(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2017-105175
(P2017-105175A)
(43)公開日　平成29年6月15日(2017.6.15)

(51)Int.Cl.		F I			テーマコード（参考）
B41J 5/30	(2006.01)	B41J	5/30	Z	2C061
G06F 3/12	(2006.01)	G06F	3/12	344	2C187
G06K 9/00	(2006.01)	G06K	9/00	Z	5B064
B41J 29/38	(2006.01)	G06F	3/12	302	
		G06F	3/12	324	

審査請求 未請求　請求項の数 10　ＯＬ　（全 17 頁）　最終頁に続く

(21)出願番号　特願2016-225518 (P2016-225518)
(22)出願日　平成28年11月18日 (2016.11.18)
(31)優先権主張番号　14/960,986
(32)優先日　平成27年12月7日 (2015.12.7)
(33)優先権主張国　米国(US)

（特許庁注：以下のものは登録商標）
１．イーサネット

(71)出願人　596170170
ゼロックス　コーポレイション
ＸＥＲＯＸ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国、コネチカット州　０６８５６、ノーウォーク、ピーオーボックス　４５０５、グローバー・アヴェニュー　４５
(74)代理人　100079049
弁理士　中島　淳
(74)代理人　100084995
弁理士　加藤　和詳
(72)発明者　ジェロウム・ポウヤドウ
フランス共和国　グルノーブル　３８０００　リュ・デ・ラ・ペ　１

最終頁に続く

(54)【発明の名称】ページ記述言語文書からの直接文字認識

(57)【要約】　（修正有）
【課題】テキストベースのＰＤＬ文書の印刷時に、そのテキストベースのＰＤＬ文書から文字を直接認識できるシステムを提供する。
【解決手段】システム１０は、印刷文書に関して印刷ドライバによって生成され、一連のグリフから形成されたテキストランを描画するための描画命令セットを含むＰＤＬ（ページ記述言語）文書２０を取り込む取り込みコンポーネント４２と、ＰＤＬフォーマットに従ってＰＤＬ文書の各グリフのすべてのテキスト関連演算子を捕捉して直接文字認識を行う解析システム４４と、を備える。解析システムは、ＰＤＬ文書を解析して各グリフの描画命令を抽出し、グリフ特徴データベース６４と比較する。一致するものが見つかった場合は、グリフに関連付けられたテキスト文字を抽出し、抽出されたテキスト文字の要約を生成する。一致するものが見つからなかった場合は、見つけ損ねたテキスト文字のログを取る。
【選択図】図１

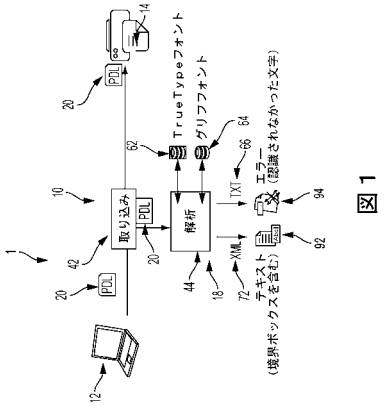

図１

【特許請求の範囲】
【請求項１】
ページ記述言語（ＰＤＬ）文書からテキストを抽出する方法であって、
印刷される文書に関して、印刷ドライバによって生成されたＰＤＬファイルを取り込むステップであって、前記ＰＤＬファイルが、前記印刷される文書にある一連のグリフから形成されたテキストランに関する描画命令セットを含むＰＤＬ文書を含む、ステップと、
前記ＰＤＬ文書を解析して、グリフセットのそれぞれに関して、描画命令を抽出する、ステップと、
前記グリフセットにある各グリフに関して、前記グリフの前記描画命令を、グリフ特性のデータベースと比較して、一致するグリフがあるかどうかを判定する、ステップと、
前記描画命令とグリフ特性の前記データベースとの間で一致するものが見つかった場合、前記グリフに関連付けられたテキスト文字を抽出するステップと、
前記一致するグリフに関連付けられた前記抽出されたテキスト文字の要約を生成するステップと、
を含む、方法。
【請求項２】
取り込む、解析する、比較する、抽出する、および生成する、前記ステップのうちの少なくとも１つが、プロセッサで実行される、請求項１に記載の方法。
【請求項３】
前記描画命令とグリフ特性の前記データベースとの間で一致するものが見つからなかった場合、見つけ損ねたテキスト文字のログを取るステップをさらに含む、請求項１に記載の方法。
【請求項４】
前記グリフ特性セットが、グリフセットのそれぞれに関するランレングス特徴ベクトルを含む、請求項１に記載の方法。
【請求項５】
コンピュータによって実行されると請求項１に記載の方法を行う命令を格納する非一時的記録媒体を備える、コンピュータプログラム製品。
【請求項６】
請求項１に記載の方法を行うための命令を格納するメモリと、前記命令を実行するために、前記メモリと通信可能なプロセッサとを備える、文書のページ記述言語（ＰＤＬ）からテキストを抽出するためのシステム。
【請求項７】
文書のページ記述言語（ＰＤＬ）からテキストを抽出するためのシステムであって、
印刷される文書に関して、印刷ドライバによって生成されたＰＤＬファイルを取り込む、取り込みコンポーネントであって、前記ＰＤＬファイルが、前記印刷される文書にある一連のグリフから形成されたテキストランに関する描画命令セットを含む、取り込みコンポーネントと、
前記ＰＤＬファイルを解析して、各グリフの前記描画命令を捕捉する、解析器と、
各グリフの前記描画命令を、グリフ特性のデータベースと比較する、比較コンポーネントと、
グリフ特性の前記データベースとの前記描画命令の前記比較に基づいて前記描画命令とグリフ特性の前記データベースとの間で一致するものが見つかった場合、各グリフに関連付けられたテキスト文字を抽出する、抽出コンポーネントと、
各グリフに関連付けられた前記抽出されたテキスト文字のテキスト要約を生成する、要約コンポーネントと、
前記取り込みコンポーネントと、解析器と、比較コンポーネントと、抽出コンポーネントと、要約コンポーネントとを実装するプロセッサと、
を備える、システム。
【請求項８】

前記描画命令の、グリフ特性の前記データベースとの比較に基づいて、前記描画命令とグリフ特性の前記データベースとの間で一致するものが見つからなかった場合、見つけ損ねたテキスト文字に関するエラーログを生成するログ取得コンポーネントをさらに備える、請求項７に記載のシステム。

【請求項９】

各グリフをビットマップとしてレンダリングする、レンダリングコンポーネントと、

各グリフの前記ビットマップに基づいて特徴ベクトルを抽出する、特徴抽出部と、

前記抽出された特徴ベクトルとグリフ特性の前記データベースとの間の類似度を計算する、類似度コンポーネントと、

をさらに備え、

前記抽出コンポーネントが、前記抽出された特徴ベクトルとグリフ特性の前記データベースとの間の前記類似度に基づいて各グリフに関連付けられたテキスト文字を抽出する、請求項７に記載のシステム。

【請求項１０】

文書のページ記述言語（ＰＤＬ）からテキストを抽出する方法であって、

テキスト文字のそれぞれに関連付けられたグリフ特性セットによって定義される前記テキスト文字の参照データベースを提供するステップと、

印刷ドライバによって生成されたＰＤＬ文書を受信するステップと、

プロセッサにより、前記ＰＤＬ文書を解析して、前記ＰＤＬ文書にあるテキスト描画プリミティブを識別する、ステップと、

前記テキスト描画プリミティブの、前記グリフ特性セットとの比較に基づいて、前記参照データベースにある前記テキスト描画プリミティブによって表現されるテキスト文字を識別するステップと、

