

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5771108号

(P5771108)

(45) 発行日 平成27年8月26日 (2015. 8. 26)

(24) 登録日 平成27年7月3日 (2015. 7. 3)

(51) Int. Cl. F I
G 0 6 K 9 / 0 3 (2006. 01)
 G 0 6 K 9 / 0 3 Z
 G 0 6 K 9 / 0 3 B
 G 0 6 K 9 / 0 3 J

請求項の数 17 (全 28 頁)

(21) 出願番号	特願2011-216500 (P2011-216500)	(73) 特許権者	390009531
(22) 出願日	平成23年9月30日 (2011. 9. 30)		インターナショナル・ビジネス・マシーンズ・コーポレーション
(65) 公開番号	特開2013-77157 (P2013-77157A)		INTERNATIONAL BUSINESS MACHINES CORPORATION
(43) 公開日	平成25年4月25日 (2013. 4. 25)		アメリカ合衆国10504 ニューヨーク州 アーモンク ニュー オーチャードロード
審査請求日	平成26年5月7日 (2014. 5. 7)	(74) 代理人	100108501 弁理士 上野 剛史
		(74) 代理人	100112690 弁理士 太佐 種一
		(74) 代理人	100091568 弁理士 市位 嘉宏

最終頁に続く

(54) 【発明の名称】 光学式文字認識で生成したテキストデータの校正を支援するためのシステム、方法、プログラム

(57) 【特許請求の範囲】

【請求項 1】

光学式文字認識で生成されたテキストデータの校正を支援するためのシステムであって、

前記テキストデータのカーペット校正を行うための第1の校正ツールと、
 前記テキストデータのサイド・バイ・サイド校正を行うための第2の校正ツールと、
 前記第1および第2の校正ツールで行った校正オペレーションのログを記録するための記憶部と、

前記第1の校正ツールのカーペット校正を行う単位である属性ごとに、前記ログに基づいて算出された前記第1および第2の校正ツールを使用して校正を行った場合に要する時間の第1の推測値と、第1の校正ツールを使用することなく第2の校正ツールを使用して校正を行った場合に要する時間の第2の推測値を比較して当該属性の校正に第1の校正ツールを使用するかどうかを決定する解析部と、

を備える、システム。

【請求項 2】

前記第1および第2の校正ツールを使用して校正を行った場合は、前記第1の校正ツールによる校正を行った後で、前記第2の校正ツールによる校正が行われる請求項1に記載のシステム。

【請求項 3】

前記カーペット校正を行う単位である属性が、文字である請求項1又は2に記載のシス

10

20

テム。

【請求項 4】

前記第 1 の校正ツールによる校正を行った後で、前記第 2 の校正ツールによる校正が行われ、前記第 1 の校正ツールによる校正において、文字認識の正誤が不明である場合に、不明であることを示すマークが付される、請求項 3 に記載のシステム。

【請求項 5】

前記不明であることを示すマークが付された文字を、カーペット校正において正誤不明であることを示すようにオペレータに提示する、請求項 4 に記載のシステム。

【請求項 6】

前記ログに基づいて、ログ統計情報を算出する算出部を備える、請求項 1 から 5 のいずれか 1 項に記載のシステム。

10

【請求項 7】

前記算出部が、各文字について、

【表 1】

a : これまでに処理した文字イメージのうち、OCRが誤認識され、
カーペット校正で誤認識を修正できた文字イメージの割合の
期待値の推定値

b : これまでに処理した文字イメージのうち、OCRが誤認識され、
カーペット校正で誤認識を修正しなかった文字イメージの
割合の期待値の推定値

20

u : 文字イメージ 1 つにつき OCR の正誤が不明であることを示す
マークを付するために要する時間の期待値の推定値

w : 文字イメージ 1 つにつき サイド・バイ・サイド 校正で修正に
要する時間の期待値の推定値

w' : 文字イメージ 1 つにつき カーペット 校正で修正に要する
時間の期待値の推定値

30

のうちの少なくとも 1 つを統計情報として算出することができる、請求項 6 に記載のシステム。

【請求項 8】

前記解析部が、前記統計情報に基づいて、

【表 2】

$\tilde{a}$:	これまでに処理した文字イメージのうち、OCRが誤認識され、カーベット校正で誤認識を修正できた文字イメージの割合の期待値の下限値の推定値
$\tilde{b}$:	これまでに処理した文字イメージのうち、OCRが誤認識され、カーベット校正で誤認識を修正しなかった文字イメージの割合の期待値の下限値の推定値
$\tilde{u}$:	文字イメージ1つにつきOCRの正誤が不明であることを示すマークを付するために要する時間の期待値の下限値の推定値
$\tilde{w}$:	文字イメージ1つにつきサイド・バイ・サイド校正で修正に要する時間の期待値の下限値の推定値
$\tilde{w}'$:	文字イメージ1つにつきカーベット校正で修正に要する時間の期待値の下限値の推定値

10

のうちの少なくとも1つを算出し、算出結果に基づいて当該文字の校正に第1の校正ツールの使用するかどうかを決定する、請求項7に記載のシステム。

20

【請求項9】

前記解析部が、前記該文字の校正に第1の校正ツールを使用するかどうかを、下記手順に従って決定する、請求項8に記載のシステム。

【数 1】

$L := \text{empty set, for each } x \text{ in } X,$

$$\tilde{a} \equiv \max \left[a - \sqrt{\frac{2 \log M}{M}}, 0 \right]$$

$$\tilde{b} \equiv \max \left[b - \sqrt{\frac{2 \log M}{M}}, 0 \right]$$

$$\tilde{u} \equiv \max \left[u - (u_U - u_L) \sqrt{\frac{2 \log N_U}{N_U}}, u_L \right]$$

$$\tilde{w} \equiv \max \left[w - (w_U - w_L) \sqrt{\frac{2 \log N_w}{N_w}}, w_L \right]$$

$$\tilde{w}' \equiv \max \left[w' - (w'_U - w'_L) \sqrt{\frac{2 \log N_{w'}}{N_{w'}}}, w'_L \right]$$

if $\tilde{b}\tilde{u} < \tilde{a}(\tilde{w} - \tilde{w}')$, add x to L

10

20

L : 対象リストに含まれる文字の集合

X : 文書に出現する可能性のある文字の集合

x : 文書に出現する文字の集合 X の要素

u_U : 文字イメージ 1 つにつき OCR の正誤が不明であることを示すマークを付するために要する時間の上限値(規定値)

u_L : 文字イメージ 1 つにつき OCR の正誤が不明であることを示すマークを付するために要する時間の下限値(規定値)

w_U : 文字イメージ 1 つにつきサイド・バイ・サイド校正に要する時間の上限値(規定値)

w_L : 文字イメージ 1 つにつきサイド・バイ・サイド校正に要する時間の下限値(規定値)

w'_U : 文字イメージ 1 つにつきカーベット校正で修正に要する時間の上限値(規定値)

w'_L : 文字イメージ 1 つにつきカーベット校正で修正に要する時間の下限値(規定値)

30

【請求項 10】

前記算出部が、複数の文字をクラスタリングし、クラスタごとにログ情報から統計情報を算出する、請求項 6 から 9 のいずれか 1 項に記載のシステム。

【請求項 11】

前記算出部が、(1) すべての文字に対する 1 つのクラスタの形成、(2) 統計情報の各パラメータの空間によるクラスタの形成、または、(3) 文字種ごとのクラスタの形成、のいずれかを実行することができる、請求項 10 に記載のシステム。

【請求項 12】

前記解析部の決定に基づいて前記属性を第 1 の校正ツールを使用する対象であることを記述するリストを備える、請求項 1 から 11 のいずれか 1 項に記載のシステム。

【請求項 13】

40

50

前記リストに基づいて、前記第1および第2の校正ツールの制御を行う制御部を備える、請求項12に記載のシステム。

【請求項14】

前記第1のツールが、コンテキスト付きカーペット校正ツールである、請求項1から13のいずれか1項に記載のシステム。

【請求項15】

前記第2の校正ツールが、ページ、文章、行のうちの少なくとも1つの単位でイメージデータとテキストデータを並べて提示する、請求項1から14のいずれか1項に記載のシステム。

【請求項16】

請求項1乃至15のいずれかに記載のシステムとしてコンピュータを機能させることができる、プログラム。

【請求項17】

光学式文字認識で生成されたテキストデータの校正を支援するための方法であって、前記テキストデータのカーペット校正を行うための第1の校正ツールを準備するステップと、

前記テキストデータのサイド・バイ・サイド校正を行うための第2の校正ツールを準備するステップと、

前記第1および第2の校正ツールで行った校正オペレーションのログを記録するステップと、

前記第1の校正ツールのカーペット校正を行う単位である属性ごとに、前記ログに基づいて算出された前記第1および第2の校正ツールを使用して校正を行った場合に要する時間の第1の推測値と、第1の校正ツールを使用することなく第2の校正ツールを使用して校正を行った場合に要する時間の第2の推測値を比較して当該属性の校正に第1の校正ツールを使用するかどうかを決定するステップと、

を含む、方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、一般的には、電子化されたテキストデータの校正技術に関し、より特定的には、光学式文字認識で生成したテキストデータの校正を支援するためのシステム、方法、プログラムに関する。

【背景技術】

【0002】

紙媒体の文書の保存、インターネット上への公開等の目的のために、既存の紙媒体の文書を電子化することが一般的に行われている。紙媒体の電子化は、単に紙媒体をイメージとして電子化する場合もあるが、後の編集や検索等の目的のために文書内の印刷テキストをデジタルのテキストに光学式文字認識（Optical Character Recognition。以下、「OCR」ということがある）を用いて変換する場合が多い。しかし、文字認識は完全に正確に行われるとは限らず、誤認識が生じることがある。そこで、誤認識を最小とするために、コンピュータにより自動的に認識されたテキストの正確性をオペレータがコンピュータ上で確認することが一般的に行われている。

