な点を考慮して動機付けられる著しい努力および多くの設計的選択事項があったとしても、最小限の実験でもってこのようなソフトウェア命令、プログラムおよびＩＣを容易に生成できると考えられる。

【００１４】

一般に、さまざまな実施例に従い、本発明は、ＯＣＲ用に画像を前処理するための方法およびシステムを提供する。画像は複数の欄を含み、各欄はアラビア文字および／または非文字項目を含む。本方法は、複数の欄の中のアラビア文字および／または非文字項目に関連付けられる複数の構成要素を定めることを含む。ここで、構成要素は１組の繋がっているピクセルを含む。つぎに、複数の構成要素に関連付けられる行高さおよび欄間が計算される。その後、複数の構成要素のうちの１つ以上の構成要素は、行の高さおよび／または欄間に基づき、複数の欄のうちのある欄に関連付けられる。さらに、複数の欄のうちの各欄に関連付けられる第１の組の特性パラメータが計算される。各欄に関連付けられる複数の構成要素は、第１の組の特性パラメータに基づいて合成されて、１つ以上の副単語および／または１つ以上の単語を形成する。

【００１５】
図１は、本発明のさまざまな実施例に従い、複数の欄を含む画像の一例を示し、複数の欄の各欄はアラビア文字および／または非文字項目を含む。画像は濃淡画像またはカラー画像のいずれかであり得る。さらに、画像はごま塩雑音を含み、歪んでいるかもしれない。ＯＣＲ用に画像を前処理する前に、画像に関連付けられるごま塩雑音および歪みは取除かれる。さらに、画像は濃淡画像またはカラー画像から２進画像に変換される。
【００１６】
画像に関連付けられる歪みは、画像に関連付けられる基線を定めて、基線の配列に基づき画像を正しい位置に置くことによって修正される。基線の配列は、変形ハフ変換によって定められ、水平射影は複数の方向で定められる。水平射影は、画像の前景に関連付けられるピクセルの数を示す。アラビア語の近似単語長さに対応する妥当なランレングスが考慮されて、最も高いピクセル密度を有する方向が決定される。最も高いピクセル密度の方向が基線の配列と一致すると考えられる。その後、画像は基線の配列に基づき正しい位置に置かれる。図２は、画像に関連付けられる歪みを修正する際の画像の一例を示す。
【００１７】
歪みを修正する際、画像に関連付けられるごま塩雑音が取除かれる。ごま塩雑音は、任意に起こる白および黒ピクセルを表わし、暗い背景上の白い点または明るい背景上の黒い点を含み得る。一実施例において、ごま塩雑音はメディアンフィルタおよび／または多数フィルタを用いることによって除去することができる。当業者にとって、ごま塩雑音は当該技術分野におけるノイズ除去技術を用いることによっても除去できることは明らかであろう。
【００１８】
その後、画像は濃淡画像またはカラー画像から２進画像に変換される。たとえば画像が濃淡画像の場合、０から２５５の各ピクセル値を０のピクセル値または１のピクセル値に変換することにより、画像は２進画像に変換される。ある実施例において、ピクセル値０は背景値を表わし、ピクセル値１は前景値を表わす。代替的に、ピクセル値０は前景値を表わし、ピクセル値１は背景値を表わしてもよい。ピクセル値０は白ピクセルに関連付けられ、ピクセル値１は黒ピクセルに関連付けられる。
【００１９】
あるピクセルのピクセル値を変換する前に、濃淡画像にしきい値が定められ、しきい値より上のピクセル値はピクセル値１に変換され、しきい値より下のピクセル値はピクセル値０に変換される。一実施例において、しきい値は濃淡画像のピクセル値のヒストグラムを作成することによって計算される。ヒストグラムは、各ピクセル値の頻度を表わす。このヒストグラムを作成する際、連続するピクセル値の頻度を加算して、その連続するピクセル値を、連続するピクセル値の結合された頻度を有する単一のピクセル値に置き換えることにより、平滑化されたヒストグラムを生成することができる。考慮される連続するピクセル値の数は、予め定めることができる。後で、平滑化されたヒストグラムの２つの最も顕著なピークが選択され、この２つの顕著なピーク間の最小の谷が定められる。最も低い谷の中で最も低い頻度を有するピクセル値がしきい値として選択される。図３は例示的に２進に変換された画像を示す。
【００２０】
別の例であって、画像がカラー画像の場合、カラー画像はまず濃淡画像に変換され、次に上記のように２進画像に変換される。一実施例において、カラー画像を濃淡画像に変換するために、全国テレビジョン方式委員会（ＮＴＳＣ）のデフォルト値を用いることができる。
【００２１】
画像を２進画像に変換する際、ピクセル値０およびピクセル値１の発生数が数えられる。より低いカウントの２進値は前景値であると考えられ、より高いカウントを有する２進値は背景値であると考えられる。
【００２２】

図４を参照すると、本発明の一実施例に従い、光学式文字認識用に画像を前処理する方法のフロー図が示される。画像は複数の欄を含み、各欄はアラビア文字および／または非文字項目を含む。前記のように、画像の２進画像への変換、ごま塩雑音の除去、および画像に関連付けられる歪みの修正のいずれか１つ以上を行なうことにより、画像が得られる。画像を前処理するために、画像の中のアラビア文字および／または非文字項目に関連付けられる複数の構成要素がステップ４０２で決定される。ここで、構成要素は１組の繋がっているピクセルを含む。構成要素は、文字が他の文字に繋がらない場合、アラビア文字の１文字を表わす。したがって、複数の文字が他の文字に繋がる場合、繋がっている文字は１つの構成要素であると考えられる。複数の構成要素を決定する方法は、図５と併せてさらに説明される。