前記テキスト描画プリミティブによって表現される前記テキスト文字を抽出するステップと、

前記抽出されたテキスト文字に基づいて情報を出力するステップと、

を含む、方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本例示的な実施形態は、テキスト抽出の分野に関し、詳細には、光学文字認識を必要とすることなく、文書からのテキスト抽出に利用できる。

【背景技術】

【０００２】

ページ記述言語（ＰＤＬ）は、印刷される文書を、印刷機に依存しないフォーマットで記述する。電子ＰＤＬ文書が送られる印刷機において、文書は紙などの印刷媒体の上にレンダリングされる。ここでは、文書のテキストの内容を、ＰＤＬフォーマットで取り込むことが望ましい事例がいくつかある。そのような事例としては、印刷されるべきではない、文書上の機密情報を検出するためのセキュリティチェックの実行、印刷コストを課金するための、文書を印刷させた顧客の検出、私的な印刷ジョブと公的な印刷ジョブとを区別するため、またはカラー印刷機が適切な文書のために使用されているかをチェックするための利用制御、印刷しているユーザおよびその理由を検出するための監査、ならびに後の検索を加速するためにすべての印刷文書をアーカイブして索引付けする状況における索引付けが挙げられる。

【０００３】

しかしながら、光学文字認識（ＯＣＲ）などの、文字認識のための既存の方法は、ＰＤＬフォーマットの文書の処理に適用できない。１つの理由としては、印刷される文書のケースにおいて、印刷レンダリングエンジン、すなわちラスタイメージプロセッサ（ＲＩＰ）が、どの文字が印刷されるかを知る必要がないことが挙げられる。ＲＩＰは、全体的な一体となった結果が人間の読み手にとって意味をなすテキストに見えるように、印刷ペー

ジ上のどこにインクのドットを置くかを分かっていればよいだけである。大半のＰＤＬおよび大半の文書で、テキストをレンダリングするための表記法およびＡＰＩが存在し得る。しかしながら、表示される文字の実際の「値」は、レンダリングに関係しない。よって、この情報は、印刷機に届くデータに含まれていない。

【０００４】

ＰＤＬをテキストにする既存のツールは、既知の識別子（通常、文字列）とそれらが表すグリフとの間に対応構造が存在すると仮定することによって、ＰＤＬ文書からテキストを抽出する。このことはよくあることとはいえ、常にというわけではなく、表現される文字セットを拡張するのにこの種の対応を使わない文書は多い。こうしたケースでは、テキスト抽出は、誤った結果を生じる。この種の対応を使わない他のケースでは、テキスト抽出は不可能であると一般的に考えられている。加えて、入手可能なツールは、ＰｏｓｔＳｃｒｉｐｔコンピュータ言語の上だけで動作する傾向にあるが、ＰｏｓｔＳｃｒｉｐｔコンピュータ言語は、現在使用されている多くの入手可能なＰＤＬのうちの１つに過ぎない。

【０００５】

テキスト抽出のための別の手法は、印刷機にＰＤＬ文書が到着した時に、または印刷機自体において、ＰＤＬ文書を画像に変換することと、テキストを再構築するためにＯＣＲ技法を適用することとを必要とする。しかしながら、レンダリングしてＯＣＲにかける手法では、文書全体に対してＯＣＲを行う前にすべてのページをレンダリングすることから、特に、文書に含まれるページ数が多い場合には、時間がかかり得る。この処理時間は、セキュリティ検出の状況において特に問題となり得、そのような状況では、ルールベースのエンジンが文書内で特定の単語を探すのであるが、その単語が出現するのが最初のページだとしても、その単語を検出できるまでには文書全体のレンダリングおよびＯＣＲを待たねばならない。

【発明の概要】

【発明が解決しようとする課題】

【０００６】

したがって、テキストベースのＰＤＬ文書の印刷時に、そのテキストベースのＰＤＬ文書から文字を直接認識できるシステムおよび方法のニーズがある。

【課題を解決するための手段】

【０００７】

本例示的な実施形態の一実施態様によれば、ページ記述言語（ＰＤＬ）文書からテキストを抽出する方法が提供される。本方法は、印刷される文書に関して、印刷ドライバによって生成されたＰＤＬファイルを取り込むステップを含む。ＰＤＬファイルは、印刷される文書において一連のグリフから形成されているテキストランに関する描画命令セットを含む、ＰＤＬ文書を含む。ＰＤＬ文書を解析して、グリフセットのそれぞれに関する描画命令を抽出する。グリフセットの各グリフに関して、グリフの描画命令を、グリフ特性のデータベースと比較して、データベースに一致するグリフがあるかどうかを判定する。描画命令とグリフ特性のデータベースとの間で一致するものが見つかった場合、一致するデータベースグリフに関連付けられたテキスト文字を抽出する。抽出した、一致するグリフに関連付けられたテキスト文字の要約を生成する。

【０００８】

本方法のステップのうちの１つ以上は、プロセッサによって実装され得る。

【０００９】

本例示的な実施形態の別の実施態様によれば、文書のページ記述言語（ＰＤＬ）からテキストを抽出するためのシステムは、印刷文書に関して印刷ドライバによって生成されたＰＤＬファイルを取り込む、取り込みコンポーネントを含む。ＰＤＬファイルは、印刷文書における一連のグリフから形成されたテキストランを描画するための描画命令セットを含む。解析器は、ＰＤＬファイルを解析して、各グリフの描画命令を捕捉する。比較コンポーネントは、各グリフの描画命令を、グリフ特性のデータベースと比較する。抽出コン

ポーネントは、描画命令の、グリフ特性のデータベースとの比較に基づいて、描画命令とグリフ特性のデータベースとの間で一致するものが見つかった場合、各グリフに関連付けられたテキスト文字を抽出する。要約コンポーネントは、抽出した、各グリフに関連付けられたテキスト文字のテキスト要約を生成する。プロセッサは、取り込みコンポーネントと、解析器と、比較コンポーネントと、抽出コンポーネントと、要約コンポーネントとを実装する。

【００１０】

本例示的な実施形態の別の実施態様によれば、文書のページ記述言語（ＰＤＬ）からテキストを抽出する方法は、テキスト文字のそれぞれに関連付けられたグリフ特性セットによって定義される、テキスト文字の参照データベースを提供するステップと、印刷ドライバによって生成されたＰＤＬ文書を受信するステップとを含む。プロセッサによって、ＰＤＬ文書を解析して、ＰＤＬ文書におけるテキスト描画プリミティブを識別する。本方法は、テキスト描画プリミティブの、グリフ特性セットとの比較に基づいて、参照データベースにおいて、テキスト描画プリミティブによって表現されるテキスト文字を識別するステップと、テキスト描画プリミティブによって表現されるテキスト文字を抽出するステップと、抽出されたテキスト文字に基づいて情報を出力するステップとをさらに含む。