【0003】

当分野の背景技術として、特開2003-099709号公報（特許文献1）は、認識結果修正画面上で誤読文字について文字修正がなされた際に、類似字形集合ファイルを参照して、中間ファイル内の上記修正された文字が含まれる類似字形のすべてに修正された文字を反映させ、一括して修正する処理を行う光学的文字認識装置を開示する。

【0004】

他の背景技術として、特開平07-057042号公報（特許文献2）は、オペレータが読取不能文字を効率良く変換することができるようにする文字読取装置を開示する。装

10

20

30

40

50

置では、文字データに変換できなかった読取不能文字がクラスタリングされ、各イメージ群に区分される。クラス化されたイメージ群は代表イメージとその他のイメージで表わされ、その候補文字が抽出される。代表イメージ群とそれに対応する候補文字に基づき、当該代表イメージ群に属する文字イメージと候補文字とを同一表示画面上に表示する。表示画面に基づいて、候補文字のいずれかが正解文字として選択されるか、新たに正解文字が入力されると、当該代表イメージ群に対応する各文字イメージ群を正解文字に対応する文字データに一括変換する。

【0005】

他の背景技術として、特開平09-179934号公報（特許文献3）は、オペレータの作業負担が小さい文字認識装置を開示する。装置は、印刷文書上の文字や図形を光学的に走査して、文字や図形を画像データとして読み込み、読み込んだ画像データの文字領域および図形領域を抽出し、文字領域について文字の認識処理を行い、認識結果を表示し、表示した認識結果が正しいかどうかの確認をオペレータに促す。該認識結果に誤りが発見された場合、オペレータの指示を受けて認識結果の誤りを修正・編集し、認識結果の全ての誤りを修正・編集して得た最終結果を出力する文字読取装置において、連続した複数行のそれぞれの行に、抽出処理の誤りによる文字の並びの誤りが存在する認識結果が表示された場合、修正・編集を、各行ごとに存在する誤った文字列から成る文字列群を指定し、文字列群に対して一括して行う。

【0006】

他の背景技術として、特開平06-290297号公報（特許文献4）は、装置は、特徴情報を位置情報とともに蓄積し、操作者が修正した修正履歴情報から画像特徴蓄積部の情報を抽出し、抽出された参照特徴と以降の特徴とのマッチングを行い、その結果に基づいて自動修正を行う文字認識装置を開示する。

【0007】

他の背景技術として、特開平05-314303号公報（特許文献5）は、オペレータの誤読修正作業を軽減し、修正時間の短縮化を実現する誤読文字修正方法を開示する。方法は、オペレータがキー操作などにより修正した文字または文字列を修正履歴として記憶し、OCRの読取結果について過去に修正された文字を検出し、誤読箇所をオペレータに強調表示し、修正候補文字を選択可能とする。

【0008】

他の背景技術として、特開平03-240183号公報（特許文献6）は、誤認識文字の確認修正作業の自動化が可能な文字認識結果の認識文字自動修正方式を開示する。文字認識結果に対する確認修正処理において、文字単位の認識を実行する文字認識手段と、文字認識結果を表示する手段と、確認の結果および修正指示を与える手段と、全文字認識結果を保持するテーブルと、正解文字候補が2位以下とされる誤認識に対して正解文字候補を1位とする修正が施された場合に他の文字位置に着いて類似の誤認識が生起しているかを識別する並べ替え判定手段と、前記の各手段の動作を制御する制御手段を備える。類似の誤認識が生起していると判断された文字位置に対して正解文字候補と推定されるも次候補の順位を1位とする入れ替えを行う。

【0009】

他の背景技術として、特開2003-223608号公報（特許文献7）は、認識結果の一括修正を可能とした上で、文字修正の精度の向上を図る認識文字列修正方法を開示する。方法では、認識結果のテキストデータの中から、作業者の操作指令に基づく任意の文字が選択し、その選択文字が、作業者の操作指令に基づいて置換文字に置き換えられる。次いで、テキストデータ上で処理対象の文字を選択文字から後方に移行して、選択文字と同一か、もしくは類似しているかの判定がなされる。同一かもしくは類似と判別された場合には、その文字を自動修正対象文字として、置換文字に一旦置き換える。その置換文字の前後についての構文を解析し、その構文が正しいと判別された場合には、置換文字への置き換えを正当なものとして確定する。

【0010】

他の背景技術として、特開 2 0 0 5 - 3 0 9 6 0 8 号公報（特許文献 8）は、同一カテゴリの文字イメージを並べた確認画面において、字形の近似する文字イメージをまとめて表示する文字認識装置を開示する。文字認識装置の出力機構において、文字認識処理の対象となる文字のイメージデータを、文字認識処理により認識された文字（カテゴリ）ごとに分類し、分類された各カテゴリのイメージデータに関して、そのイメージデータに含まれる文字の形状に関する特徴量を求め、この特徴量に基づきイメージデータをさらに複数のクラスタに分類し、クラスタごとにイメージデータを表示する確認画面を生成し表示出力する。

【0011】

他の背景技術として、松村季樹ほか（非特許文献 1）は、文字認識率等の認識精度に応じて最も効率の良い確認修正手段を、複数の手段の中から自動選択する入力方式を提案する。認識精度ごとに最大入力効率を示す確認修正手段を、測定によりあらかじめ認識精度と対応づけておく。入力効率として、例えば、単位時間あたりの入力可能文字数を表す入力速度を想定する。次に、得られた対応関係を参照して、文書の文字認識結果に基づいて推定された認識精度から、入力速度が最大となる確認修正手段を自動選択する。入力作業は、提示された確認修正手段に従って行う。

10

【0012】

他の背景技術として、宮原末治ほか（非特許文献 2）は、日本語 OCR のための効率的な確認修正法として、パターン間照合を利用した多段型の一括修正法を開示する。本方法は、初段のリジェクト処理でリジェクト文字の同一カテゴリを集めて一括修正し、次段の誤り修正処理で初段の処理結果を利用して誤り文字を検出し、再度一括修正を行う。

20

【先行技術文献】

【特許文献】

【0013】

【特許文献 1】特開 2 0 0 3 - 0 9 9 7 0 9 号公報

【特許文献 2】特開平 0 7 - 0 5 7 0 4 2 号公報

【特許文献 3】特開平 0 9 - 1 7 9 9 3 4 号公報

【特許文献 4】特開平 0 6 - 2 9 0 2 9 7 号公報

【特許文献 5】特開平 0 5 - 3 1 4 3 0 3 号公報

【特許文献 6】特開平 0 3 - 2 4 0 1 8 3 号公報

30

【特許文献 7】特開 2 0 0 3 - 2 2 3 6 0 8 号公報

【特許文献 8】特開 2 0 0 5 - 3 0 9 6 0 8 号公報

【非特許文献】

【0014】

【非特許文献 1】松村季樹ほか、「文字認識精度推定による確認修正手段の選択とその結果」、情報処理学会論文誌、V o l . 4 0 , N o . 4 , p p . 1 8 4 0 - 1 8 5 0 、 1 9 9 9 年 4 月

【非特許文献 2】宮原末治ほか、「パターン間照合を利用した文字認識結果の確認修正法」、電子情報通信学会論文誌、D - I I , V o l . J 7 7 - D - I I N o . 9 , p p . 1 7 3 5 - 1 7 4 3 、 1 9 9 4 年 9 月

40

【非特許文献 3】P.Auer, N.Cesa-Bianchi and P. Fischer, “Finite-time analysis of the multiarmed bandit problem”, Machine Learning, 47(2/3):p.235-256, 2002

【非特許文献 4】National Center for Biomedical Communications National Library of Medicine, “A report to the Board of Scientific Counselors Communications Engineering Branch Lister Hill”, October 2001, 2 0 1 1 年 9 月 2 9 日 検 索、URL: <http://archive.nlm.nih.gov/pubs/reports/mars2001-DIAU/mars2001-DIAU.php>

【発明の概要】

【発明が解決しようとする課題】

【0015】

OCR により電子化されたテキストデータの校正を行う場合、ある共通の属性（例えば

50

、文字)を有する部分として認識された紙媒体のイメージを切り出して、OCRの正誤をまとめて確認することで効率的に校正作業を行う手法が存在する(以下、この手法を「カーペット校正」といい、カーペット校正を実現するためのソフトウェア・ツールを「カーペット校正ツール」ということがある)。この手法は、現在の検証対象の共通属性とイメージが一致しているかどうかだけを判断すればよく、また、誤認識を発見した場合に一度に共通属性を有する複数の部分を修正することができる等のメリットを得られる。

【0016】

カーペット校正は、例えば、共通の属性が「文字」である場合に対象の文字がどの文字として認識されるかを判断するために前後の文脈の情報が必要である場合など、OCRの正誤の判断が困難である場合がある。そのようなOCRの正誤を判断するために、カーペット校正を実施した後で、共通の属性での切り出しを行わず元のイメージとテキストを対比して校正を行う手法(以下、この手法を「サイド・バイ・サイド校正」といい、サイド・バイ・サイド校正を実現するためのソフトウェア・ツールを「サイド・バイ・サイド校正ツール」ということがある))により電子化されたテキストの校正を行うことが行う必要であることがある。

