

(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2010-26667

(P2010-26667A)

(43) 公開日 平成22年2月4日(2010.2.4)

(51) Int.Cl.
G06F 17/21 (2006.01)F I
G06F 17/21 530Aテーマコード (参考)
5B109

審査請求 有 請求項の数 11 O L (全 20 頁)

(21) 出願番号 特願2008-185451 (P2008-185451)
(22) 出願日 平成20年7月17日 (2008.7.17)(71) 出願人 000005496
富士ゼロックス株式会社
東京都港区赤坂九丁目7番3号
(74) 代理人 100115129
弁理士 清水 昇
(74) 代理人 100102716
弁理士 在原 元司
(74) 代理人 100122275
弁理士 竹居 信利
(72) 発明者 久保田 聡
神奈川県足柄上郡中井町境430 グリー
ンテクノikai 富士ゼロックス株式会社内
(72) 発明者 木村 俊一
神奈川県足柄上郡中井町境430 グリー
ンテクノikai 富士ゼロックス株式会社内
Fターム(参考) 5B109 NA01 QA06

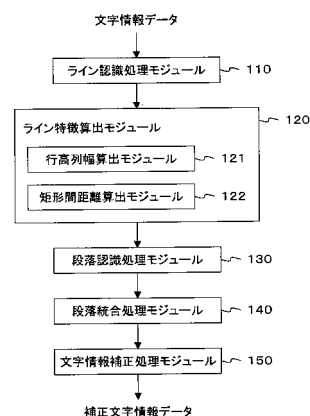
(54) 【発明の名称】 情報処理装置及び情報処理プログラム

(57) 【要約】

【課題】電子文書内の画素塊を囲む矩形の位置及び大きさを揃えて、ラインの形状を整えるようにした情報処理装置を提供する。

【解決手段】情報処理装置のライン抽出手段は、電子文書内の画素塊を囲む矩形に関する情報を用いて、該電子文書内の行又は列であるラインを抽出し、段落抽出手段は、前記ライン抽出手段によって抽出されたラインに応じて、前記電子文書内の段落を抽出し、段落統合手段は、前記段落抽出手段によって抽出された段落を統合し、矩形算出手段は、前記段落統合手段で統合された段落中のラインである行の高さ又は列の幅、及びラインを構成する画素塊間の間隔に基づいて、該統合された段落内の画素塊を囲む矩形の位置及び大きさを算出する。

【選択図】図1



【特許請求の範囲】**【請求項 1】**

電子文書内の画素塊を囲む矩形に関する情報を用いて、該電子文書内の行又は列であるラインを抽出するライン抽出手段と、

前記ライン抽出手段によって抽出されたラインに応じて、前記電子文書内の段落を抽出する段落抽出手段と、

前記段落抽出手段によって抽出された段落を統合する段落統合手段と、

前記段落統合手段で統合された段落中のラインである行の高さ又は列の幅、及びラインを構成する画素塊間の間隔に基づいて、該統合された段落内の画素塊を囲む矩形の位置及び大きさを算出する矩形算出手段

を具備することを特徴とする情報処理装置。

10

【請求項 2】

前記電子文書内の画素塊を囲む矩形に関する情報として、該画素塊を囲む矩形の高さ又は幅方向の位置を含み、

前記ライン抽出手段は、前記画素塊の矩形の高さ又は幅方向の位置を用いて、該画素塊を含むラインである各行の高さ又は各列の幅を抽出する

ことを特徴とする請求項 1 に記載の情報処理装置。

【請求項 3】

前記段落抽出手段は、前記ライン抽出手段によって抽出されたラインである各行の高さ又は各列の幅、及び該ラインの高さ又は幅方向の位置を用いて段落を抽出する

ことを特徴とする請求項 1 又は 2 に記載の情報処理装置。

20

【請求項 4】

前記段落抽出手段は、前記ライン抽出手段によって抽出されたラインと、処理対象としている段落との位置関係に基づいて段落を抽出する

ことを特徴とする請求項 1 から 3 のいずれか一項に記載の情報処理装置。

【請求項 5】

前記段落抽出手段は、抽出した段落に関する情報として、該段落を囲む外接矩形の位置に関する情報を算出する

ことを特徴とする請求項 1 から 4 のいずれか一項に記載の情報処理装置。

【請求項 6】

前記段落抽出手段は、同一行又は同一列に属するラインが複数存在する場合には、該ラインを順序付けする

ことを特徴とする請求項 1 から 5 のいずれか一項に記載の情報処理装置。

30

【請求項 7】

前記段落抽出手段は、抽出した段落に関する情報として、該段落に含まれるラインである各行の高さ又は各列の幅を用いて、該段落の代表値を算出し、

前記段落統合手段は、前記段落抽出手段で抽出された段落の代表値を用いて段落を統合する

ことを特徴とする請求項 1 から 6 のいずれか一項に記載の情報処理装置。

【請求項 8】

前記矩形算出手段は、前記段落統合手段によって統合された段落内で、ラインである行の高さ又は列の幅を統一し、画素塊間に隙間が生じないように、該統合された段落内の画素塊を囲む矩形の位置及び大きさを算出する