【００１１】

本方法のステップのうちの１つ以上は、プロセッサによって実装され得る。

【図面の簡単な説明】

【００１２】

【図１】図１は、ページ記述言語（ＰＤＬ）文書からの直接文字認識のためのシステムおよび方法の概略図である。

【図２】図２は、本例示的な実施形態の一実施態様による、ページ記述言語（ＰＤＬ）文書からの直接文字認識のためのシステムの機能ブロック図である。

【図３】図３は、図２のシステムの追加的な実施態様を示すブロック図である。

【図４】図４は、グリフ特性によって定義されるテキスト文字の参照データベースの記号的表現の図である。

【図５】図５は、本例示的な実施形態の別の実施態様による、ＰＤＬ文書からの直接文字認識のための例示的な方法を示す流れ図である。

【発明を実施するための形態】

【００１３】

本例示的な実施形態の実施態様は、一般的な事務文書の印刷時または印刷時付近に、一般的な事務文書にローカル文字認識を適用するための、方法、装置、およびコンピュータ可読媒体に関する。本方法は、同様のフォント技法が使用される状況およびグリフを文字にする情報が失われている場合の他の状況において使用可能であるほど十分に汎用的である。

【００１４】

本例示的な方法は、ページ記述言語（ＰＤＬ）文書からの直接文字認識を提供する。本方法は、ユーザが印刷ジョブを開始した時、例えばアプリケーションから「印刷」をクリックすることによって作成されるＰＤＬファイルを解析するステップを含み得る。解析されるＰＤＬ文書は、描画命令から構成され得る。処理される記号（グリフ）の各視覚的表現に関して、グリフ特性を、参照データベースと照合することによって、それが表現する文字を見つける。参照データベースは、フォントのデータベースにおけるグリフルックアップに基づくものであり得る。フォントは、グリフの２次スプライン制御点によって直接認識が行われる、ＴｒｕｅＴｙｐｅ／ＯｐｅｎＴｙｐｅ技術を使用できる。他のケースでは、半厳密な、視覚的類似度によって認識が行われる。

【００１５】

一実施形態では、本方法は、ＰＤＬファイルからテキスト関連演算子セットを捕捉するステップを含み得る。

【００１６】

図１は、ページ記述言語（ＰＤＬ）文書からの直接文字認識のためのシステム１０が動作する、印刷ネットワーク１における文書処理の概略図である。一般的に、システム１０では、開示される実施形態と整合するコンピュータソフトウェアが使用され得る。ネットワークは、コンピュータ１２などの、１つ以上のネットワーク化されたＰＤＬ生成装置と、従来の通信プロトコルおよび／またはデータポートインターフェースを用いて情報を交換できるようにする通信リンク１６（図２）を通して接続される、１つ以上の印刷機１４とを含み得る。

【００１７】

ネットワーク１内の各計算装置１２は、コンピュータワークステーションもしくはデスクトップコンピュータ、ラップトップもしくはポータブルコンピュータ、携帯機器、またはネットワーク化環境で使用可能な任意の他の計算装置であり得る。

【００１８】

図２に示すように、ＰＤＬ文書２０は、ユーザが計算装置１２において印刷ジョブを要求した時に、印刷ドライバ２２によって作成される。元の文書２４を印刷するには、計算装置１２のユーザは、例えば、印刷機１４などの送信先印刷機と、元の文書を印刷するためのいくつかの仕上げオプション（例えば、両面モード、ホチキス留め、カラーもしくは単色、およびページ選択）とをユーザが選択できる、通常「印刷ダイアログ」の形態である、特定のアプリケーションコマンドを利用する。印刷オプションの選択がなされた時（例えば、ユーザが、アイコンまたはキーボードのキーなどの「印刷」セレクタをクリックした時）、アプリケーションは、印刷ドライバ２２によって提供される仮想ディスプレイの形態を用いて、文書の各ページをデジタル形式でレンダリングする。特に、印刷ドライバ２２は、アプリケーションによって提供された描画命令を、印刷機固有の描画命令に変換する。これらの印刷機固有の描画命令は、それら独自の構文を有し、その構文は、元のアプリケーションの描画構文とは異なる。このような印刷機固有の描画命令セットはＰＤＬ文書２０を構成し、ＰＤＬ文書２０は、ＰＤＬファイル２６で印刷機に送信され得る。

【００１９】

印刷ドライバ２２は、コンピュータ１２の一部であってもよいし、計算装置１２と印刷機１４との中間にある別個の計算装置上に配置されてもよい。

【００２０】

従来のネットワーク印刷システムのように、ＰＤＬファイル２６は、特定のアプリケーションにおいてオープンな、固有のフォーマットの元の文書２４から、印刷ドライバによって、ＰＳまたはＰＣＬ６などの特定のＰＤＬフォーマットで作成される。例えば、事務文書は、その種類に応じて、ワードプロセッサ、スプレッドシートハンドラ、またはスライドプレゼンテーションアプリケーションで開くことができる。この元の文書２４のフォーマットは、公知であってもよいし、そうでなくてもよいが、いずれにせよ、元の文書を作成、修正、および／または閲覧するアプリケーションに固有のものである。大半の印刷機は、そのような固有のフォーマットを直接理解できない。

【００２１】

元の文書をＰＤＬフォーマットに変換した後、ＰＤＬデータ２０は、選択された印刷機１４に送信され、必ずしもそうではないが、各ページの画像表現を含み得る。ＰＤＬフォーマットどうしは構文および表現が異なるが、すべてのＰＤＬフォーマットは、コンピュータグラフィクスプリミティブに基づいている。これらのプリミティブは、その最終的なゴールが、この場合紙である「表示面」上にグラフィックエレメントを描画するというものである、プログラム命令である。このようなプリミティブは、様々なフォーマットのビットマップを入力画像とし、これらを紙の上にレンダリングさせる、画像描画プリミティブを含む。他のプリミティブは、例えば、線分、円、四角形、およびベジェ曲線などの幾何形状を表現するために使用されるグラフィックプリミティブを含む。ＰＤＬの高度化レベルに応じて、これらの命令はまた、なんらかの形態のテクスチャ表現（すなわち、幾何学的形状を埋めるために使用されるパターン）ならびに様々な線の属性（例えば、色、幅、および線種）も含み得る。また、テキスト文字列（「ラン」ともいう）のグリフを描画

するために使用される、テキスト描画プリミティブも含む。一部の言語は、描画プリミティブの挙動を修正するために使用される、非描画プリミティブを含むこともある。一般的に、非描画プリミティブは、縦向きから横向きへと印刷の向きを切り替えるために使用される、行列演算やページ回転演算子などであり、座標系を修正するために使用される。非描画プリミティブはまた、様々な装置制御演算子、例えば、出力トレイ選択または仕上げオプションも修正できる。

【００２２】

大半のＰＤＬフォーマットにはテキスト描画演算が存在するが、だからといって、以下に詳述するように、所与の印刷物シート上のすべてのテキストがテキスト命令によって描画されているとは言えない。しかしながら、テキスト命令が使用されるならば、レンダリング装置１４は、いかなる他のグラフィクスでもなくテキストを印刷していることを一般的に把握している。