10

【0017】

しかし、例えば、ほとんど間違いがない文字に対してカーペット校正を行うと、サイド・バイ・サイド校正のみを行うよりも時間がかかる場合がある。すなわち、カーペット校正を行うことなく、サイド・バイ・サイド校正を行ったほうが効率的な場合があり得る。しかし、従来はかかる問題を解決する手法は存在しなかった。

20

【0018】

したがって、本発明の目的の1つは、光学式文字認識で生成されたテキストデータの校正作業をより効率的に支援するための装置、プログラム、および、方法を提供することである。

【課題を解決するための手段】

【0019】

上記の目的を達成するために、光学式文字認識で生成されたテキストデータの校正を支援するためのシステムが提供される。システムは、テキストデータのカーペット校正を行うための第1の校正ツールと、テキストデータのサイド・バイ・サイド校正を行うための第2の校正ツールと、第1および第2の校正ツールで行った校正オペレーションのログを記録するための記憶部と、第1の校正ツールのカーペット校正を行う単位である属性ごとに、ログに基づいて算出された第1の校正ツールを使用して校正を行った場合に要する時間の第1の推測値と、第1の校正ツールを使用することなく第2の校正ツールを使用して校正を行った場合に要する時間の第2の推測値を比較して当該属性の校正に第1の校正ツールを使用するかどうかを決定する解析部を備える。

30

【0020】

カーペット校正を行う単位である属性が、文字であることが好ましい。

【0021】

第1の校正ツールによる校正を行った後で、第2の校正ツールによる校正が行われ、第1の校正ツールによる校正において、文字認識の正誤が不明である場合に、不明であることを示すマークが付されることが好ましい。さらに、不明であることを示すマークが付された文字を、カーペット校正において正誤不明であることを示すようにオペレータに提示することが好ましい。

40

【0022】

ログに基づいて、ログ統計情報を算出する算出部を備えることが好ましい。さらに、算出部が、各文字について、

【表 1】

a :	これまでに処理した文字イメージのうち、OCRが誤認識され、カーペット校正で誤認識を修正できた文字イメージの割合の期待値の推定値
b :	これまでに処理した文字イメージのうち、OCRが誤認識され、カーペット校正で誤認識を修正しなかった文字イメージの割合の期待値の推定値
u :	文字イメージ 1 つにつき OCR の正誤が不明であることを示すマークを付するために要する時間の期待値の推定値
w :	文字イメージ 1 つにつき サイド・バイ・サイド 校正で修正に要する時間の期待値の推定値
w' :	文字イメージ 1 つにつき カーペット 校正で修正に要する時間の期待値の推定値

10

のうちの少なくとも 1 つを統計情報として算出することができることが好ましい。

【0 0 2 3】

解析部が、統計情報に基づいて、

20

【表 2】

$\tilde{a}$:	これまでに処理した文字イメージのうち、OCRが誤認識され、カーペット校正で誤認識を修正できた文字イメージの割合の期待値の下限値の推定値
$\tilde{b}$:	これまでに処理した文字イメージのうち、OCRが誤認識され、カーペット校正で誤認識を修正しなかった文字イメージの割合の期待値の下限値の推定値
$\tilde{u}$:	文字イメージ 1 つにつき OCR の正誤が不明であることを示すマークを付するために要する時間の期待値の下限値の推定値
$\tilde{w}$:	文字イメージ 1 つにつき サイド・バイ・サイド 校正で修正に要する時間の期待値の下限値の推定値
$\tilde{w}'$:	文字イメージ 1 つにつき カーペット 校正で修正に要する時間の期待値の下限値の推定値

30

のうちの少なくとも 1 つを算出し、算出結果に基づいて当該文字の校正に第 1 の校正ツールを使用するかどうかを決定することが好ましい。

40

【0 0 2 4】

さらに、解析部が、該文字の校正に第 1 の校正ツールを使用するかどうかを、下記手順に従って決定することが好ましい。

【数 1】

 $L := \text{empty set, for each } x \text{ in } X,$

$$\tilde{a} \equiv \max \left[a - \sqrt{\frac{2 \log M}{M}}, 0 \right]$$

$$\tilde{b} \equiv \max \left[b - \sqrt{\frac{2 \log M}{M}}, 0 \right]$$

$$\tilde{u} \equiv \max \left[u - (u_U - u_L) \sqrt{\frac{2 \log N_U}{N_U}}, u_L \right]$$

$$\tilde{w} \equiv \max \left[w - (w_U - w_L) \sqrt{\frac{2 \log N_w}{N_w}}, w_L \right]$$

$$\tilde{w}' \equiv \max \left[w' - (w'_U - w'_L) \sqrt{\frac{2 \log N_{w'}}{N_{w'}}}, w'_L \right]$$

if $\tilde{b}\tilde{u} < \tilde{a}(\tilde{w} - \tilde{w}')$, add x to L

10

20

 L : 対象リストに含まれる文字の集合 X : 文書に出現する可能性のある文字の集合 x : 文書に出現する文字の集合 X の要素 u_U : 文字イメージ1つにつきOCRの正誤が不明であることを示すマークを付するために要する時間の上限値(規定値) u_L : 文字イメージ1つにつきOCRの正誤が不明であることを示すマークを付するために要する時間の下限値(規定値) w_U : 文字イメージ1つにつきサイド・バイ・サイド校正に要する時間の上限値(規定値) w_L : 文字イメージ1つにつきサイド・バイ・サイド校正に要する時間の下限値(規定値) w'_U : 文字イメージ1つにつきカーベット校正で修正に要する時間の上限値(規定値) w'_L : 文字イメージ1つにつきカーベット校正で修正に要する時間の下限値(規定値)

30

【0025】

算出部が、複数の文字をクラスタリングし、クラスタごとにログ情報から統計情報を算出することが好ましい。さらに、算出部が、(1)すべての文字に対する1つのクラスタの形成、(2)統計情報の各パラメータの空間によるクラスタの形成、または、(3)文字種ごとのクラスタの形成、のいずれかを実行することができることが好ましい。

40

【0026】

解析部の決定に基づいて属性を第1の校正ツールを使用する対象であることを記述するリストを備えることが好ましい。さらに、リストに基づいて、第1および第2の校正ツールの制御を行う制御部を備えることが好ましい。

【0027】

第1の校正ツールを、コンテキスト付きカーベット校正ツールとすることができる。また、第2の校正ツールが、ページ、文章、行のうちの少なくとも1つの単位でイメージデータとテキストデータを並べて提示することが好ましい。

50

【0028】

以上、テキストデータの校正を支援するためのシステムとして本発明の概要を説明したが、本発明は、方法、プログラム、プログラム製品、ソフトウェア、ソフトウェア製品、装置などとして把握することもできる。プログラム製品ないしソフトウェア製品は、例えば、前述のプログラム、ソフトウェアを格納した記憶媒体を含め、あるいはプログラム、ソフトウェアを伝送する媒体を含めることができる。

【0029】

上記の発明の概要は、本発明の必要な特徴の全てを列挙したものではなく、これらの構成要素のコンビネーションまたはサブコンビネーションもまた、発明となり得ることに留意すべきである。

【図面の簡単な説明】

【0030】

【図1】本発明の実施形態におけるデータ処理システムの外観イメージの一例を示した図である。

【図2】本発明の実施形態におけるデータ処理システムの機能ブロック図である。

【図3】本発明の実施形態におけるデータ処理システムの全体動作を表現するフローチャートである。

【図4】本発明の実施形態におけるデータ処理システムのカーペット校正処理の動作を表現するフローチャートである。

【図5】本発明の実施形態におけるデータ処理システムのサイド・バイ・サイド校正処理の動作を表現するフローチャートである。

【図6】本発明の実施形態におけるデータ処理システムのカーペット校正処理のインターフェースのスナップショットの一例である。

【図7】本発明の実施形態におけるデータ処理システムのサイド・バイ・サイド校正処理のインターフェースのスナップショットの一例である。

【図8】コンテキスト付きカーペット校正処理のインターフェースのスナップショットの一例である。

【図9】本発明の実施形態におけるデータ処理システムを実現するのに好適な情報処理装置のハードウェア構成の一例を示した図である。

【発明を実施するための形態】

【0031】

以下、本発明を実施するための形態を図面に基づいて詳細に説明するが、以下の実施形態は特許請求の範囲に記載された発明を限定するものではなく、また実施形態の中で説明されている特徴の組み合わせの全てが発明の解決手段に必須であるとは限らない。

【0032】

また、本発明は多くの異なる態様で実施することが可能であり、実施の形態の記載内容に限定して解釈されるべきものではない。実施の形態の説明の全体を通じて同じ要素には同じ番号を付している。

【0033】

図1は、本発明の実施形態におけるデータ処理システムの全体像の一例を示した図である。データ処理システム100は紙媒体文書を電子化して得られたテキストデータの校正を実施するオペレータが操作する端末110および紙媒体文書をスキャンし、電子イメージデータとすることができるスキャナ120を含む。

【0034】

本発明の実施の形態では、大量の紙媒体文書がスキャナ120によって読み込まれ、生成されたイメージデータが端末110に送信される。本発明の実施形態の端末110には光学式文字認識（OCR）ソフトウェアがインストールされており、受け取った文書のイメージデータから文書中に存在する文字を認識し、テキストデータに変換する。

【0035】

本発明の実施形態では、端末110にインストールされたOCRソフトウェアの性質上

10

20

30

40

50

、変換されたテキストデータが完全に正確であるとは限らないものとする。そして、オペレータは、端末110にインストールされた校正ソフトウェアを操作して文字認識の正誤の確認および誤って認識された文字の修正を行う。テキストデータの校正のために、ユーザは、端末110の表示装置に作業状況の情報などを視覚的に提示するウィンドウ・ベースのグラフィカル・ユーザ・インターフェースを通じて、キーボード、マウスなどの入力デバイスを用いてアプリケーション・プログラムを操作する。