【００２３】

複数の構成要素を決定する際、複数の構成要素に関連付けられる行高さおよび欄間がステップ４０４で計算される。複数の構成要素に関連付けられる行の高さは、複数の構成要素の各構成要素の高さに対応する高さのヒストグラムを作成することによって計算される。行高さおよび／または欄間は、画像の複数の構成要素のすべての構成要素の平均値に基づき計算される。たとえば、行高さは、複数の構成要素のすべての構成要素に対して平均化された頻出高さである。行高さを計算する方法は、図６と併せてさらに詳しく説明される。欄間は、行高さの関数として動的に計算される。ステップ４０６において、複数の構成要素のうちの１つ以上の構成要素は、図７で例示的に示されるように、行高さおよび／または欄間に基づき、複数の欄のうちのある欄に関連付けられる。すなわち、複数の構成要素は、図７の７０２、７０４および７０６で例示的に示されるように、複数の構成要素に関連付けられる行高さおよび欄間に基づき、複数の欄に分離される。たとえば、２つの横方向に連続する構成要素間の間隔が欄間よりも小さければ、その構成要素は同じ欄のものであると考えられ、それに応じて分離される。

【００２４】

複数の構成要素がある欄に関連付けられると、ステップ４０８において、第１の組の特性パラメータが各欄について計算される。ある実施例において、第１の組の特性パラメータは、各欄に関連付けられる行高さ、各欄に関連付けられる語間、各欄に関連付けられる行間、各構成要素のピクセルの数、各構成要素の幅、各構成要素の高さ、各構成要素の座標、各構成要素の密度、および各構成要素のアスペクト比を含む。行高さ、語間、および行間を計算する方法は、それぞれ図６、図８、および図９と併せて説明される。

【００２５】

その後、ステップ４１０において、各欄に関連付けられた複数の構成要素は、第１の組の特性パラメータに基づき合成されて、１つ以上の副単語および／または１つ以上の単語を形成する。複数の構成要素を合成する方法は、図１０と併せてさらに説明される。

【００２６】

図５は、本発明の一実施例に従い、複数の構成要素を定める方法のフロー図を示す。ステップ５０２において、ラスタスキャンが画像に対して行なわれる。ラスタスキャンでは複数の掃引を行ない、複数の構成要素に対応する各ピクセル列に対して１回掃引される。ラスタスキャンの１回以上の掃引の実行により、画像の前景に関連付けられる１つ以上のピクセルがステップ５０４において特定される。画像の前景は、複数の構成要素に対応する。その後、ステップ５０６において、複数のピクセル間の相互接続に基づき、その複数のピクセルは統合されて、１つ以上の組の接続ピクセルを形成する。一実施例において、複数のピクセルは、８個の隣接ピクセルと１つ以上繋がっている場合に相互接続していると考えられる。こうして、アラビア文字の連続する文字は、連続する文字に関連付けられる１つ以上のピクセルが互いに相互接続されている場合に、単一の構成要素を形成する。

【００２７】

たとえば、ラスタスキャンの現行の掃引で特定されたピクセルは、当該ピクセルが前回の掃引で特定されたピクセルと繋がる場合には、そのピクセルと統合される。現行の掃引で特定されたピクセルが、前回の掃引で特定された複数のピクセルと繋がる場合、当該ピ

クセルはその複数のピクセルと統合される。別の例では、現行の掃引で特定された複数のピクセルが繋がっている場合、その複数のピクセルは統合される。同様に、ラスタスキャンの後続の掃引で特定される１つ以上のピクセルは、その１つ以上のピクセルが互いに繋がる場合にも統合される。統合されたピクセルは、１組の繋がっているピクセルを形成し、複数の構成要素のうちのある構成要素に対応付けられる。その結果、１つ以上の組の繋がっているピクセルは複数の構成要素に関連付けられる。

【００２８】

接続するピクセルの組を決定する際に、各構成要素のピクセルの数、各構成要素の幅、各構成要素の高さ、各構成要素の座標、各構成要素の密度、および各構成要素のアスペクト比のいずれか１つ以上は、各構成要素に関連付けられる接続ピクセルをトラッキングすることにより計算される。

【００２９】

図６を参照すると、本発明の一実施例に従い、行の高さを計算する方法のフロー図が示される。本方法はステップ６０２において、複数の構成要素の各々の高さに対応する高さのヒストグラムを作成することを含む。ヒストグラムは、複数の構成要素の各々の高さの頻出を表わす。ヒストグラムを作成する際、連続する高さ値の頻度を加算して、その連続する高さ値を、連続する高さ値の結合された頻度を有する単一の高さ値と置き換えることにより、平滑化されたヒストグラムを生成することができる。考慮される連続する高さ値の数は予め定められてもよい。たとえば、連続する高さ値の数が３個であると定められたのなら、高さが２０ピクセルの頻度は、高さが１９ピクセルの頻度プラス高さが２０ピクセルの頻度プラス高さが２１ピクセルの頻度となる。