【００２３】

多くのケースでは、テキスト演算子は、テキストラン（すなわち、単語の境界にかかり得る、あるいはかかり得ない、文字シーケンス）を構成する実際の文字ではなく、紙などの物理的媒体２８上にレンダリングされるグリフの順番に割り当てられた識別子を引数とすることによって作用する。例えば、印刷機１４は、次の疑似コード：ｏｐｅｒａｔｏｒＤｒａｗＴｅｘｔ（「ｈｅｌｌｏ」）といった命令を受信するのではなく、ｏｐｅｒａｔｏｒＤｒａｗＴｅｘｔ（１２，１，１５，１５，２１）」などの、文字のグリフコードを識別する命令を受信し得る。この例では、１２は、その時使用中のフォントで文字「ｈ」を表現するグリフのインデックスであり、１は「ｅ」のインデックスであり、１５は「ｌ」のインデックスであり、２１は「ｏ」のインデックスである。

【００２４】

対応する文字をレンダリングするためのインデックスおよび命令を格納する、グリフインデックス３０を、例えばＰＤＬファイル２６で印刷機１４に送信してもよい。このインデックスは、フォントに固有でないことが多く、したがって、簡素な、インデックスをグリフにする対応表を使用できない。大半の印刷ドライバは、ＰＤＬ文書２０を生成する際に、グリフインデックス３０を徐々に構築する。一般的に、インデックスは、ジョブの際、ある文字を初めて所与のフォントで描画しなければならない時はいつでも割り当てられる。例えば、上記の例では、「ｅ」は、ジョブの際、選択されたフォントに関して、表示される最初の文字になるため、インデックス１を得、以下、ｈは１２番目であるためインデックス１２を得るなどとなる。このことから明らかなように、テキスト演算子をフックすることができる場合であっても、グリフインデックスだけからでは、描画される文字に戻る実質的な方法はない。

【００２５】

単一のテキスト描画演算子によって処理されたとしても、グリフは、多くのフォーマットで格納できることに留意されたい。典型的なフォーマットとしては、純粋なビットマップ、ＴｒｕｅＴｙｐｅ輪郭（これは、基本的には、グリフ曲線を描画するために使用されるベジェ点セットである）、および／またはＰｏｓｔＳｃｒｉｐｔ　ＰＤＬとＰＤＦなどのそのバリエーションとで使用されるＰｏｓｔＳｃｒｉｐｔ命令が挙げられる。このケースでは、グリフ記述に組み込まれたＰｏｓｔＳｃｒｉｐｔ命令サブセットを用いてグリフ曲線を描画する。他のバリエーションとしては、例えば、Ａｄｏｂｅ　Ｔｙｐｅ２フォント、ＣＦＦフォント、およびＣｈａｍｅｌｅｏｎフォントが挙げられ、これらのフォーマットは、独自仕様であってもよいし、そうでなくてもよい。

【００２６】

図示の直接文字認識システム１０は、本例示的な方法を行うための命令１８を格納するメモリ３１と、命令を実行するためにメモリと通信可能なプロセッサ３２（または複数のプロセッサ）とを含む。システムは、サーバコンピュータ３４などの１つ以上の計算装置にわたって分散させてもよい。他の実施形態では、システム１０は、計算装置１２もしくは印刷機１４でホストされてもよいし、ネットワーク１上に分散させてもよい。１つ以上

の入出力（「Ｉ／Ｏ」）装置３６、３８によって、システムがコンピュータ１２および／または印刷機１４などの外部装置と通信できる。システムのハードウェアコンポーネント３１、３２は、データ／制御バス４０によって通信可能に接続される。コンピュータ１２、印刷機１４、および／またはサーバコンピュータ３４との間で送受信される情報には、使用されるプロトコルに従って、データ、コマンド、言語データファイルのロケーションおよびフォーマッティングに関する情報、ケーパビリティ要求、ステータス要求、応答、および／または肯定確認を含むことができる。

【００２７】

例示的な命令１８は、取り込みコンポーネント４２と、解析システム４４と、テキスト要約コンポーネント４６と、ログ取得コンポーネント４８と、キャッシュコンポーネント５０と、検証コンポーネント５２と、テキスト処理コンポーネント５４と、出力コンポーネント５６とを含むが、提供されるコンポーネントは、より少なくてもよいし、より多くてもよいし、または異なっていてもよい。

【００２８】

取り込みコンポーネント４２は、印刷ドライバ２２と印刷機１４との中間の、ネットワーク１上、例えば、サーバコンピュータ３４上、またはサーバコンピュータと通信可能に接続される別個の計算装置上に配置される。取り込みコンポーネントは、ＰＤＬファイル２６を、これが印刷機に向かう途中で捕捉し、ＰＤＬ文書２０のコピーとグリフインデックス３０とを解析システムに転送する。解析システム４４は、ＰＤＬフォーマットに従って取り込んだＰＤＬ文書２０を解析する。

【００２９】

解析システム４４は、ＰＤＬフォーマットに従ってＰＤＬ文書２０の各グリフのすべてのテキスト関連演算子を捕捉し、直接文字認識を行う。図示の解析システムは、ＴｒｕｅＴｙｐｅフォントデータベース６２およびグリフ特徴データベース６４のうちの少なくとも１つを含む、参照データベース６０にアクセスできる。

【００３０】

一部の実施形態では、解析システム４４は、図３に示すように構成され得る。この実施形態では、解析システムは、解析器７８と、比較コンポーネント８０と、抽出コンポーネント８２と、レンダリングコンポーネント８４と、特徴抽出部８６と、類似度コンポーネント８８とを含む。ただし、含まれているコンポーネントは、より少なくてもよいし、より多くてもよいし、または異なっていてもよい。

【００３１】

ＴｒｕｅＴｙｐｅフォントデータベース６２は、既定の文字セットのそれぞれに関して、ＴｒｕｅＴｙｐｅ輪郭のリストを含み得る。例えば、ＴｒｕｅＴｙｐｅデータベース６２は、多数の、例えば、少なくとも１０の、または少なくとも１００、または少なくとも１０００の入手可能なＴｒｕｅＴｙｐｅフォントから生成されたグリフ点座標の集まりであり得る。ＴｒｕｅＴｙｐｅデータベース６２に格納されたＴｒｕｅＴｙｐｅフォントは、多くのビジネス環境で使用されるフォントの大半を表現する、一般的な事務文書において一般的に利用されているフォントである。ＴｒｕｅＴｙｐｅフォントにおけるグリフ９０のアウトラインは、図４に示すように、直線線分と２次ベジェ曲線点とから作られる。

【００３２】

グリフ特徴データベース６４は、既定の文字セットのそれぞれに関して、ストックフォントから生成されたグリフ特徴セットを含み得る。よって、グリフ特徴データベース６４は、予め計算したベクトルの集まりを含み得る。予め計算した特徴は、標準的なストックフォントから生成できる、または顧客固有のフォントに関して生成できる。グリフ特徴データベース６４を作成するためには、図３に示すように、入力グリフがそれぞれ、各入力グリフの最外輪郭を正確に取り囲むビットマップ９２としてレンダリングされる。ビットマップは、バイナリ値を有してもよいし、非バイナリ値を有してもよい。文字「ａ」のビットマップ９２は、図４に描かれ、アレイの形態を取っている。各要素の値は、画像のその部分の色に対応する。文字「ａ」は、１２×１４のマトリクスで表現でき、ここで、マ

トリクス内の値は、ピクセル（画素）の輝度を示している。より大きな値は、より明るい領域に対応し、より低い値はより暗い領域に対応する。連続する同じ色のピクセルの数の計数、すなわちランレングス特徴は、各グリフに関してビットマップから抽出され、特徴ベクトルとしてグリフ特徴データベース６４に格納され得る。