【0036】

本発明の実施形態の校正ソフトウェアは、カーペット校正機能およびサイド・バイ・サイド校正機能を有するものとする。本発明の実施形態のカーペット校正機能は、共通の文字として認識されたイメージデータの部分を切り出して、OCRの正誤をまとめて確認することで効率的に校正作業を行う環境を提供するものである。

10

【0037】

本発明の実施形態のサイド・バイ・サイド校正機能は、カーペット校正機能のように共通の文字を有するイメージの切り出しを行うのではなく、元のイメージとテキストデータを所定の単位で対比して校正を行う。その一例として、本発明の実施形態では、サイド・バイ・サイド校正機能が文書内のページのイメージと、それに対応する認識されたテキストをオペレータに同一画面に提示して校正をさせる機能を含むシナリオを開示するが、サイド・バイ・サイド校正機能によって表示される単位はページに限られず、文章単位、行単位などでイメージとテキストデータの対比をするようにしてもよい。

20

【0038】

本発明の実施形態のカーペット校正機能を用いた作業では、オペレータは、一覧表示されているイメージのそれぞれが、現在検証対象の共通の文字であるかどうかだけを判断すればよいので、オペレータの負担が軽減され、作業効率が向上する場合がある。また、共通の文字を有するイメージは同じようにOCRで誤認識をされることが多いので、カーペット校正機能を用いた校正においてそのような同様の誤認識を複数発見した場合に一度に複数箇所を修正することができる等のメリットを得られる。一方、本発明の実施形態のカーペット校正機能については、例えば、対象のイメージがどの文字として認識されるかを判断するために前後のコンテキストの情報が必要である場合など、OCRの正誤が判断できない場合などがあり得る。

30

【0039】

本発明の実施形態では、オペレータがカーペット校正でOCRの正誤が判断できない場合、対応するイメージに正誤不明であることを示すマークを付しておき、その後、最終的に正誤を判断し必要に応じて修正を行うためにサイド・バイ・サイド校正を行うものとする。しかし、本発明の実施形態では、一部のイメージについては、カーペット校正を行わず、最初からサイド・バイ・サイド校正を行うようにされている。すなわち、本発明の実施形態では、システムは、カーペット校正およびサイド・バイ・サイド校正の両方を行うか、カーペット校正を実施せずにサイド・バイ・サイド校正を行うかを制御する機構を備える。

【0040】

本発明の実施形態では、事前指定のリスト、文字認識出力の解析、テキスト構造からの推論、および、過去の校正のログから得られた統計情報に基づいて、かかる制御が自動的に行われる。かかる制御を行う結果、効率的な校正作業が実現される。

40

【0041】

図2は、本発明の実施形態のデータ処理システム100の機能ブロック図である。なお、図2の機能ブロック図に示す各要素は、図9に例示したハードウェア構成を有する情報処理装置およびスキャナ装置を組み合わせ、ハードディスク装置13などに格納されたオペレーティング・システム、OCRソフトウェア、校正ソフトウェアなどのコンピュータ・プログラムをメインメモリ4にロードした上でCPU1に読み込ませ、ハードウェア資源とソフトウェアを協働させることによって実現することができる。具体的には、イメージ読取部205はスキャナ120において実現され、その他の要素は端末110において

50

実現される。

【0042】

本発明の実施形態のデータ処理システム100は、イメージ読取部205、OCR部210、テキストデータ保存部215、カーペット校正部220、サイド・バイ・サイド校正部225、校正ツール制御部230、入力部235、出力部240を備える。

【0043】

本発明の実施形態において、イメージ読取部205は紙媒体の文書を光学的に読み取ってイメージデータを生成し、OCR部210に引き渡す。文書のイメージデータを受けたOCR部210は、イメージデータに存在する文字を認識し、テキストデータに変換する。また、イメージデータからその他の属性を抽出するようにしてもよい。生成されたテキストデータに含まれる文字が、対応するイメージ片、イメージデータにおける位置の情報等と関連付けられ、テキストデータ保存部215に保存される。その他の属性が抽出されている場合は、あわせて対応する文字と関連付けて保存するようにしてもよい。イメージ読取部205、OCR部210およびテキストデータ保存部215は、既存のスキャニング技術、OCR技術、ストレージ技術によって当業者は適宜実施をすることができるので、ここではこれ以上詳細には説明されない。

10

【0044】

本発明の実施形態のカーペット校正部220は、ある共通の文字として認識された紙媒体のイメージデータの部分を切り出して一覧表示し、オペレータがOCRの正誤をまとめて確認することで効率的な校正作業を行う環境を提供する機能を有する。

20

【0045】

図6は、本発明の実施形態におけるカーペット校正部220によって提供される環境においてカーペット校正を行うためのインターフェース600のスナップショットの一例である。カーペット校正のインターフェース600の上部605に文書を特定する識別子が表示されている。オペレータは、共通の属性610（ここでは文字r）を有する部分として認識された文書のイメージデータの一部が切り出され、フィールド615にカーペット状に並べられている。このインターフェースを通じて、並べられたイメージ片が文字rとしてOCRされているかどうかをオペレータが判断する。例えば、OCRが誤っていた場合（例えば、文字γ（ガンマ）のイメージが、誤って文字rとして認識されている場合など）は、誤っているイメージを1回クリックして選択し、入力フィールド620に正しい文字（ここではγ）を入力してボタン625を押すことによって修正を行う。複数のイメージについてOCRが誤っている場合は、当該複数のイメージを一括選択して同様の操作を行うことで、一度に複数のOCRの認識誤りを修正することができる。

30

【0046】

本発明の実施形態のサイド・バイ・サイド構成部225は、カーペット校正部220のように共通の文字での切り出しを行うのではなく、元のイメージとテキストデータを所定の単位で対比して校正を行う環境をオペレータに提供する機能を備える。その一例として、サイド・バイ・サイド校正部225は、文書内のページのイメージと、それに対応する認識されたテキストをオペレータに同一画面に提示して校正をさせる機能を、図7を用いて示すが、本発明の実施形態では、サイド・バイ・サイド校正機能によって表示される単位はページに限られず、文章単位、行単位などでイメージとテキストデータの対比をすることができるものとする。

40

【0047】

図7は、本発明の実施形態におけるサイド・バイ・サイド校正部225によって提供される環境においてサイド・バイ・サイド校正を行うためのインターフェース700のスナップショットの一例である。カーペット校正のインターフェース700の左側に紙文書のページ単位のイメージ705が表示され、右側にOCRの結果得られたテキストデータのうち、イメージ705に対応する部分がイメージ705と類似するレイアウト710にしたがって表示されている。オペレータは、左右を見比べながら一致していない部分が見つかった場合は、レイアウト710に表示されたテキストを修正する。本発明の実施形態に

50

おけるサイド・バイ・サイド校正では、左側のイメージと右側テキストを見比べなければならないこと、また一括修正ができないことなどから時間がカーペット校正よりも時間がかかる場合がある。他方、サイド・バイ・サイド校正では、前後のコンテキストを参照して正誤が判断できるのでカーペット校正で発見されないOCRの認識誤りが発見できる可能性がある。

【0048】

図2に戻り、本発明の実施形態の校正ツール制御部230は、各文字について、事前指定、OCR出力の解析、テキスト構造からの推論、および、過去の校正ログから抽出された統計情報に基づいて形成されたカーペット校正対象リスト245を参照して、カーペット校正部220によるカーペット校正を行うかどうかを判断する。すなわち、校正ツール制御部230は、カーペット校正対象リスト245を参照して、各文字について、カーペット校正部220およびサイド・バイ・サイド校正機能225の両方を用いて校正をするか、サイド・バイ・サイド校正機能225のみを用いて校正をするか、または、カーペット校正およびサイド・バイ・サイド校正のいずれも行わないか、の校正モードを制御する機能を含む。

10

【0049】

また、校正ツール制御部230は、ユーザが入力部235を通じて入力した校正オペレーションをカーペット校正部220およびサイド・バイ・サイド校正機能225に引渡す。また、校正ツール制御部230は、カーペット校正部220およびサイド・バイ・サイド校正機能225から受けたユーザに提示する情報を出力部240に引き渡す。さらに、校正ツール制御部230は、カーペット校正部220およびサイド・バイ・サイド校正機能225を監視し、それぞれにおいて修正に要した時間および文字認識の正誤を含むログ情報を生成し、校正ログ保存部245に引き渡す。入力部235はユーザからの入力を受け付けて校正ツール制御部230に引渡し、出力部240は校正ツール制御部230からユーザに提示すべき情報を受けて、ユーザに視覚的に提示する。

20

【0050】

さらに本発明の実施形態のデータ処理システム100は、校正ログ保存部245、処理時間推定部250、ログ統計情報保存部255、カーペット校正対象リスト260、OCR出力解析部265、および、構造推論エンジン270を含む。これらの要素は、校正ツール制御部230に制御のための情報を提示するために設けられている。

30

【0051】

本発明の実施形態のカーペット校正対象リスト260は、検証対象の文書に含まれる文字とその文字に適用される校正モードを表すタグのペアを含む。本発明の実施形態では、タグは、<accept>（その文字については、カーペット校正部220およびサイド・バイ・サイド校正機能225の両方を用いて校正をすることを意味する）、<reject>（その文字については、サイド・バイ・サイド校正機能225のみを用いて校正をすることを意味する）、<no-correction>（その文字については、カーペット校正およびサイド・バイ・サイド校正のいずれも行わないことを意味する）の三種類が存在するものとする。