【００３０】

平滑化されたヒストグラムが得られると、頻出高さがステップ６０４で特定される。頻出高さを特定するために、アラビア文字に対応するアクセント記号や句読点の小さな構成要素の高さは除外される。これは、しきい値高さを設定し、頻出高さを特定するのに、しきい値高さより大きい高さを有する構成要素のみを考慮することによって行なわれる。頻出高さは、画像が複数の文字サイズを有する場合には、画像の主要文字サイズを表わす。

【００３１】

頻出高さを特定する際、行高さはステップ６０６において頻出高さに基づき計算される。行高さは、頻出高さおよび乗率の積として計算される。乗率は頻出高さに依存する。行高さは、１つ以上の単語および／または１つ以上の副単語を、アラビア文字の１つ以上の横行に分割するのに用いることができる。さらに、行高さは、図８と併せて説明したように、語間を計算するために用いられる。

【００３２】

図８は、本発明の一実施例に従い、語間を計算する方法のフロー図を示す。本方法は、ステップ８０２において、複数の構成要素のうちの連続する構成要素間の間隔のヒストグラムを作成することを含む。一実施例において、縦方向に重なり、かつ他の構成要素によって分けられていない２つの構成要素はすべて連続する構成要素であると考えられる。２つの構成要素は、縦軸に沿って１つ以上の共通の座標を共有する場合、縦方向に重なる。すなわち、連続する構成要素はアラビア文字１行に属する。代替的に、２つの構成要素が縦方向に重ならない場合、２つの構成要素は予め定められた距離で分けられている場合に、連続する構成要素であると考えられる。

【００３３】

連続する構成要素間の間隔のヒストグラムを作成する際、平滑化されたヒストグラムは、連続する間隔値の頻度を加算することにより生成できる。連続する間隔値は、連続する間隔値の結合された頻度を有する単一の間隔値と置き換えられる。たとえば、１０ピクセルの間隔値の頻度は、９ピクセルの間隔値の頻度と、１０ピクセルの間隔値の頻度と、１１ピクセルの間隔値の頻度との合計と置き換えられる。

【００３４】

ステップ８０４において、平滑化されたヒストグラムから頻出間隔が特定される。頻出

間隔は、行高さによって定められるしきい値範囲内から特定される。たとえば、５分の１の行高さと半分の行高さとの間にある頻出間隔値を対象とすることができる。ステップ８０６において、語間は頻出間隔に基づき計算される。語間は、アラビア文字の２つの連続する単語の間の間隔である。

【００３５】

図９は、本発明の一実施例に従い、行間を計算する方法のフロー図を示す。ステップ９０２において、前景に対応する複数の構成要素の複数の水平射影のヒストグラムが作成される。水平射影は、ラスタスキャンの掃引に対応する複数の構成要素に関連付けられるピクセルの数を示す。たとえば、ラスタスキャンの掃引が、複数の構成要素に関連付けられる１５個のピクセルを特定すると、その掃引に対するピクセル列の水平射影は１５である。

【００３６】

その後、ステップ９０４において、２つの連続する最大水平射影間の平均距離が計算される。最大水平射影は、最も高い密度の領域を表わす。その後、ステップ９０６において、行間は、平均距離に基づき計算される。

【００３７】

第１の組の特性パラメータを計算する際、複数の構成要素は合成されて、図１０と併せて説明されるように、１つ以上の副単語および／または１つ以上の単語を形成する。

【００３８】

図１０は、本発明の一実施例に従い、１つ以上の副単語および／または１つ以上の単語を形成するために、複数の構成要素を合成する方法を示す。ステップ１００２において、ある欄に関連付けられる連続する構成要素間の間隔が語間の係数未満である場合に、当該連続する構成要素が結合される。語間に加えて、連続する構成要素の座標も、連続する構成要素が結合されるか否かを定めることができる。連続する構成要素の語間および／または座標に基づいてある欄に関連付けられる連続する構成要素を結合することは、アラビア文字のある単語または副単語に対応する異なる構成要素の結合を引起す。

【００３９】

たとえば、アクセント記号に関連付けられる構成要素は、構成要素の語間および位置に基づき、属する単語と結合される。１つの単語は１つ以上の構成要素を含み得る。構成要素の位置は、構成要素の座標によって定められる。ある構成要素に関連付けられる第１の組の特性パラメータのうち１つ以上が、アラビア文字の句読点またはアクセント記号と類似しており、かつアラビア語の文字に対応する構成要素に対して適切に隣接している場合、その構成要素は文字とともにグループ化されて単語または副単語を形成する。さもなければ、構成要素はノイズであると考えられ、除去される。

【００４０】

アラビア文字に関連付けられる構成要素を結合することに加えて、非文字項目に関連付けられる構成要素も、語間に基づき結合される。非文字項目に関連付けられる構成要素は結合されて、１つ以上のより大きい構成要素を形成する。

【００４１】

ステップ１００４において、非文字項目に関連付けられる構成要素は、第１の組の特性パラメータに基づき、アラビア文字に関連付けられる構成要素からふるい落とされる。たとえば、大きい高さ、大きい幅、および低い密度を有する構成要素は取除かれる。これらの構成要素は、ある欄の周りまたは他の非文字項目の周りの枠またはボーダーに対応し得る。同様に、大きい高さ、小さい幅、および高い密度を有する構成要素は縦線として認識され、除去される。横線は小さい高さ、大きい幅、および高い密度を有するものとして認識される。