【００３３】

一部の実施形態では、解析器７８は、個々のテキスト文字を表現するグリフなどの個々のオブジェクトを識別し得る。グリフおよび関連しているテキスト文字は、一般的に、ＰＤＬ文書２０内においてテキスト描画プリミティブで符号化される。テキスト描画プリミティブは、テキストランにおけるグリフを描画するために使用される描画命令セットである。解析器７８は、例えばＰｏｓｔＳｃｒｉｐｔ（「ＰＳ」）またはＰｒｉｎｔｅｒ Ｃｏｍｍａｎｄ Ｌａｎｇｕａｇｅ（「ＰＣＬ６」）などの、使用されるＰＤＬフォーマットに従ってＰＤＬ文書２０を解析でき、すべてのテキスト関連演算子またはグリフ描画命令を捕捉できる。他のＰＤＬフォーマットとしては、限定するものではないが、ＰＣＬ５、ＢＢＪＬ、Ｐｏｒｔａｂｌｅ Ｄｏｃｕｍｅｎｔ Ｆｏｒｍａｔ（「ＰＤＦ」）、およびＸＭＬ Ｐａｐｅｒ Ｓｐｅｃｉｆｉｃａｔｉｏｎ（「ＸＰＳ」）が挙げられる。

【００３４】

直接文字認識のための一実施形態では、テキスト演算子文字列引数における各グリフに関して、解析器７８は、グリフインデックス３０に基づいてグリフ形状を取得する。多くの文書がＴｒｕｅＴｙｐｅフォントを利用していることから、グリフインデックス３０は、ＴｒｕｅＴｙｐｅフォントを使用し得る。本実施形態では、グリフ識別器７８が、ＰＤＬ文書２０におけるＴｒｕｅＴｙｐｅグリフに関する各描画命令を識別する。取り込んだＰＤＬ文書２０から取得したグリフ形状がＴｒｕｅＴｙｐｅ輪郭リストである場合、比較コンポーネント８０は、描画命令を、グリフ特性のＴｒｕｅＴｙｐｅデータベース６８と比較して、ＴｒｕｅＴｙｐｅフォントデータベース６２に一致するリストがあるかどうかを判定する。特に、比較コンポーネント８０は、ＰＤＬ文書２０の描画命令にある入力グリフの直線線分および２次ベジェ曲線点を、ＴｒｕｅＴｙｐｅデータベース６２に格納された各文字９０の直線線分および２次ベジェ曲線点と比較する。点のリスト間で一致するものが見つかった場合、テキスト文字が見つかり、抽出コンポーネント８２によって抽出できる。

【００３５】

直接文字認識のための別の実施形態では、例えば、取得したグリフ形状がＴｒｕｅＴｙｐｅフォントではない（例えば、解析器７８が、ＴｒｕｅＴｙｐｅデータベース６８にあるＴｒｕｅＴｙｐｅ輪郭ではない、グリフ形状に関する描画命令を識別する）場合、または一致するものが見つからなかった場合、グリフ特徴データベース６４にアクセスできる。本実施形態では、レンダリングコンポーネント８４は、ストックフォントに関しては、グリフを、その輪郭を正確に取り囲むビットマップ９２（図３）としてレンダリングする。例えば、ビットマップの各ピクセルは、「ｏｎ」または「ｏｆｆ」のバイナリ値を有し、「ｏｎ」は、グリフの輪郭内に実質的に含まれるピクセルを表し、「ｏｆｆ」は背景ピクセルを表す。特徴抽出部８６は、レンダリングコンポーネントによって生成されたビットマップを受信し、データベース６４のストックフォントと同様の仕方で、生成されたビットマップ９２から特徴セットを、例えば、１つ以上の特徴ベクトルの形態で抽出する。特徴ベクトルは、ビットマップからベクトルとして収集された、連続する同じ色のピクセルの数の計数、すなわちランレングス特徴であり得る。特徴ベクトルは、図４に示すアレイと同様に、アレイとして格納され得る。類似度コンポーネント８８は、グリフ特徴データベース６４の文字セットのそれぞれに関して、抽出された特徴ベクトルとグリフ特徴データベース６４にある特徴ベクトルとの間の類似度を計算する。類似度の計算は、抽出された特徴ベクトルとグリフ特徴データベース６４に格納された、予め計算されたベクトルの集まりとの間の距離の計算であり得る。例えば、ユークリッド距離またはバッタチャリャ距離が計算され得る。ただし、他の標準的な特徴抽出または距離の方法を代替的に使用できる。

【００３６】
ビットマップから抽出されたグリフ特徴セットとの類似度が閾値を超える、少なくとも１つの特徴セットがデータベース６４において見つかった場合、グリフ特徴セットに対して最も類似の特徴セットを有する文字が識別される。これで、文字が見つかり、抽出できる。したがって、ＰＤＬ文書２０の描画命令において符号化されたグリフ特徴とグリフ特性のデータベースとの間で一致するものが見つかった場合、類似度コンポーネントによって計算された距離の計算に基づいて、抽出コンポーネント８２は、最も類似のグリフに関連付けられたテキスト文字を抽出する。最初未知の入力グリフ形状とグリフ特徴データベース６４との間におけるこの対応は、キャッシュコンポーネント５０によってキャッシュ９６に格納され得る。この次に解析器４４が、同じ入力グリフ形状に関して、ＰＤＬ文書２０において同じ描画命令を識別した時には、システム１０は、特徴抽出および距離比較に頼らずとも、キャッシュを使用して、グリフ特徴データベース６４において直接一致するものを判定し得る。

【００３７】
文字が識別された時または一度解析器システム４４がＰＤＬ文書２０の解析を完了しすべてのテキスト文字を抽出した時、各識別された文字に関して、要約コンポーネント４６は、各グリフに関連付けられた抽出されたテキスト文字のテキスト要約９２を生成する。所与のテキスト関連演算子に関して解析器によって識別されたそれぞれの一致するものに関して、テキスト要約コンポーネント４６は、抽出された文字と文書ページ内の場所とをテキスト要約に加える。テキスト要約９２は、計算装置１２などの計算装置、および／または装置１２を操作するユーザによって読み出し可能とするフォーマットであり得る。テキスト要約９２は、拡張マークアップ言語（「ＸＭＬ」）ファイルとして格納され得る。

【００３８】
一致するものが見つからなかった場合、所与の場所の候補の文字が識別できなかったという事実が解析器システムによって記録され得る。特に、なんらテキスト文字が認識されない場合、ログ取得コンポーネント４８は、見つけ損ねたテキスト文字の記録をつける。ログ取得コンポーネントは、エラーログ９４に見つけ損ねたテキスト文字を格納できる。エラーログ９４は、例えばテキスト（「ＴＸＴ」）ファイルとして、メモリ３１に格納できる。

【００３９】
検証コンポーネント５２は、識別された文字シーケンスにおける候補の単語を識別し、辞書９８にアクセスして、候補の単語が存在するかを判定する。存在する場合、候補の単語が検証され、検証された単語シーケンスが出力され得る。