ことを特徴とする請求項 1 から 7 のいずれか一項に記載の情報処理装置。

40

【請求項 9】

前記矩形算出手段は、前記電子文書内に同等の形状の画素塊がある場合には、該画素塊を囲む矩形の位置及び大きさを同等の値とする

ことを特徴とする請求項 1 から 8 のいずれか一項に記載の情報処理装置。

【請求項 10】

前記矩形算出手段は、前記電子文書内の文字の言語に基づいて、前記画素塊を囲む矩形

50

の大きさを算出する

ことを特徴とする請求項 1 から 9 のいずれか一項に記載の情報処理装置。

【請求項 11】

コンピュータを、

電子文書内の画素塊を囲む矩形に関する情報を用いて、該電子文書内の行又は列であるラインを抽出するライン抽出手段と、

前記ライン抽出手段によって抽出されたラインに応じて、前記電子文書内の段落を抽出する段落抽出手段と、

前記段落抽出手段によって抽出された段落を統合する段落統合手段と、

前記段落統合手段で統合された段落中のラインである行の高さ又は列の幅、及びラインを構成する画素塊間の間隔に基づいて、該統合された段落内の画素塊を囲む矩形の位置及び大きさを算出する矩形算出手段

として機能させることを特徴とする情報処理プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、情報処理装置及び情報処理プログラムに関する。

【背景技術】

【0002】

電子ドキュメントを記述できる電子ドキュメントフォーマットが存在する。例えば、PDF (Portable Document Format) (登録商標) といわれるものがある。

このような電子ドキュメントでは、PC 上で、その電子ドキュメントを表示することが行われる。

そして、その電子ドキュメントに記述されているテキスト情報を、操作者の操作に応じて PC 上で選択し、コピー & ペースト等の処理が行われる。テキスト情報を PC 上で選択する (例えば、電子ドキュメントを表示しているディスプレイ上に表示されているテキスト位置でマウスを左クリックしながらテキスト位置を右に移動させる等の動作でテキスト情報を選択することができる) 場合、選択したテキスト位置が反転して、どのテキストを選択しているかを示すようなビューワが存在している。

一方、画像を文字認識して、電子ドキュメントを生成することも行われている。

【0003】

これに関連する技術として、例えば、特許文献 1 には、入力画像より文字もしくは文字の要素に外接する矩形を抽出する矩形抽出手段、該矩形抽出手段により抽出された矩形内の画像に対し、指定可能な複数のモード中より指定されたモードにしたがった変倍処理を施す変倍処理手段、前記矩形抽出手段により抽出された矩形の座標を、指定可能な複数のモード中より指定されたモードにしたがって変換する座標変換手段、及び、前記変倍後の矩形内画像を前記座標変換手段により変換後の座標にしたがって印字位置を制御し印刷する出力手段を具備する文字列整形装置が開示されている。

【0004】

また、例えば、特許文献 2 には、認識された文字を、記載されているテキストと文字の大きさや位置等を等しく、アウトフォントで描くことを実現することを目的とし、テキストコード情報と文字のレイアウト情報を有する認識文字に関する情報を得る認識手段と、文字のアウトフォントデータを保持するアウトフォントテーブルと、これを参照して、得られた認識文字に外接する矩形を、アウトラインフォントの文字ボックスと描かれる文字部領域との比で拡大し、拡大された外接矩形を文字ボックスデータとして用いて情報を修正する文字ボックス拡大手段と、認識された文字列中の区切りを判定する手段と、当該区切りの間の文字数で文字ボックスの幅を均等に割り付けて得られた幅で拡大された外接矩形からなる文字ボックスデータを修正する手段と、を有し、情報に基づいてアウトフォントにより文字を描いたときに小さくなることを回避し、また、区切られた文字列をまとめ

10

20

30

40

50

て取り扱うことを可能としていることが開示されている。

【特許文献 1】特開平 0 4 - 1 6 7 1 8 8 号公報

【特許文献 2】特開平 0 6 - 1 7 6 1 8 8 号公報

【発明の開示】

【発明が解決しようとする課題】

【 0 0 0 5 】

ところで、電子文書から画素塊である例えば文字を選択するような場合に、従来技術を用いると、その文字を選択して表示する反転形状が文字毎に個別になり、ラインとして不揃いの反転形状になってしまう。