【0052】

オペレータは、入力部235を通じて、カーペット校正対象リスト245を編集することができる。例えば、ユーザは文字だけ見ても判別が困難なアルファベットの“1”（エル）と、数字の“1”（いち）をカーペット校正の対象外とし、<reject>タグを付することができる。

40

【0053】

本発明の実施形態のOCR出力解析部265は、OCRの結果を解析し、所定のルールに従って、各文書においてカーペット校正を実施することは効率的であると判断される文字をカーペット校正の対象外として特定し、カーペット校正対象リストを更新する。具体的には、本発明の実施形態では、OCR出力解析部265は、ある文書において所定の閾値（例えば、1回）以下しか出現しない文字をカーペット校正の対象外として特定する（すなわち、<no-correction>タグが付される）。

50

【0054】

本発明の実施形態の構造推論エンジン270は、ある文書に表れる文字がその文書のどこに現れるかを構造的に解析し、校正してもほとんどあるいは全く意味のない文字を校正の対象外として特定し、カーペット校正対象リストを更新する。例えば、本文ではないページ番号など後で使用されない部分にある文字を対象外に指定する（すなわち、<no-correction>タグが付される）。

【0055】

本発明の実施形態の校正ログ保存部245は、カーペット校正部220およびサイド・バイ・サイド校正部225を監視し、それぞれにおいて修正に要した時間およびOCRの正誤を含むログ情報を校正ツール制御部230から受けて記憶する。

10

【0056】

本発明の実施形態のログ統計情報算出部250は、校正ログ保存部260に保存されたログ情報に基づいて、ログ統計情報を算出する。具体的には、本発明の実施形態では、以下の統計情報が文字ごとに算出されるが、これら以外の統計情報を算出するようにしてもよいことは勿論である。

【0057】

【表3】

a : これまでに処理した文字イメージのうち、OCRが誤認識され、
カーペット校正で誤認識を修正できた文字イメージの割合の
期待値の推定値

20

b : これまでに処理した文字イメージのうち、OCRが誤認識され、
カーペット校正で誤認識を修正しなかった文字イメージの
割合の期待値の推定値

u : 文字イメージ1つにつきOCRの正誤が不明であることを示す
マークを付するために要する時間の期待値の推定値

w : 文字イメージ1つにつきサイド・バイ・サイド校正で修正に
要する時間の期待値の推定値

w' : 文字イメージ1つにつきカーペット校正で修正に要する
時間の期待値の推定値

30

N : これまでの処理において誤認識を修正した回数

M : これまでの処理においてカーペット校正の対象となった回数

N_w : これまでの処理においてカーペット校正で誤認識を修正した回数

N_U : これまでの処理においてOCRの正誤不明マークが付された回数

$N_{w'}$: これまでの処理においてサイド・バイ・サイド校正で誤認識を修正した回数

40

【0058】

本発明の実施形態では、ログ情報におけるそれぞれのイベントの回数をカウントすることによって、 N 、 M 、 N_w 、 N_U 、 $N_{w'}$ をそれぞれ算出することができる。

【0059】

ログ統計情報解析部255は、所定のイベントの生起（例えば、本発明の実施形態では、1つの文書の校正作業が終わるなど）を受けて、ログ統計情報算出部265によって算出された統計情報に基づいて、カーペット校正部220およびサイド・バイ・サイド校正機能225の両方で校正を行った場合に要する時間の推測値と、カーペット校正部220を使用せずにサイド・バイ・サイド校正機能225で校正を行った場合に要する時間の推

50

測値を対比し、いずれのモードで校正を行うのが最も効率がよいかどうかを判断する。本発明の実施形態では、ログ統計情報解析部 270 の判断結果に基づいて、カーペット校正対象リスト 260 が更新されることとなる。

【0060】

具体的には、本発明の実施形態では、以下の処理が行われることによって、ログ統計情報解析部 255 は、ある文字 x をカーペット校正部 220 およびサイド・バイ・サイド校正機能 225 の両方を用いて校正をするかどうかを判定し、両方を用いて校正をすると判定された場合、<accept>のタグを付してカーペット校正対象リスト 260 に追加されることとなる。

【0061】

【数 2】

$L := \text{empty set, for each } x \text{ in } X,$

$$\tilde{a} \equiv \max \left[a - \sqrt{\frac{2 \log M}{M}}, 0 \right]$$

$$\tilde{b} \equiv \max \left[b - \sqrt{\frac{2 \log M}{M}}, 0 \right]$$

$$\tilde{u} \equiv \max \left[u - (u_U - u_L) \sqrt{\frac{2 \log N_U}{N_U}}, u_L \right]$$

10

$$\tilde{w} \equiv \max \left[w - (w_U - w_L) \sqrt{\frac{2 \log N_w}{N_w}}, w_L \right]$$

$$\tilde{w}' \equiv \max \left[w' - (w'_U - w'_L) \sqrt{\frac{2 \log N_{w'}}{N_{w'}}}, w'_L \right]$$

if $\tilde{b}\tilde{u} < \tilde{a}(\tilde{w} - \tilde{w}')$, add x to L

20

L : 対象リストに含まれる文字の集合

X : 文書に出現する可能性のある文字の集合

x : 文書に出現する文字の集合 X の要素

$\tilde{a}$: これまでに処理した文字イメージのうち、OCRが誤認識され、カーペット校正で誤認識を修正できた文字イメージの割合の期待値の下限値の推定値

$\tilde{b}$: これまでに処理した文字イメージのうち、OCRが誤認識され、カーペット校正で誤認識を修正しなかった文字イメージの割合の期待値の下限値の推定値

$\tilde{u}$: 文字イメージ 1 つにつき OCR の正誤が不明であることを示すマークを付するために要する時間の期待値の下限値の推定値

30

$\tilde{w}$: 文字イメージ 1 つにつき サイド・バイ・サイド 校正で修正に要する時間の期待値の下限値の推定値

$\tilde{w}'$: 文字イメージ 1 つにつき カーペット 校正で修正に要する時間の期待値の下限値の推定値

u_U : 文字イメージ 1 つにつき OCR の正誤が不明であることを示すマークを付するために要する時間の上限値(規定値)

u_L : 文字イメージ 1 つにつき OCR の正誤が不明であることを示すマークを付するために要する時間の下限値(規定値)

40

w_U : 文字イメージ 1 つにつき サイド・バイ・サイド 校正に要する時間の上限値(規定値)

w_L : 文字イメージ 1 つにつき サイド・バイ・サイド 校正に要する時間の下限値(規定値)

w'_U : 文字イメージ 1 つにつき カーペット 校正で修正に要する時間の上限値(規定値)

w'_L : 文字イメージ 1 つにつき カーペット 校正で修正に要する時間の下限値(規定値)

【 0 0 6 2 】

50

なお、本発明の実施形態の上記手順においては、サンプル数が少ないことに起因した統計上の偏りの影響を少なくするために、それぞれのパラメータについて、「パラメータの統計情報」から「パラメータの取り得る幅」に「単調減少関数（ルート部分）」を乗じたもの（「補正統計情報」という）を算出し、それに基づいて「カーペット校正部２２０およびサイド・バイ・サイド校正機能２２５の両方で校正を行った場合に要する時間の推測値」と「カーペット校正部２２０を使用せずにサイド・バイ・サイド校正機能２２５で校正を行った場合に要する時間の推測値」を算出して比較していることに留意されたい。ただし、補正統計情報が当該パラメータの下限值よりも小さい場合は下限値が使用される。

【００６３】

図３は、本発明の実施形態におけるデータ処理システム１００の動作を表現するフローチャート３００である。処理はステップ３０５でスタートし、ステップ３１０でユーザの入力部２３５を介した入力を処理することで初期設定が実施される。かかる初期設定は、カーペット校正対象リスト２６０への手動の編集（特定の文字を<accept>、<reject>、<no-correction>タグを付して登録すること）を含む。

【００６４】

処理はステップ３１５に進み、イメージ読取部２０５によって電子化される紙媒体の文書のスキャンが実施される。本発明の実施形態では、スキャンは一度に一文書ずつ行われるようにしてもよいし、一度に多量の文書を対象として実施されるようにしてもよい。スキャンされた文書は文書を特定する情報およびページ番号の情報と関連付けられて、イメージデータが形成され、ＯＣＲ部２１０に引き渡される。

【００６５】

次いで、ステップ３２０において、ＯＣＲ部２１０によって、ＯＣＲ処理が実施される。ＯＣＲ処理の結果、認識された文字を含むテキストデータが形成され、テキストデータ保存部２１５に記憶される。そして、ステップ３２５において、ＯＣＲ出力解析部２６５および構造推論エンジン２７０によって、ＯＣＲ出力解析および構造推論が実施され、カーペット校正対象リスト２６０が更新される。

【００６６】

処理はステップ３３０に進み、ステップ３１５でスキャンした文書で未校正のものうち、いずれを校正するかを選択する。この選択は、スキャンをした順番で、順次文書を選択していてもよいし、他のルールに従ってシステムが選択するようにしてもよい。また、オペレータにユーザ・インターフェースを通じて選択させてもよい。

【００６７】

次いで、ステップ３３５で過去の校正オペレーションのログからログ統計情報算出部２６５が算出した統計情報に基づいて、ログ統計情報解析部２７０がカーペット校正対象リストの更新を行う。かかる更新の詳細については、ログ統計情報解析部２７０の説明において既に述べたので、ここでは詳述されない。