【００４０】
処理コンポーネント５４は、検証された単語シーケンスの少なくとも一部を処理し、それに基づいて決定をレンダリングし得る。例えば、処理コンポーネントは、１つ以上のキーワードを検索し、文書が印刷許可されているかどうか、および／またはその印刷に関して顧客に課金するかどうかなどを決定する。

【００４１】
出力コンポーネント５６は、検証された単語シーケンス、および／またはあらゆるエラー／検証されていない単語、および／または処理コンポーネント５４によって出力された決定などの情報をシステム１０から出力する。文書を印刷すべきである／すべきではないという決定の場合、情報が、印刷機に送信されて、印刷を許可／阻止し得る。決定が、課金すべき顧客である場合、情報は、課金システム（図示せず）に送信され得る。他の実施形態では、情報は、処理のために別のコンピュータ装置に送信され得る識別された単語シーケンスであり得る。

【００４２】
ネットワークリンク１６は、サブネット、ローカルエリアネットワーク（ＬＡＮ）、および／またはインターネットを含み得る。

【００４３】

入出力コンポーネント３６、３８はそれぞれ、変復調器（モデム）、ルータ、ケーブル、イーサネットポート、および／またはネットワーク１に接続された周辺機器が、例えばネットワーク管理者によって設定されたポリシーに従って有線または無線接続を介して他の装置と通信できるようにする他の通信装置（図示せず）を含み得る。計算装置１２は、１つ以上のネットワーク連携システム１６を介して接続される、複数のＰＣまたは複数のワークステーションなどの複数の装置を含み得ることに留意されたい。

【００４４】

印刷機または複数の印刷機１４は、レーザ印刷機、インクジェット印刷機、ＬＥＤ印刷機、プロッタ、および／またはインクまたはトナーなどのレンダリング媒体を用いて紙などの物理的媒体上に画像をレンダリングできる、任意の他の装置であり得る。印刷機１４は、コンピュータ印刷機、ファクシミリ装置、デジタル複写機、多機能装置、および／または文書を印刷できる他の装置の形態を取り得る。

【００４５】

接続部１６は、計算装置１２と印刷機１４とをネットワーク１に接続する。接続部１６は、適切な従来の通信プロトコルおよび／またはデータポートインターフェースを用いる有線または無線接続部として実装され得る。一般的に、接続部１６は、装置間のデータ伝送を可能にする任意の通信チャネルであり得る。一実施形態では、例えば、装置には、適切な接続部１６を通してデータを伝送するための、ＵＳＢ（商標）、ＳＣＳＩ、ＦＩＲＥＷＩＲＥ（登録商標）、および／またはＢＮＣポートなどのデータポート３６を設けることができる。通信リンクは、無線リンクもしくは有線リンクまたは計算装置１２と印刷機１４との間の通信を可能にする任意の組み合わせであり得る。

【００４６】

印刷機１４は、ハードウェア、ファームウェア、もしくはソフトウェア、またはこれらの組み合わせによって制御され得る。ＰＤＬ文書２０からの直接文字認識のためのシステム１０は、例示的なコンピュータ１２および／または印刷機１４、あるいは図２に示すような別個の装置のうちの１つ以上に配備され得る。例えば、印刷機１４は、開示される実施形態と整合する仕方で印刷機１４が直接文字認識の処理を最適化できるようにするソフトウェアまたはファームウェアを実行し得る。別の実施形態では、システム１０は、コンピュータ１２に常駐し、印刷機１４のためのＰＤＬデータに対して動作するのであってもよい。一般的に、システムコンポーネントは、システム内の１つ以上のコンピュータ１２、３４および／または印刷機１４上において全体的または部分的に実施され得る。

【００４７】

デジタルプロセッサ３２は、汎用プロセッサ、専用プロセッサ、または組み込みプロセッサであり得る。プロセッサ３２は、制御情報と命令とを含むデータをメモリ３１と交換できる。メモリ３１は、ＳＤＲＡＭまたはＲＤＲＡＭなどの、任意のタイプのダイナミックランダムアクセスメモリ（「ＤＲＡＭ」）および／または読み出し専用メモリ（ＲＯＭ）であり得る。命令１８は、限定されるものではないが、ブートアップシーケンス、ＰＤＬの解析、プログラミング言語のためのコンパイラ、自動コード生成ルーチン、解釈されたページ記述言語における機能ベースのオブジェクト操作の処理のための最適化ルーチンなどのＰＤＬを用いて記述された文書を処理するルーチン、受信する要求およびメッセージを処理するためのルーチン、発信する応答およびメッセージを作成するためのルーチンを含む、１つ以上の既定のルーチン、ならびに構成管理のためのルーチン、文書処理のためのルーチン、および他のコードを含む命令を保持し得る。一部の実施形態では、命令１８のコードは、プロセッサ３２によって作動させられる前にメモリ３１にコピーされ得る。任意のＰＤＬ処理ルーチンおよび最適化ルーチンを含む命令１８は、コンピュータ１２とネットワークリンク１６とのうちの１つ以上を用いてアップグレード可能であり得る。

【００４８】

一部の実施形態では、計算装置１２は、印刷ドライバ２２を介して印刷機１４に、ＰＤＬを用いて指定した文書２０の印刷可能データを送信し得る。印刷機１４は、ルーチンを呼び出して、ＰＤＬ文書２０を解析器４４で解析する。取り込みコンポーネント４２は、

印刷ドライバ２２によってＰＤＬ文書２０が生成された後であるが、印刷のために印刷機１４に送信される前に、ＰＤＬ文書２０を捕捉する。解析器４４は、ＰＤＬデータ２０にある様々なオブジェクト、演算子、および構造を識別し、認識されたオブジェクト、演算子、および／または構造に関連付けられた動作を行うまたはアクションを開始することができる。

【００４９】

図５は、図２のコンピュータネットワークにおいて行われ得る、ＰＤＬ文書からの直接文字認識のための例示的な方法を示している。本方法は、Ｓ１００で開始する。Ｓ１０２において、テキスト文字のそれぞれに関連付けられたグリフ特性セットによって定義されるテキスト文字の参照データベース６０へアクセスできるようにする。参照データベースは、メモリに格納され得る。テキスト文字の参照データベースは、上述のＴｒｕｅＴｙｐｅフォントデータベースおよび／またはグリフ特徴データベースを含むことができる。

【００５０】

Ｓ１０４において、印刷ドライバによって、ＰＤＬフォーマットにおいて指定された文書の印刷ジョブが開始される。

【００５１】

Ｓ１０６において、ＰＤＬ文書２０が図１および図２に示すシステム１０によって受信または取り込まれる。ＰＤＬ文書２０は、印刷文書において一連のグリフから形成されたテキストランを描画するための描画命令セットを含み得る。

【００５２】

Ｓ１０８において、ＰＤＬ文書２０が解析されて、複数のグリフからなるテキストに関する任意の描画命令を捕捉または識別する。描画命令は、「ラン」としても知られるテキスト文字列のグリフを描画するために使用されるテキスト描画プリミティブの形態であり得る。ＰＤＬ解析は、印刷ドライバの出力と印刷機での文書の印刷との間にあるパイプライン上の任意の場所で実行され得る。ＰＤＬ解析は、印刷ドライバによってＰＤＬ文書が生成された後、例えば、印刷ドライバ変換の最後の段階であるが、コンピュータまたは印刷サーバのポートモニタで文書が印刷される前に、実行され得る。必要に応じて、ＰＤＬ文書は、例えばスイッチまたはルータを介してネットワークから直接取り込まれてもよい。このことは、ネットワークが暗号化されていない場合または暗号化システムが分かっている場合に可能である。