【００４２】

同様に、他の非文字項目も１つ以上のフィルタに基づき除去される。この１つ以上のフィルタは、画像の共通に起こる構成要素の長さ、構成要素の幅、構成要素のアスペクト比、構成要素の密度、および構成要素の合計数を用いて、非文字項目をアラビア文字からふ

るい落とす。２つ以上の欄にわたる非文字項目も、非文字項目に関連付けられる構成要素の寸法を、アラビア文字に関連付けられる構成要素の最もよく起こる寸法と比較することによって除去される。図１１は、非文字項目が除去された画像を例示的に示す。

【００４３】

非文字項目をアラビア文字からふるい落とした後、１つ以上の単語および１つ以上の副単語の第２の組の特性パラメータが計算される。第２の組の特性パラメータは、各副単語および各単語の少なくとも一方に関連付けられる行高さ、各副単語および各単語の少なくとも一方に関連付けられる語間、ならびに各副単語および各単語の少なくとも一方に関連付けられる行間を含む。第２の組の特性パラメータは、１つ以上の副単語および／または１つ以上の単語を形成するために、複数の構成要素を結合するプロセスの精度をさらに上げるために、計算される。第２の組の特性パラメータに基づき、１つ以上の副単語をグループ化して、１つ以上の副単語および／または１つ以上の単語を形成する。

【００４４】

その後、縦方向に重なり、かつ複数の欄のうちのある欄に関連付けられる１つ以上の副単語および１つ以上の単語は、分割されてアラビア文字の横行を形成する。一実施例において、１つ以上の副単語および１つ以上の単語は、行高さおよび／または行間に基づいても分割されてもよい。たとえば、互いに重なる縦方向の構成要素を１つ以上有するので２本の横行が一緒に分割されると、その２本の横行は、行高さおよび／または行間に基づき分けられる。図１２は行高さおよび行間に依存して、１つ以上の単語および１つ以上の副単語を１つ以上の横行に分割する例を示す。

【００４５】

こうして、ＯＣＲ用に画像を前処理する方法が開示される。画像は複数の欄を含み、各欄はアラビア文字および／または非文字項目を有する。この方法は、アラビア文字および非文字項目に関連付けられる複数の構成要素を定めることを含む。複数の構成要素のうちのある構成要素は、アラビア語の１つ以上の文字または１つ以上の非文字項目を表わす。複数の文字が相互接続されるのなら、構成要素は２つ以上の文字を表わす。

【００４６】

複数の構成要素を決定する際、複数の構成要素に関連付けられる行高さおよび欄間が計算される。行高さおよび欄間は、すべての欄にわたるすべての構成要素の平均値を表わす。複数の構成要素は、平均行高さおよび平均欄間に基づき、１つ以上の欄に分離される。後で、各欄の複数の構成要素に関連付けられる第１の組の特性パラメータが計算される。各欄に関連付けられる複数の構成要素は、後で第１の組の特性パラメータに基づき合成されて、１つ以上の副単語および／または１つ以上の単語を形成する。

【００４７】

ここに開示されている方法は、繋がっている文字を含み、かつ複数の欄を含むアラビア文字を正確に前処理して分割することを可能にする。本方法は、構成要素がノイズであるのかアラビア文字の一部であるのかを判断する場合に、アラビア文字に関連付けられるアクセント記号および句読点を考慮する。さらに、本方法は画像が複数の欄を含むか否かを特定し、それらを分離する。

【００４８】

図１３は本発明の一実施例に従い、光学式文字認識（ＯＣＲ）のために、アラビア文字および／または非文字項目を含む画像を前処理するためのシステム１３００のブロック図を示す。画像は複数の欄を含み、各欄はアラビア文字および／または非文字項目を含む。画像は濃淡画像およびカラー画像のどちらかであり得る。さらに、画像はごま塩雑音を含み、歪んでいるかもしれない。図１３に示されるように、システム１３００は、メモリ１３０２と、メモリ１３０２に結合されるプロセッサ１３０４とを含む。ＯＣＲ用に画像を前処理する前に、プロセッサ１３０４は、図１と併せて説明したように、変形ハフ変換を用いて画像に関連付けられる歪みを除去する。その後、プロセッサ１３０４はごま塩雑音を除去し、濃淡画像またはカラー画像を２進画像に変換する。一実施例において、ごま塩雑音は、メディアンフィルタおよび／または多数フィルタを用いて除去され得る。ここで

は、画像を前処理するために、プロセッサ１３０４はアラビア文字および／または非文字項目に関連付けられる複数の構成要素を定める。構成要素は接続されたピクセルの組を含む。構成要素は、文字が他の文字と繋がらない場合、アラビア語文字の１つの文字を表わす。したがって、複数の文字が他の文字と繋がる場合、繋がっている文字は１つの構成要素であると考えられる。

【００４９】

一実施例において、複数の構成要素を定めるために、プロセッサ１３０４は画像に対してラスタスキャンを行なう。ラスタスキャンでは複数の掃引を行ない、複数の構成要素に対応する各ピクセル列に対して１回掃引される。ラスタスキャンの１回以上の掃引の実行により、画像の前景に関連付けられる１つ以上のピクセルが特定される。画像の前景は、複数の構成要素に対応する。その後、プロセッサ１３０４は、複数のピクセル間の相互接続に基づき、複数のピクセルを統合して、１つ以上の組の接続ピクセルを形成する。統合されたピクセルは１組の接続ピクセルを形成し、複数の構成要素のうちのある構成要素に関連付けられる。