【００６８】

次いで、ステップ３３５において、ステップ３３０で選択された文書内のテキストについて、カーペット校正処理が実行される。カーペット校正処理の詳細は、図４を用いて後で詳細に説明される。次いで、ステップ３４０において、サイド・バイ・サイド校正処理が実行される。サイド・バイ・サイド校正処理の詳細は、図５を用いて後で詳細に説明される。なお、ステップ３４０および３４５では、校正オペレーションのログがシステム上で記録されているものとする。かかるログ採取の手法は、当業者は適宜実施することができるので、ここではこれ以上詳述されない。

【００６９】

ステップ３５０において、校正対象となる文書のイメージデータがまだ存在するかどうか判断される。ステップ３５０で存在すると判断された場合、処理はＹＥＳの矢印を介してステップ３３０に戻り、校正対象文書の選択（ステップ３３５）および校正処理（ステップ３４０、３４５）が繰り返される。ステップ３５０で存在しないと判断された場合、処理はＮＯの矢印を介してステップ３５５に進み、スキャン対象となる紙媒体の文書が

さらに存在するかどうか判断される。ステップ350で存在すると判断された場合、処理はYESの矢印を介してステップ315に進み、以降の処理が繰り返される。ステップ355で存在しないと判断された場合、処理はNOの矢印を介してステップ360へ進み終了する。

【0070】

次に、図4のフローチャート335を使用して、図3のステップ335におけるカーペット校正処理の一例をより詳細に説明する。処理は、ステップ405でスタートし、ステップ410でカーペット校正対象リストを参照して、校正対象文書に含まれる文字からカーペット校正の対象となる文字（すなわち、<accept>タグが付された文字）を1つ選択する。この選択は、オペレータが表示されたリストの中から手動で選択することで行ってもよいし、コンピュータが所定のルールに従って自動で選択するようにしてもよい。

10

【0071】

処理はステップ415に進み、ステップ410において選択された文字について、校正対象の文書に含まれる対応する文字イメージの部分の切り出し、図6を用いて既に説明されたインターフェースを通じて一覧表示する。次にステップ420で、オペレータが一覧表示された文字イメージのうちの1以上について、当該文字として正しく認識されているかどうかを判断する対象として選択し、システムが当該選択を受け付ける。

【0072】

処理はステップ425に進み、ステップ420において選択された文字イメージについて文字認識が正しいかどうか判断される。ステップ425で正しいと判断された場合は、YESの矢印を通じてステップ430へ進み、選択イメージに文字認識が正しく行われたことを確認したことを示す「確認済み」マークが付され、次いで、ステップ445へ進む。

20

【0073】

ステップ425で正誤不明と判断された場合、「正誤不明」の矢印からステップ435へ進み、カーペット校正で文字認識の正誤が不明であったことを示す「正誤不明」マークが付され、次いで、ステップ445へ進む。

【0074】

ステップ425で文字認識が誤っていると判断された場合は、NOの矢印を通じてステップ435へ進み、当該イメージに対応する文字認識の修正を行い、誤っていた文字認識を修正したことを示す「修正済」マークが付され、次いで、ステップ445へ進む。これらの一連のカーペット校正処理については、すべてのオペレーションについてログが保存される。

30

【0075】

ステップ445では、一覧表示されているイメージに未確認イメージが存在するかどうか判断される。ステップ445で存在すると判断された場合、処理はYESの矢印からステップ420に戻り、以降のステップが繰り返される。ステップ445で存在しないと判断された場合は、NOの矢印を介してステップ450に進む。ステップ450では、校正対象の文書中に存在する文字に、カーペット校正の対象文字がまだ存在するかどうか判断される。ステップ450において存在すると判断された場合は、YESの矢印を介してステップ410に戻り、以降のステップを繰り返す。ステップ450において存在しないと判断された場合は、処理はNOの矢印を通じて、ステップ455に進む。

40

【0076】

次に、ステップ455において、ログ統計情報算出部250が、カーペット校正のログ情報に基づいて、校正オペレーションに関する統計情報を更新する。具体的には、本発明の実施形態では、以下の擬似コードによって表現される手順を実行するプログラムによって、ステップ455における統計情報の更新が行われる。

【0077】

【数 3】

for each x in X # 文字集合 X の要素 x すべてに対してそれぞれ以下の処理を行う
 $M(x) \mathrel{+}= \text{count}(x)$ # 今回の校正の対象となった回数を加算する
 $N_W(x) \mathrel{+}= \text{count_corrected}(x)$ # 今回の校正で誤認識を修正した回数を加算する
 $N_U(x) \mathrel{+}= \text{count_marked}(x)$ # 今回の校正で正誤不明マークが付された回数を加算する
 $T_W(x) \mathrel{+}= \text{total_time_for_correction}(x)$ # 今回の校正で修正に要した時間を加算する
 $T_U(x) \mathrel{+}= \text{total_time_for_mark}(x)$ # 今回の校正で正誤不明マークを付けるために要した時間を加算する
 $a(x) = N_W(x) / M(x)$
 $b(x) = N_U(x) / M(x)$
 $w'(x) = T_W(x) / M(x)$
 $u(x) = T_U(x) / N_U(x)$

10

【0078】

ステップ455において統計情報の更新が完了したことに応じて、ステップ460において処理は終了する。

【0079】

次に、次に、図5のフローチャート340を使用して、図3のステップ340におけるサイド・バイ・サイド校正処理の一例をより詳細に説明する。図5は、本発明の実施形態におけるデータ処理システム100のサイド・バイ・サイド処理を表現するフローチャート340である。処理はステップ505から開始し、次いで、ステップ510でサイド・バイ・サイド校正の対象ページを選択する。

20

【0080】

処理はステップ515に進み、ステップ510において選択されたページについて、当該ページの読み取られたイメージと、それに対応する文字認識されたテキストデータを、図7に例示したユーザ・インターフェースを通じてオペレータに同時に提示する。このとき、既にカーペット校正において文字認識が正しいことが確認され、また、カーペット校正で誤認識が修正されたものについてはサイド・バイ・サイド校正の対象でないということがオペレータわかるように表示がされる。また、カーペット校正において、「正誤不明」マークが付された文字はそのことがオペレータにわかるように表示される。かかる表示は様々な方法が考えられ、例えば、文字の色を変える等が考えられるがこれに限られない。なお、既に述べたようにサイド・バイ・サイド校正において表示される単位はページに限られず、文章単位、行単位などでイメージとテキストデータの対比をすることもできることに留意されたい。

30

【0081】

処理はステップ520に進み、ステップ510で選択され、ステップ515で表示されたページについて、サイド・バイ・サイド校正処理がなされる。これらの一連のサイド・バイ・サイド校正処理についても、すべての校正オペレーションについてログが記録される。

【0082】

次に、ステップ525において、まだサイド・バイ・サイド校正が完了していないページが校正対象の文書に存在するかどうか判断される。ステップ525において存在すると判断された場合、処理はYESの矢印を介してステップ510に戻り、以降のステップが繰り返される。ステップ525において存在しないと判断された場合、処理はNOの矢印からステップ530に進む。

40

【0083】

ステップ530では、ログ統計情報算出部250が、サイド・バイ・サイド校正のログ情報に基づいて校正オペレーションに関する統計情報を更新する。具体的には、本発明の実施形態では、以下の擬似コードによって表現される手順を実行するプログラムによって、ステップ530における統計情報の更新が行われる。

【0084】

50

【数 4】

for each x in X # 文字集合 X の要素 x すべてに対してそれぞれ以下の処理を行う
 $N_w(x) += \text{count_corrected}(x)$ # 今回の校正で誤認識を修正した回数を加算する
 $T_w(x) += \text{total_time_for_correction}(x)$ # 今回の校正で修正に要した時間を加算する
 $w(x) = T_w(x) / N(x)$

【0085】

ステップ530において統計情報の更新が完了したことに応じて、ステップ535において処理は終了する。

10

【0086】

図9は、本発明の実施の形態による端末110を実現するのに好適な情報処理装置のハードウェア構成の一例を示した図である。情報処理装置は、バス2に接続されたCPU（中央処理装置）1とメインメモリ4を含んでいる。ハードディスク装置13、30、およびCD-ROM装置26、29、フレキシブル・ディスク装置20、MO装置28、DVD装置31のようなリムーバブル・ストレージ（記録メディアを交換可能な外部記憶システム）がフロッピーディスクコントローラ19、IDEコントローラ25、SCSIコントローラ27などを経由してバス2へ接続されている。

【0087】

フレキシブル・ディスク、MO、CD-ROM、DVD-ROMのような記憶メディアが、リムーバブル・ストレージに挿入される。これらの記憶メディアやハードディスク装置13、30、ROM14には、オペレーティング・システムと協働してCPU等に命令を与え、本発明を実施するためのコンピュータ・プログラムのコードを記録することができる。メインメモリ4にロードされることによってコンピュータ・プログラムは実行される。コンピュータ・プログラムは圧縮し、また複数に分割して複数の媒体に記録することもできる。

20

【0088】

情報処理装置は、キーボード／マウス・コントローラ5を経由して、キーボード6やマウス7のような入力デバイスからの入力を受ける。情報処理装置は、視覚データをユーザに提示するための表示装置11にDAC/LCDC10を経由して接続される。

30

【0089】

情報処理装置は、ネットワーク・アダプタ18（イーサネット（R）・カードやトークンリング・カード）等を介してネットワークに接続し、他のコンピュータ等と通信を行うことが可能である。図示はされていないが、パラレルポートを介してプリンタと接続することや、シリアルポートを介してモデムを接続することも可能である。