【００５３】

Ｓ１１０において、Ｓ１０８において識別した描画命令を、グリフ特性のデータベースと比較して、描画命令によって表現されるテキスト文字を見つける。解析するステップが、ＴｒｕｅＴｙｐｅグリフの描画命令を識別する場合、描画命令を、グリフ特性のＴｒｕｅＴｙｐｅデータベースと比較する。上述のように、ＴｒｕｅＴｙｐｅデータベースは、単純に、数千の入手可能なＴｒｕｅＴｙｐｅフォントから容易に生成され得るグリフ点座標の集まりである。入力グリフ点は、一致するものが見つかるまで参照データベースのレコードと比較される。有利なことに、多くのビジネスでは、一般的な事務文書にはＴｒｕｅＴｙｐｅフォントを使用している。必要に応じて、ＴｒｕｅＴｙｐｅデータベースは、グリフ検索をさらに加速するために、顧客固有のフォントと補足されまたは置き換えられ得る。

【００５４】

Ｓ１１２において、描画命令とグリフ特性のデータベースにあるテキスト文字との間で一致するものが見つかった場合、次いで、Ｓ１１４において、テキスト文字が抽出される。そうでない場合、本方法は、Ｓ１１６に進んでよい。

【００５５】

ＴｒｕｅＴｙｐｅグリフ以外の他の技法が使用される場合、またはグリフ技法がＴｒｕｅＴｙｐｅであるが、データベースの所与の文書グリフに関して、一致するＴｒｕｅＴｙｐｅ点がない場合、グリフ形状が、さらなる処理のために、ビットマップとしてレンダリングされる（Ｓ１１６）。

【００５６】
Ｓ１１８において、ビットマップに基づいて特徴ベクトルが抽出される。例えば、ビットマップから、連続する同じ色のピクセルの数を計数し、ランレングス特徴ベクトルとして格納する。
【００５７】
Ｓ１２０において、抽出された特徴ベクトルとグリフ特徴データベースとの間の類似度が計算される。上述のグリフ特徴データベースは、標準的なストックフォントから容易に生成され得る、または必要に応じて顧客固有のニーズのために生成され得る、予め計算されたベクトルの集まりである。Ｓ１２０の類似度計算は、自動画像分類ソフトウェアにおいて使用されるものなどの従来の特徴距離計算を用いる、単純な画像類似度予測であり得る。例えば、類似度の計算は、抽出された特徴ベクトルとグリフ特徴データベースに格納された、予め計算されたベクトルの集まりとの間の、ユークリッド距離またはバッタチャリャ距離などの距離の計算であり得る。テキスト文字の単純な単色の形状に他の標準的な特徴抽出または距離の方法を使用してもよいが、ランレングスベクトルおよびユークリッドもしくはバッタチャリャ距離は、非常に良好な結果をもたらすのに十分であることが分かっている。
【００５８】
Ｓ１２０において、計算された距離の計算結果に基づいて、入力グリフ形状に十分に類似した少なくとも１つの特徴セットが見つかった場合、本方法は、Ｓ１１４に進んで、類似度に基づいて入力グリフ形状に関連付けられたテキスト文字を抽出する。場合により、Ｓ１２２において、最初未知の入力グリフ形状とグリフ特徴データベースとの間におけるこの対応は、キャッシュに格納され得る。この次にＳ１０８における解析が、同じ入力グリフ形状に関して、ＰＤＬ文書２０において同じ描画命令を識別した時には、システム１０は、Ｓ１１８の特徴抽出およびＳ１２０の類似度計算に頼らずとも、グリフ特徴データベースにおいて直接一致するものを判定し得る。
【００５９】
Ｓ１２０において、一致するテキスト文字が見つからない場合、本方法はＳ１２４に進み、ここで、エラーのログが取られる。見つけ損ねたテキスト文字は、例えばＴＸＴファイルとしてメモリにログを取られ、格納され得る。
【００６０】
本方法は、Ｓ１１４からＳ１２６に進み、ここで、Ｓ１１４において抽出されたテキスト文字とＳ１２４においてログを取ったエラーとに基づいてテキスト要約が生成される。生成されるテキスト要約は、システム１０を操作する機械およびユーザの両方に可読であり得、ＸＭＬフォーマットで格納され得る。
【００６１】
場合により、Ｓ１２８において、抽出されたテキスト文字を含む、Ｓ１２６で生成されたテキスト要約は、辞書を用いてクロスチェックされて、抽出されたテキスト文字から形成された単語を検証し得る。例えば、ＰＤＬ文書は、大文字の「ｉ」の描画命令と同じものになる、小文字の「Ｌ」の描画命令を含むことがある。これが当てはまるのは、両方のテキスト文字がまっすぐな垂直の線「Ｉ」で描画されるフォントタイプの場合である。この場合、辞書を用いてクロスチェックをすることにより、無効な文字認識を抑制し、Ｓ１２６において生成されるテキスト要約において正しいテキスト文字が表現されることを確実にすることによって両義性を低下させることができる。必要に応じて、一般的な辞書または専門辞書を用いてクロスチェックを行って、抽出されたテキスト文字から形成された特定の単語を認識することができる。専門辞書は、専門的なテキスト文字および単語の認識を要求している顧客から提供され得る。Ｓ１２４においてログを取ったエラーは、候補の単語を、辞書と比較する際には「任意の文字」とみなされ得る。
【００６２】
Ｓ１３０において、場合によって検証されるテキスト要約などの情報が出力される。
【００６３】

本方法は、Ｓ１３２で終了する。

【００６４】

本方法は、説明される機能を行うためのソフトウェアで実装され得る。例示的な疑似コードを以下のアルゴリズム１に示す。

【００６５】

【表１】

アルゴリズム１
テキスト演算子「文字列」引数における各グリフに関して ｛引数として与えられるグリフインデックスに基づいてグリフ形状を取得する Ｉｆ（グリフ形状がＴｒｕｅＴｙｐｅ輪郭リストである） ｛ＴｒｕｅＴｙｐｅデータベースにおいて一致するものを探す Ｉｆ（一致するものが見つかった） ｛文字が発見された｝｝ Ｉｆ（グリフ形状がＴｒｕｅＴｙｐｅ輪郭リストではないまたは一致するものがない） ｛グリフをその輪郭を正確に取り囲むビットマップとしてレンダリングする与えられたビットマップの特徴を抽出する 参照特徴データベースを用いてこれらの特徴の類似度を計算する Ｉｆ（候補に十分に近い少なくとも１つの特徴セットが見つかった） ｛文字が発見された。これは、いかなる参照特徴よりも最も近い｝｝ Ｉｆ（一致するものがない） ｛所与の場所で文字を見つけ損ねたという事実の痕跡を保持する｝｝