【００５０】

こうしてプロセッサ１３０４によって定められた複数の構成要素は、メモリ１３０２に記憶することができ、プロセッサ１３０４によって用いられて、複数の構成要素に関連付けられる行高さおよび欄間が計算される。行高さおよび欄間を用いて、複数の構成要素のうちの１つ以上の構成要素を、複数の欄のある欄に関連付ける。すなわち、複数の構成要素が行高さおよび／または欄間を満たすのなら、複数の構成要素はプロセッサ１３０４によって複数の欄に分離される。たとえば、２つの縦にまたは横に連続する構成要素間の間隔が欄間よりも小さければ、それらの構成要素は同じ欄のものであると考えられて、分離される。その後、複数の構成要素に関連付けられる第１の組の特性パラメータが計算される。ある実施例において、第１の組の特性パラメータは、各欄に関連付けられる行高さ、各欄に関連付けられる語間、各欄に関連付けられる行間、各構成要素のピクセルの数、各構成要素の幅、各構成要素の高さ、各構成要素の座標、各構成要素の密度、および各構成要素のアスペクト比を含む。その後、プロセッサ１３０４は第１の組の特性パラメータに基づき、複数の構成要素を合成する。合成された構成要素は、１つ以上の副単語および／または１つ以上の単語を形成する。

【００５１】

一実施例において、プロセッサ１３０４は、複数の構成要素の各々の高さに対応する高さのヒストグラムを作成することによって行高さを計算する。ヒストグラムから頻出高さがプロセッサ１３０４によって特定される。その後、プロセッサ１３０４は頻出高さおよび乗率の積として、行高さを計算する。乗率は頻出高さに依存する。行高さを用いて、１つ以上の単語および／または１つ以上の副単語を、アラビア文字の１つ以上の横行に分割することができる。さらに、プロセッサ１３０４は行高さを用いて語間を計算する。

【００５２】

次に、プロセッサ１３０４は、複数の構成要素の連続する構成要素間の間隔のヒストグラムを作成することにより、語間を計算する。プロセッサ１３０４は、ヒストグラムから頻出間隔を特定する。頻出間隔はしきい値範囲内から特定され、そのしきい値範囲は行高さに基づいている。その後、語間はプロセッサ１３０４によって頻出間隔に基づき計算される。語間は、アラビア文字の２つの連続する単語の間の間隔である。

【００５３】

プロセッサ１３０４は、複数の構成要素の複数の水平射影のヒストグラムを作成することによって行間を計算するようにも構成されている。水平射影は、ラスタスキャンの各掃引に対応する複数の構成要素に関連付けられるピクセルの数を示す。次に、２つの連続する水平射影間の平均距離は、プロセッサ１３０４によって計算される。その後、プロセッサ１３０４は平均距離に基づき、行間を計算する。

【００５４】

さらに、プロセッサ１３０４は、各構成要素のピクセル数、各構成要素の幅、各構成要

素の高さ、各構成要素の座標、各構成要素の密度、および各構成要素のアスペクト比を定める。

【００５５】

前述のように、プロセッサ１３０４は、第１の組の特性パラメータに基づき、各欄に関連付けられる複数の構成要素を合成する。これを行なうため、プロセッサ１３０４は、構成要素間の間隔がその欄に関連付けられる語間の係数よりも小さい場合に、連続する構成要素を結合する。各欄に関連付けられる語間に加えて、連続する構成要素の座標も連続構成要素が結合されるか否かを定めることができる。さらに、プロセッサ１３０４は、図１０と併せて説明されたように、第１の組の特性パラメータに基づき、アラビア文字に関連付けられる構成要素から、非文字項目に関連付けられる構成要素をふるい落とす。非文字項目をふるい落とすことは、１つ以上の副単語および／または１つ以上の単語をもたらす。

【００５６】

プロセッサ１３０４は、１つ以上の副単語および／または１つ以上の単語に関連付けられる第２の組の特性パラメータを計算するようさらに構成されている。第２の組の特性パラメータは、各副単語および／または各単語に関連付けられる行高さ、各副単語および／または各単語に関連付けられる語間、ならびに各副単語および／または各単語に関連付けられる行間を含む。次に、２つ以上の副単語は、第２の組の特性パラメータに基づきプロセッサ１３０４によってグループ化されて、１つ以上の副単語および／または１つ以上の単語を形成する。すなわち、２つ以上の副単語は第２の組の特性パラメータに基づきグループ化されて、完全な単語またはより大きい副単語を形成する。

【００５７】

１つ以上の副単語および１つ以上の単語を形成する際、プロセッサ１３０４は、縦方向に重なりかつ複数の欄のうちのある欄に関連付けられる１つ以上の副単語および１つ以上の単語を分割して、アラビア文字の横行を形成する。一実施例において、１つ以上の副単語および１つ以上の単語は、行高さおよび／または行間に基づき、プロセッサ１３０４によって分割されてもよい。