【0090】

以上の説明により、本発明の実施の形態によるデータ処理システムを実現するのに好適な情報処理装置は、通常のパーソナルコンピュータ、ワークステーション、メインフレームなどの情報処理装置、または、これらの組み合わせによって実現されることが容易に理解されるであろう。ただし、これらの構成要素は例示であり、そのすべての構成要素が本発明の必須構成要素となるわけではない。

40

【0091】

本発明の実施の形態において使用される情報処理装置の各ハードウェア構成要素を、複数のマシンを組み合わせ、それらに機能を配分し実施する等の種々の変更は当業者によって容易に想定され得ることは勿論である。それらの変更は、当然に本発明の思想に包含される概念である。

【0092】

本発明の実施の形態のデータ処理システムは、マイクロソフト・コーポレーションが提供するWindows（R）オペレーティング・システム、アップル・コンピュータ・インコーポレイテッドが提供するMac OS（R）、X Window Systemを備

50

えるUNIX(R)系システム(例えば、インターナショナル・ビジネス・マシーンズ・コーポレーションが提供するAIX(R))のような、GUI(グラフィカル・ユーザー・インターフェース)マルチウインドウ環境をサポートするオペレーティング・システムを採用する。

【0093】

以上から、本発明の実施の形態において使用されるデータ処理システムは、特定のマルチウインドウ・オペレーティング・システム環境に限定されるものではないことを理解することができる。

【0094】

以上、本発明の実施形態によれば、カーペット校正ツールとサイド・バイ・サイド校正ツールを適切に制御することができる、効率的な校正システム、ソフトウェア、および方法が実現される。従って、校正システムを操作してOCRによる誤認識の修正を行うユーザの生産性を高めることができることが容易に理解できる。

【0095】

また、本発明は、ハードウェア、ソフトウェア、またはハードウェア及びソフトウェアの組み合わせとして実現可能である。ハードウェアとソフトウェアの組み合わせによる実行において、所定のプログラムを有するデータ処理システムにおける実行が典型的な例として挙げられる。かかる場合、該所定プログラムが該データ処理システムにロードされ実行されることにより、該プログラムは、データ処理システムを制御し、本発明にかかる処理を実行させる。このプログラムは、任意の言語・コード・表記によって表現可能な命令群から構成される。そのような命令群は、システムが特定の機能を直接、または1. 他の言語・コード・表記への変換、2. 他の媒体への複製、のいずれか一方もしくは双方が行われた後に、実行することを可能にするものである。

【0096】

もちろん、本発明は、そのようなプログラム自体のみならず、プログラムを記録した媒体もその範囲に含むものである。本発明の機能を実行するためのプログラムは、フレキシブル・ディスク、MO、CD-ROM、DVD、ハードディスク装置、ROM、MRAM、RAM等の任意のコンピュータ読み取り可能な記録媒体に格納することができる。かかるプログラムは、記録媒体への格納のために、通信回線で接続する他のデータ処理システムからダウンロードしたり、他の記録媒体から複製したりすることができる。また、かかるプログラムは、圧縮し、または複数に分割して、単一または複数の記録媒体に格納することもできる。また、様々な形態で、本発明を実施するプログラム製品を提供することも勿論可能であることにも留意されたい。

【0097】

上記の実施の形態に、種々の変更または改良を加えることが可能であることが当業者に明らかである。例えば、上述の本発明の実施形態では、図6に示すユーザ・インターフェースを備える特定のカーペット校正ツールを使用する態様を開示しているが、他のいかなるカーペット校正ツールを用いる変更を加えることができる。そのような他のカーペット校正ツールには、図8に示すユーザ・インターフェース800を備えるコンテキスト付きカーペット校正ツールが含まれる。

【0098】

具体的には、図8に示すインターフェース800は、現在、図6のインターフェースのコンポーネント605~625に対応する805~825のほか、選択中の文字イメージのコンテキスト(すなわち、文字イメージの周辺)をユーザに提示する小さなフィールド835およびコンテキスト表示機能をアクティベートするためのチェックボックス830がさらに含まれている。この場合、上述の実施形態と異なり、カーペット校正ツールを使用するだけで(すなわち、サイド・バイ・サイド校正ツールを使用することなく)高精度の校正ができる可能性がある。

【0099】

しかし、上述の実施形態におけるカーペット校正ツールと異なり、コンテキスト付きカ

10

20

30

40

50

カーペット校正ツールを使用する場合は一括修正ができず、また、コンテキストを確認する必要があるため、サイド・バイ・サイド校正ツールよりも常に速く校正ができるとは限らない。従って、この場合も、ログ情報に基づいて算出したカーペット校正ツールを使用して校正を行った場合に要する時間の第1の推測値と、カーペット校正ツールを使用することなくサイド・バイ・サイド校正ツールを使用して校正を行った場合に要する時間の第2の推測値を対比して、カーペット校正ツールを使用するかどうかを決定する意義がある。

【0100】

他の変更例として、本発明の実施形態のシステムの運用を開始する前に、ログ情報に相当する教師データが得られる場合に、その教師データを利用して運用開始から校正時間を最適化することができる。

10

【0101】

他の変更例として、ある文字について、「カーペット校正ツールを使用しない校正に要する時間」が「カーペット校正ツールを使用する校正に要する時間」よりも大きい確率が、ほぼ1に近い場合に、当該文字をカーペット校正ツールの対象としてリストに追加することができる。

【0102】

他の変更例として、文字ごとに十分なサンプル数が得られない可能性があることに対応するために、複数の文字をクラスタリングして、クラスタごとにログ情報から統計情報を算出することが考えられる。クラスタリングの例としては、(1)すべての文字に対して1つのクラスタを形成する、(2)「OCRが誤認識され、カーペット校正で誤認識を修正できた文字イメージの割合」、「OCRが誤認識され、カーペット校正で誤認識を修正しなかった文字イメージの割合」、「文字イメージ1つにつきOCRの正誤が不明であることを示すマークを付するために要する時間」、「文字イメージ1つにつきサイド・バイ・サイド校正で修正に要する時間」、「文字イメージ1つにつきカーペット校正で修正に要する時間」などの統計情報の各パラメータの空間で文字をクラスタを形成する、(3)ひらがな、カタカナ、漢字、数字、英字、記号などの文字種ごとにクラスタを形成する、などの手法があり得る。

20

【0103】

ログ情報またはログ情報に相当する教師データから統計情報の各パラメータは、以下のようにして算出することもできる。まず、 a は「カーペット校正にて修正された回数」を「これまで処理した文字イメージの数」で除算することによって算出することができる。 b は「サイド・バイ・サイド校正にて修正した回数」を「これまで処理した文字イメージの数」で除算することによって算出することができる。 c は、1からその時点での a 、 b を減じる（すなわち、 $c = 1 - a - b$ ）ことによって算出することができる。

30

【0104】

u 、 w' は以下のように求めることができる。すなわち、ある1回のカーペット校正について、当該校正に要した時間を T 、確認した文字イメージの数を α 、正誤不明のマークが付された文字イメージの数を β 、修正された文字イメージの数を γ とする。また、当該1回のカーペット校正において、文字イメージ1つにつきOCRの正誤が不明であることを示すマークを付するために要する時間を U 、文字イメージ1つにつきOCRの正誤が不明であることを示すマークを付するために要する時間を W' 、イメージの文字認識の正誤を確認するための時間を V とする。すると、 $T = \alpha V + \beta U + \gamma W'$ が成り立ち、この式を最小二乗法で線形回帰をすることによって、 V 、 U 、 W' を求めることができる。そして、すべてのカーペット校正について、こうして求められた U 、 W' の平均値を求めることによって、 u 、 w' をそれぞれ算出することができる。

40

【0105】

また、本発明の実施形態では、 w は、 $\alpha = 0$ 、 $\beta = 0$ 、 $\gamma = 1$ であるカーペット校正のログのみの平均値を求めるか、または、 $\gamma = 1$ となるサンプルについて、それまでに算出された U 、 V の最新の推定値を用いて W' を求め、それらの平均値を求めることによって算出することができる。

50

【0106】

なお、オペレータが作業の途中で休憩した場合などは、Tの値が極端に大きくなっている場合があり、かかる異常値を有するサンプルを考慮すると、u、w'、wの値が正しく求められない可能性がある。したがって、本発明の実施形態では、所定の閾値よりも大きなTを有するものが、u、w'、wの算出の基礎から除外されるものとする。

【0107】

また、本発明の実施形態では、カーペット校正ツールの校正の単位を「文字」として説明をしたが、複数の文字や単語などのより大きな単位を校正の単位としてもよいし、また、そのほかの属性を校正の単位とすることができる。本発明の実施形態では、カーペット校正ツール対象のイメージに対して「確認済み」「正誤不明」「修正済」のいずれかのマークを付することとしたが、当該カーペット校正完了の時点でいずれのマークも付されていないイメージがある場合はそれらに特定のマーク（例えば、「確認済み」マーク）が追加されるようにしてもよい。

10

【0108】

また、本発明の実施形態では、サンプル数が少ないことに起因した統計上の偏りの影響を少なくするために、特定の補正統計情報を算出し、それに基づいて「カーペット校正部220およびサイド・バイ・サイド校正機能225の両方で校正を行った場合に要する時間の推測値」と「カーペット校正部220を使用せずにサイド・バイ・サイド校正機能225で校正を行った場合に要する時間の推測値」を算出した。しかし、かかる算出は例示に過ぎず、かかる2つの比較される推測値は、補正がされない統計上情報を用いて算出されてもよいし、他のいかなる方法によって算出されてもよい。上記のような変更または改良を加えた形態も当然に本発明の技術的範囲に含まれる。