【００６６】

図５に示す本方法は、コンピュータ上で実行され得るコンピュータプログラム製品に実装され得る。コンピュータプログラム製品は、ディスク、ハードドライブなどの、制御プログラムが記録される（格納される）非一時的コンピュータ可読記録媒体を含み得る。一般的な形態の非一時的コンピュータ可読媒体は、例えば、フロッピーディスク、フレキシブルディスク、ハードディスク、磁気テープまたは任意の他の磁気記憶媒体、ＣＤ－ＲＯＭ、ＤＶＤ、または任意の他の光学式媒体、ＲＡＭ、ＰＲＯＭ、ＥＰＲＯＭ、ＦＬＡＳＨ－ＥＰＲＯＭ、または他のメモリチップもしくはカートリッジ、あるいはコンピュータが読み出しおよび使用し得る任意の他の有形の媒体を含む。

【００６７】

あるいは、本方法は、音波もしくは光波など、ならびに無線および赤外線データ通信中に生成される媒体などの伝送媒体を使用して、制御プログラムがデータ信号として埋め込まれている送信可能な搬送波などの一時的媒体に実装され得る。

【００６８】

本例示的な方法は、１つ以上の汎用コンピュータ、専用コンピュータ（複数可）、プログラムされたマイクロプロセッサもしくはマイクロコントローラおよび周辺集積回路素子、ＡＳＩＣもしくは他の集積回路、デジタル信号プロセッサ、離散要素回路などのハードウェアにより実現されている電子回路または論理回路、ＰＬＤ、ＰＬＡ、ＦＰＧＡ、グラフィカルカードＣＰＵ（ＧＰＵ）、またはＰＡＬなどのプログラマブルロジックデバイスなどの上に実装され得る。一般的に、図５に示す流れ図を実施できる任意の装置を用いて、本方法を実施できる。

【図１】

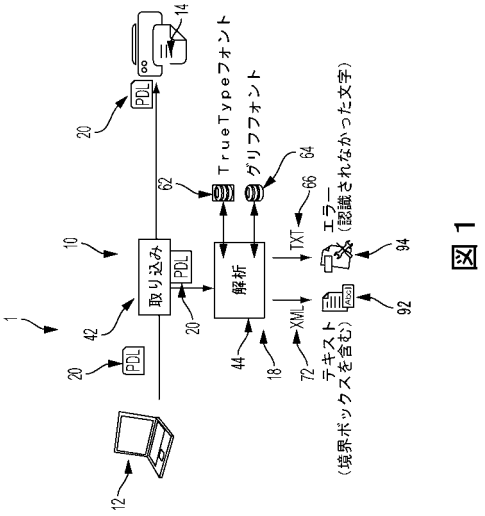

図１

【図２】

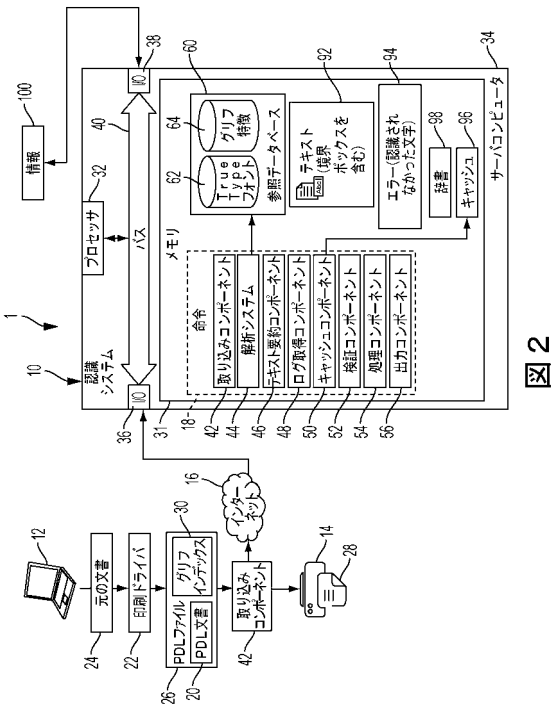

図２

【図３】

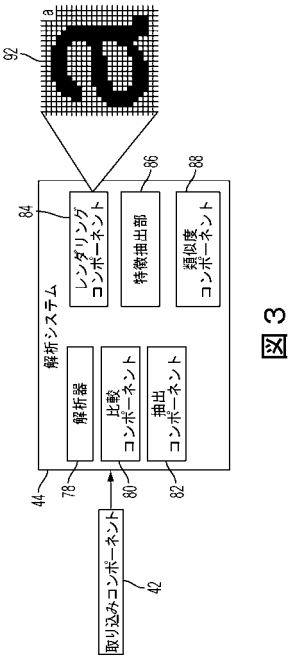

図３

【図４】

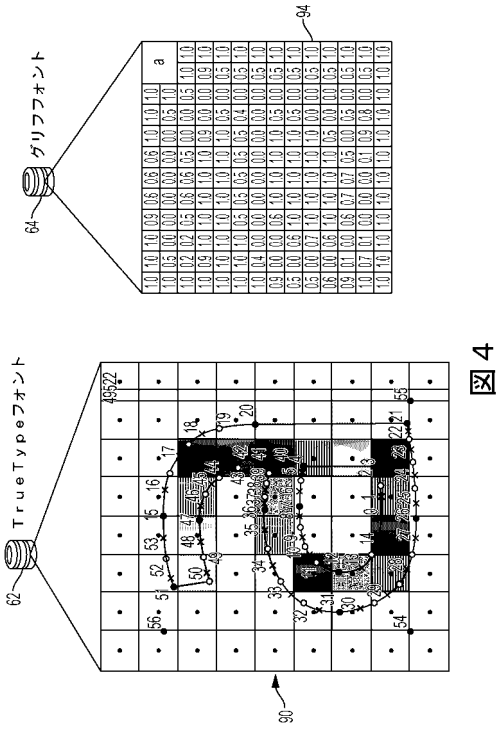

図４

【図５】

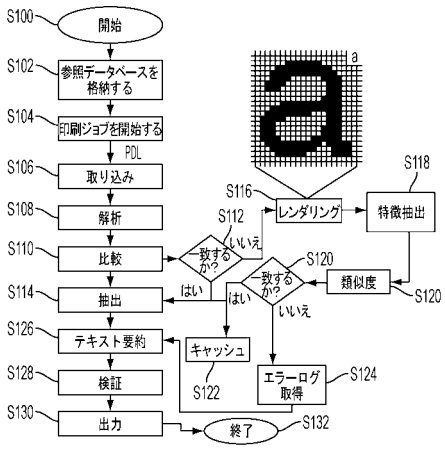
S100
開始
S102
参照データベースを格納する
S104
印刷ジョブを開始する
PDL
S106
取り込み
S108
解析
S110
比較
S114
抽出
S126
テキスト要約
S128
検証
S130
出力
S112
一致するか?
いいえ
はい
S116
レンダリング
S118
特徴抽出
S120
一致するか?
類似度
S120
はい
いいえ
キャッシュ
S122
エラーログ取得
S124
終了
S132

図５

フロントページの続き

(51)Int.Cl.	FI		テーマコード（参考）
	Ｇ０６Ｆ 3/12	３４５	
	Ｇ０６Ｆ 3/12	３７３	
	Ｂ４１Ｊ 29/38	Ｚ	

(72)発明者 イヴ・オプノ
フランス共和国 ノートルーダムードーメザージュ ３８２２０ ル・シャンボール １５２

Ｆターム(参考) 2C061 AP01 AQ05 AQ06 AR01 AR03 HK03
2C187 AC06 AC08 AE07 BG03 BG19 BG49 CD04 JA07
5B064 AA10 DC26 EA08 EA19