【００５８】

図１４および図１５は、文字欄（または領域）の一部が他の欄内に入るまたは他の欄と重なるある事例を図示する。たとえば、文字の１つの矩形のまたは円形の区域が文字のより大きな矩形のまたは円形の区域内に入ってもよい。

【００５９】

図１４は、文字領域「テキスト２」が文字領域「テキスト１」内にあり、文字領域「テキスト３」が「テキスト４」と部分的に縦方向に重なる１つの例を示す。これらの文字領域は、文章の１つ以上の行または列を有してもよい。以上で述べた技術と同様に、図１４および図１５の例が図示する技術は分割および前処理エンジンの使用に係る。複数の欄、文字および非文字、起こり得る歪み、起こり得るノイズを有する、濃淡またはカラーの完全未加工スキャン文書が分割および前処理エンジンに与えられる。この例では、このシステムの出力は分類された欄のシーケンスであり、かつ２進または濃淡形式の、最適な行認識エンジンに渡すべき、正しい側が上になった、歪んでいない、ノイズのない各欄内の行画像の一覧である。

【００６０】

文書は、分割および前処理エンジンがページ構造もしくは内容またはフォントまたはフォントのサイズの事前の情報を有していなくても、適切に正しい位置に置かれ、ノイズのないきれいな分割された文字行画像に分割される。

【００６１】

上述のように、濃淡画像は、それが元々であってもまたはカラー画像から変換されたものであっても、ページ毎に異なる動的しきい値化を用いて２進化画像に変換される。異なる背景および／またはシェーディングを有する領域が特定され、別々に２進化される。正確さのために、各々の領域は別々に２進化されてもよい。以上のように、文字ページの前

景が背景よりも少ないピクセルを有する各領域に対して、前景／背景のチェックが行なわれる。各領域の２進化は濃淡のヒストグラムを用いて行なわれる。

【００６２】

繋がっている構成要素は、単一ランレングスのアルゴリズムを用いて２進画像から計算される。実現例に従うと、スキャンは、２進画像に対して左上から右へ文書の下方に移動して行なわれる。各々のランレングス上でセグメントが見出され、長いセグメントは不純物として無視される。このプロセスを次の行以降について繰返して、上述のように、すべての新しい構成要素が形成されるまで行同士の間で接続するセグメントを合成する。古い列に対して新しい列がチェックされる。あるセグメントまたは未完了の構成要素に何も接続していなければ、これは完了済み構成要素に変換される。１つ以上の未完了構成要素に接続している新しいセグメントは、それらに加えられるか、または１つよりも多くの未完了構成要素を１つに合成するように用いられる。

【００６３】

１つ以上の寸法、アスペクト比、密度、および周などの構成要素パラメータを用いて、繋がっている構成要素に対して統計的分析を行なう。明白な非文字構成要素をふるい落とすように規則が加えられる。

【００６４】

次に残余の文字状の構成要素に対して統計的分析を行なって、ページが他のサイズの文字および領域を有するかもしれないとしても、支配的な文字サイズの単語または行の共通の高さを見出す。各々の単一の欄に属する構成要素に対して再び統計的分析を行なって、異なる欄が異なるサイズの文字を有する場合は、新たな欄特有の単語の共通の高さを用いてさらに文字および非文字を分ける。語間および欄間はこの共通高さを用いて動的に計算される。このように、変更できない値は存在せず、ページ計算は動的であり、ページサイズが拡大または縮小されたとしても変化しない。これは、サイズとともに変化する無次元パラメータを用いて達成される。計算された距離が計算された語間よりも大きくかつ欄間よりも小さければ、単語を合成してもよい。

【００６５】

構成要素を欄に分離するのに用いられる単語高さおよび欄間に基づいて距離が計算される。欄のうちいくつかは、「テキスト１」領域内に入る「テキスト２」領域などの他の欄内、または「テキスト３」領域の部分の下に入る「テキスト４」領域などの文字欄が横部分の下または上にある任意の方向のＬ字型欄内に入り得るので、各々の欄を規定する矩形は横および／または縦方向に重なる。区別される重なる欄または領域に構成要素を分けることは、各々の構成要素を分析すること、およびこの構成要素に最も近い文字構成要素がその欄内に入る場合は構成要素を欄と合成することによってなされ、その距離は以上で計算された余白よりも小さい。最も近い欄と構成要素とを合成すれば、１つの大きな欄となるすべての重なり合う欄となるため、合成は、現在の構成要素に最も近い構成要素を包含する欄に対してなされる。

【００６６】

図１５は、図１４に示すような、領域「テキスト３」として処理されるように次に合成される領域「テキスト３ａ」－「テキスト３ｅ」などの１つ以上の構成要素を分析する例示的な技術を示す。同様のフォントサイズおよび近似した寸法を有する互いの上にある層である「テキスト３ａ」－「テキスト３ｅ」などの欄は、高さまたは幅などの寸法が異なる１つの欄または領域に合成されてもよい。欄または領域は、ページ上のそれらの場所（たとえばそれらの座標）に基づいて分類されて、認識された結果的に得られた文字を正しい順序で流すのを助ける。文書の上から下へかつ左から右へ文字を処理して、ページの幅または高さ全体にわたり得るいくつかの広い欄を利用して、正しい位置に置くことおよび文書の文字をどのように処理するかについての視覚的手がかりを与えるように、後処理アルゴリズムを構成してもよい。当該アルゴリズムは、処理のために欄を分けるのに非文字行またはボーダーも利用してもよい。平均行高さを求めるのに統計的分析をさらに用いてもよい。行の厚みが平均行高さよりも数倍大きければ、行は連続した行同士の間の最も重