20

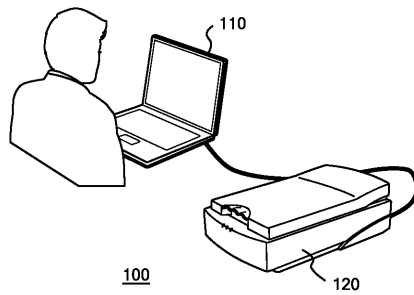
【符号の説明】

【0109】

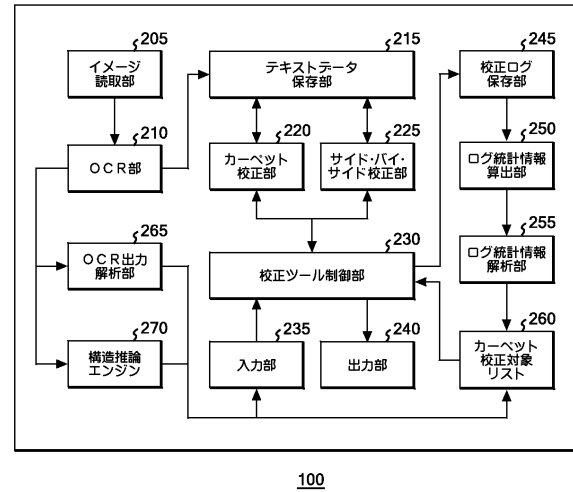
110…オペレータ端末

120…スキャナ

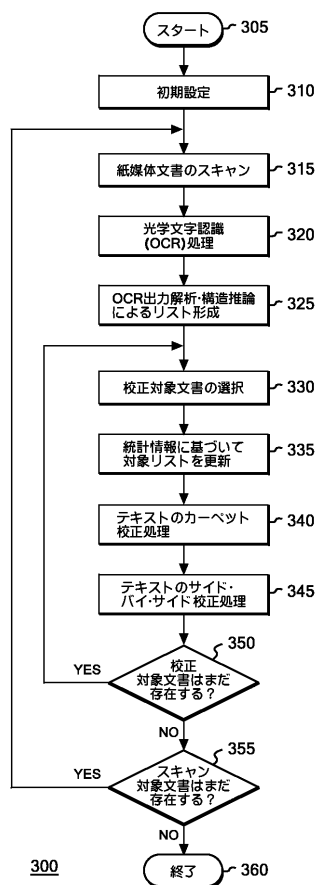
【図 1】



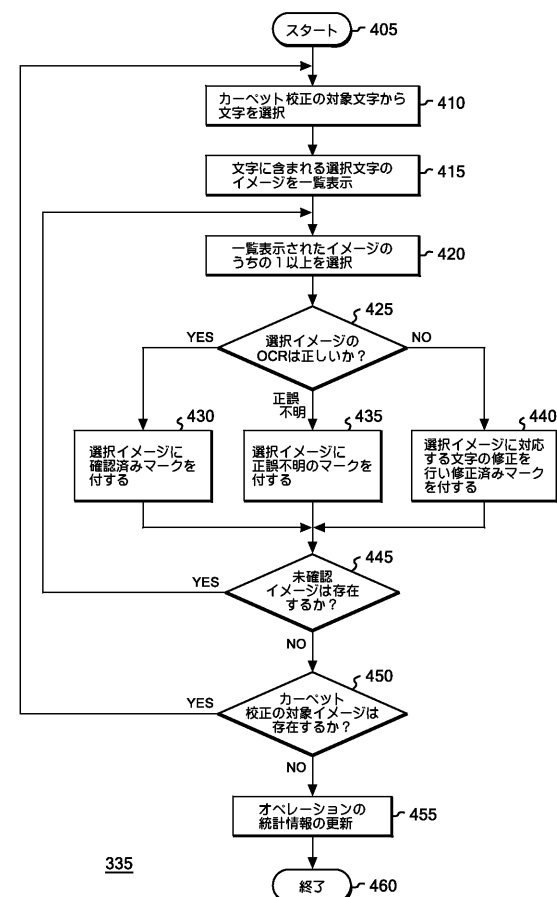
【図 2】



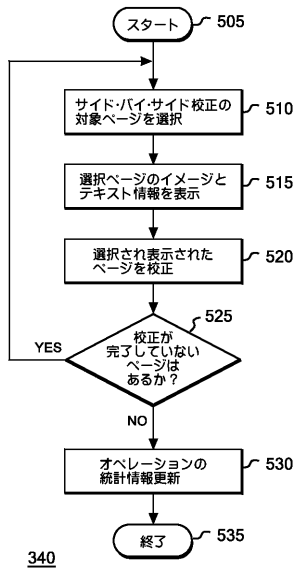
【図 3】



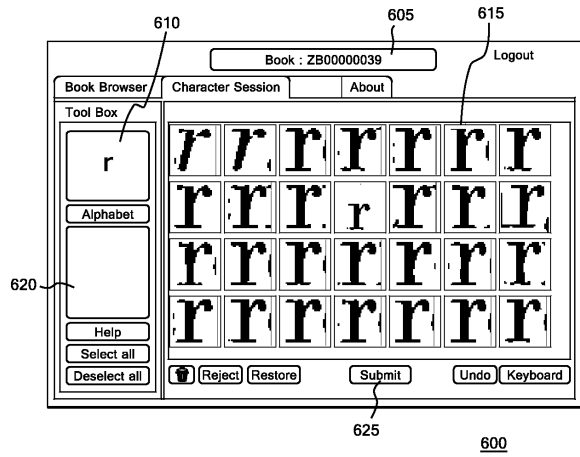
【図 4】



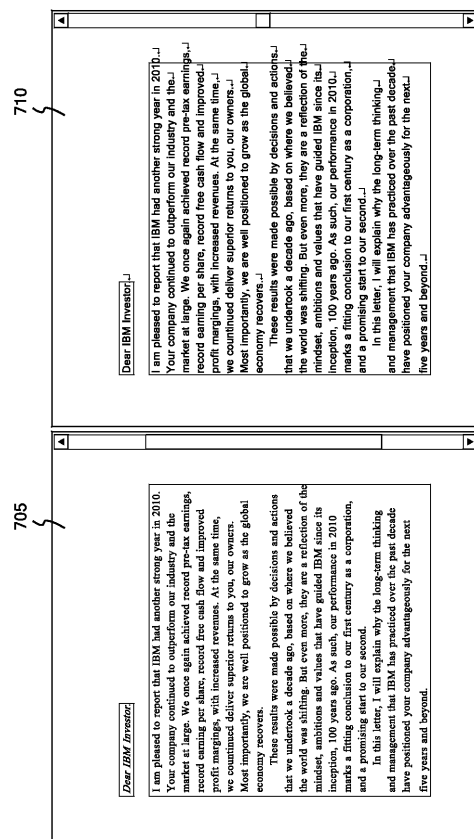
【図 5】



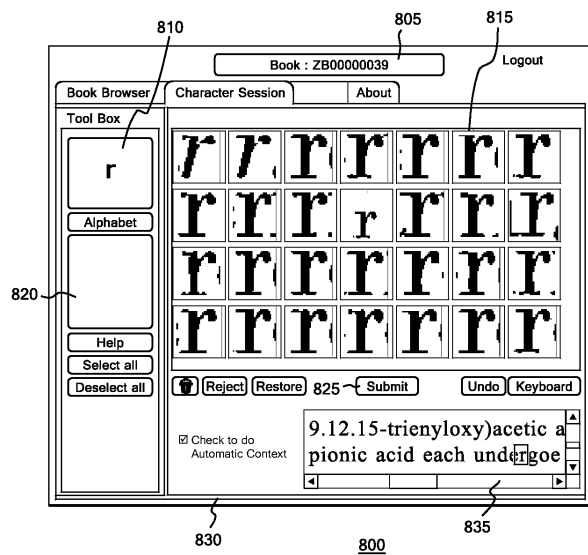
【図 6】



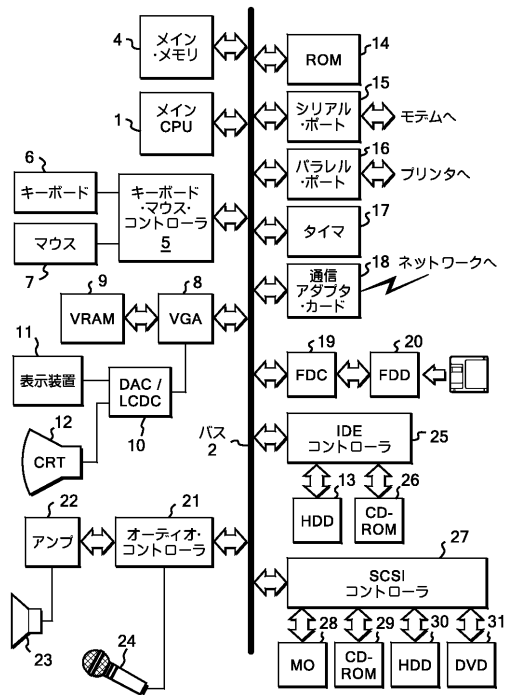
【図 7】



【図 8】



【図 9】



フロントページの続き

- (72)発明者 井床 利生
神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 東京基礎研究所内
- (72)発明者 伊藤 隆
神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 東京基礎研究所内
- (72)発明者 恐神 貴行
神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 東京基礎研究所内

審査官 鹿野 博嗣

- (56)参考文献 特開2005-309608 (JP, A)
特開平11-110480 (JP, A)
特開2002-073827 (JP, A)
特開2004-341754 (JP, A)
特開平11-353311 (JP, A)

- (58)調査した分野(Int.Cl., DB名)
G06K 9/03