なっていない横方向の場所でいくつかの行に分けられる。

【００６７】

本発明の多様な実施例は、ＯＣＲ用に画像を前処理するための方法およびシステムを提供する。画像は複数の欄を含み、各欄はアラビア文字および／または非文字項目を含む。本発明は、アラビア文字を、ＯＣＲシステムによって正確に処理することができる分割された行の副単語および単語に分割する。本方法は、構成要素がノイズであるのかアラビア文字の一部であるのかを判断する場合に、アラビア文字に関連付けられるアクセント記号および句読点も考慮に入れる。

【００６８】

当業者は、ここに記載される上記認識される利点および他の利点は一例であって、本発明のさまざまな実施例の利点すべてを含むことは意図されていないと認識するであろう。

【００６９】

以上で、本発明の具体的実施例が説明された。しかし、当業者なら、さまざまな変形および変更が、添付の請求項に記載されている本発明の範囲から逸脱することなく行なうことができると理解するであろう。したがって、明細書本文および図面は限定するのではなく例示するものであり、変形はすべて本発明の範囲内に含まれることが意図される。利益、利点、問題の解決、および利益、利点または解決を引起すまたは顕著にする要素は、請求項のいずれかまたはすべてにおける重大な、必要な、または必須の特徴もしくは要素であると考えるべきではない。本発明は、本願の係属中になされた補正を含む添付の請求項およびこれら請求項の均等物すべてによってのみ規定される。

【符号の説明】

【００７０】

１３００　システム

１３０２　メモリ

１３０４　プロセッサ

４０２　アラビア文字および／または非文字項目に関連付けられる複数の構成要素を定める

４０４　複数の構成要素に関連付けられる行高さおよび欄間を計算する

４０６　複数の構成要素のうちの１つ以上の構成要素をある欄に関連付ける

４０８　各欄について、第１の組の特性パラメータを計算する

４１０　第１の組の特性パラメータに基づいて各欄の複数の構成要素を合成して、１つ以上の副単語および１つ以上の単語のうちの少なくとも１つ以上を形成する

【図１】


أخبار المدينة
مدينة الملك عبدالعزيز للعلوم والتقنية KACST
التأسيس لعلاقة استراتيجية بين المدينة و جامعة الملك عبدالله للعلوم والتقنية
معالي وزير الثقافة والإعلام يزور مدينة «العلوم والتقنية»
د. تركي بن سعود : دعم ٣٠ مشروع بحثي استراتيجي بتكلفة تجاوزت ٥٩٤ مليون ريال خلال عام واحد

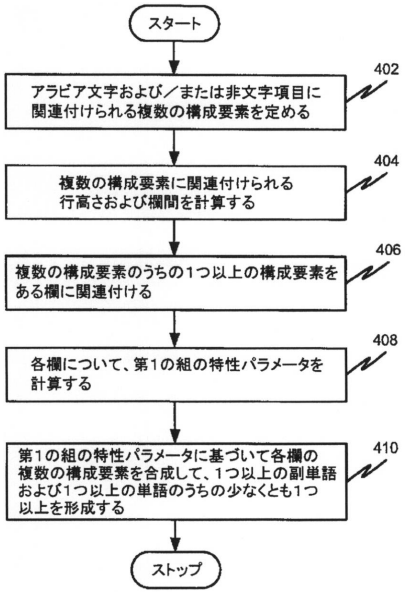
【図２】


أخبار المدينة
مدينة الملك عبدالعزيز للعلوم والتقنية KACST
التأسيس لعلاقة استراتيجية بين المدينة و جامعة الملك عبدالله للعلوم والتقنية
معالي وزير الثقافة والإعلام يزور مدينة «العلوم والتقنية»
د. تركي بن سعود : دعم ٣٠ مشروع بحثي استراتيجي بتكلفة تجاوزت ٥٩٤ مليون ريال خلال عام واحد

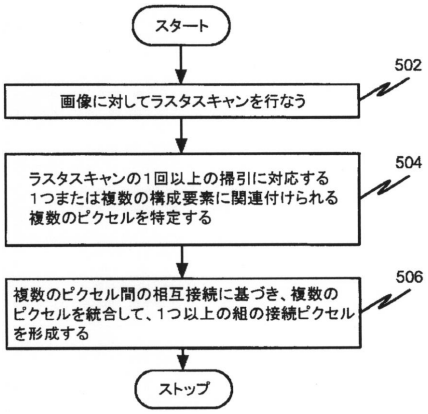
【図３】


أخبار المدينة
مدينة الملك عبدالعزيز للعلوم والتقنية KACST
التأسيس لعلاقة استراتيجية بين المدينة و جامعة الملك عبدالله للعلوم والتقنية
معالي وزير الثقافة والإعلام يزور مدينة «العلوم والتقنية»
د. تركي بن سعود : دعم ٣٠ مشروع بحثي استراتيجي بتكلفة تجاوزت ٥٩٤ مليون ريال خلال عام واحد

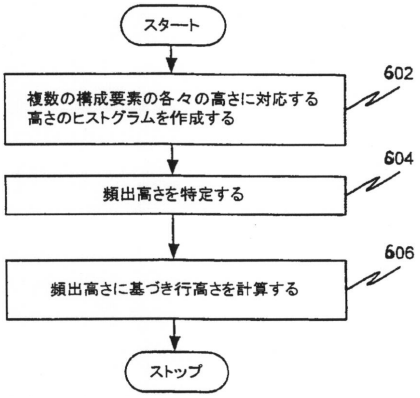
【図４】


スタート
アラビア文字および／または非文字項目に関連付けられる複数の構成要素を定める
402
複数の構成要素に関連付けられる行高さおよび欄間を計算する
404
複数の構成要素のうちの1つ以上の構成要素をある欄に関連付ける
406
各欄について、第1の組の特性パラメータを計算する
408
第1の組の特性パラメータに基づいて各欄の複数の構成要素を合成して、1つ以上の副単語および1つ以上の単語のうちの少なくとも1つ以上を形成する
410
ストップ

【図５】

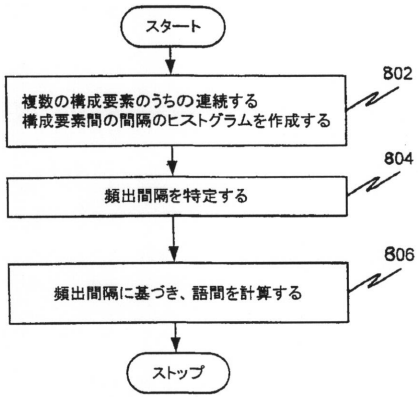


【図６】

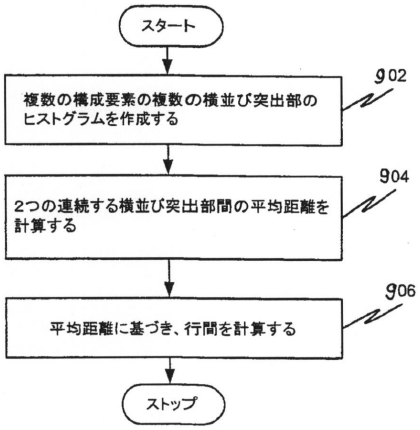


【図７】

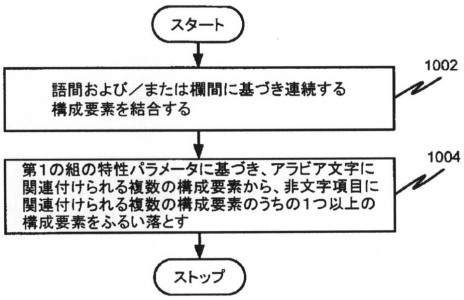
【図８】



【図９】



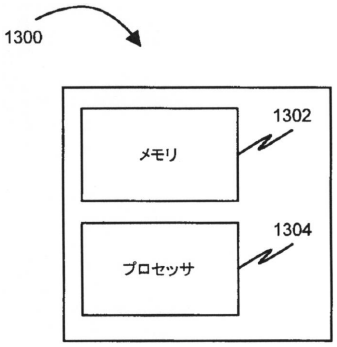
【図１０】



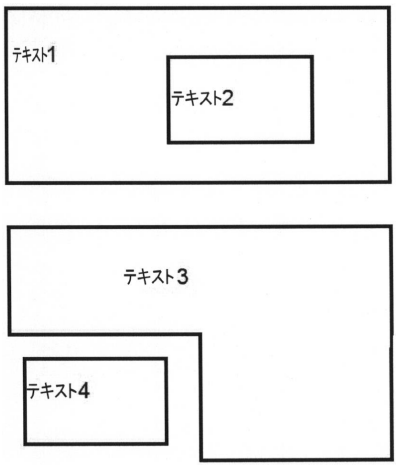
【図１１】

【図１２】

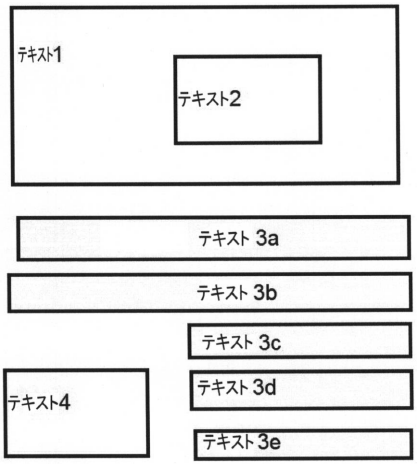
【図１３】



【図１４】



【図１５】



フロントページの続き

(72)発明者　フセイン・ハリド・アルーオマリ
ヨルダン、１１９５３　アンマン、ピィ・オゥ・ボックス・２７５０
(72)発明者　モハメド・スレイマン・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、ビルディング・１７、ブロック・エフ、オフィス・４１５、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

審査官　松浦　功

(56)参考文献　特開２０１２－００３７５６（ＪＰ，Ａ）
特表２００９－５４５８０７（ＪＰ，Ａ）
国際公開第２００８／１３８３５６（WO，Ａ１）
米国特許出願公開第２０１０／０２４６９６３（ＵＳ，Ａ１）
米国特許出願公開第２０１０／０２７２３６１（ＵＳ，Ａ１）
米国特許出願公開第２０１２／００８７５８４（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｋ　　９／２０　－　９／６０