本発明は、電子文書内の画素塊を囲む矩形の位置及び大きさを揃えて、ラインの形状を整えるようにした情報処理装置及び情報処理プログラムを提供することを目的としている。

10

【課題を解決するための手段】

【 0 0 0 6 】

かかる目的を達成するための本発明の要旨とするところは、次の各項の発明に存する。

請求項 1 の発明は、電子文書内の画素塊を囲む矩形に関する情報を用いて、該電子文書内の行又は列であるラインを抽出するライン抽出手段と、前記ライン抽出手段によって抽出されたラインに応じて、前記電子文書内の段落を抽出する段落抽出手段と、前記段落抽出手段によって抽出された段落を統合する段落統合手段と、前記段落統合手段で統合された段落中のラインである行の高さ又は列の幅、及びラインを構成する画素塊間の間隔に基づいて、該統合された段落内の画素塊を囲む矩形の位置及び大きさを算出する矩形算出手段を具備することを特徴とする情報処理装置である。

20

【 0 0 0 7 】

請求項 2 の発明は、前記電子文書内の画素塊を囲む矩形に関する情報として、該画素塊を囲む矩形の高さ又は幅方向の位置を含み、前記ライン抽出手段は、前記画素塊の矩形の高さ又は幅方向の位置を用いて、該画素塊を含むラインである各行の高さ又は各列の幅を抽出することを特徴とする請求項 1 に記載の情報処理装置である。

【 0 0 0 8 】

請求項 3 の発明は、前記段落抽出手段は、前記ライン抽出手段によって抽出されたラインである各行の高さ又は各列の幅、及び該ラインの高さ又は幅方向の位置を用いて段落を抽出することを特徴とする請求項 1 又は 2 に記載の情報処理装置である。

30

【 0 0 0 9 】

請求項 4 の発明は、前記段落抽出手段は、前記ライン抽出手段によって抽出されたラインと、処理対象としている段落との位置関係に基づいて段落を抽出することを特徴とする請求項 1 から 3 のいずれか一項に記載の情報処理装置である。

【 0 0 1 0 】

請求項 5 の発明は、前記段落抽出手段は、抽出した段落に関する情報として、該段落を囲む外接矩形の位置に関する情報を算出することを特徴とする請求項 1 から 4 のいずれか一項に記載の情報処理装置である。

【 0 0 1 1 】

40

請求項 6 の発明は、前記段落抽出手段は、同一行又は同一列に属するラインが複数存在する場合には、該ラインを順序付けすることを特徴とする請求項 1 から 5 のいずれか一項に記載の情報処理装置である。

【 0 0 1 2 】

請求項 7 の発明は、前記段落抽出手段は、抽出した段落に関する情報として、該段落に含まれるラインである各行の高さ又は各列の幅を用いて、該段落の代表値を算出し、前記段落統合手段は、前記段落抽出手段で抽出された段落の代表値を用いて段落を統合することを特徴とする請求項 1 から 6 のいずれか一項に記載の情報処理装置である。

【 0 0 1 3 】

請求項 8 の発明は、前記矩形算出手段は、前記段落統合手段によって統合された段落内

50

で、ラインである行の高さ又は列の幅を統一し、画素塊間に隙間が生じないように、該統合された段落内の画素塊を囲む矩形の位置及び大きさを算出することを特徴とする請求項 1 から 7 のいずれか一項に記載の情報処理装置である。

【0014】

請求項 9 の発明は、前記矩形算出手段は、前記電子文書内に同等の形状の画素塊がある場合には、該画素塊を囲む矩形の位置及び大きさを同等の値とすることを特徴とする請求項 1 から 8 のいずれか一項に記載の情報処理装置である。

【0015】

請求項 10 の発明は、前記矩形算出手段は、前記電子文書内の文字の言語に基づいて、前記画素塊を囲む矩形の大きさを算出することを特徴とする請求項 1 から 9 のいずれか一項に記載の情報処理装置である。

10

【0016】

請求項 11 の発明は、コンピュータを、電子文書内の画素塊を囲む矩形に関する情報を用いて、該電子文書内の行又は列であるラインを抽出するライン抽出手段と、前記ライン抽出手段によって抽出されたラインに応じて、前記電子文書内の段落を抽出する段落抽出手段と、前記段落抽出手段によって抽出された段落を統合する段落統合手段と、前記段落統合手段で統合された段落中のラインである行の高さ又は列の幅、及びラインを構成する画素塊間の間隔に基づいて、該統合された段落内の画素塊を囲む矩形の位置及び大きさを算出する矩形算出手段として機能させることを特徴とする情報処理プログラムである。

【発明の効果】

20

【0017】

請求項 1 の情報処理装置によれば、電子文書内の画素塊を囲む矩形の位置及び大きさを揃えて、ラインの形状を整えることができる。

【0018】

請求項 2 の情報処理装置によれば、予め定めた値ではなく、その電子文書内のラインに合わせて各行の高さ又は各列の幅を抽出することができる。

【0019】

請求項 3 の情報処理装置によれば、予め定めた値ではなく、その電子文書内のラインに合わせて段落を抽出することができる。

【0020】

30

請求項 4 の情報処理装置によれば、本構成を有していない場合に比較して、段落抽出の誤りを低減することができる。

【0021】

請求項 5 の情報処理装置によれば、抽出した段落に関する情報を抽出することができる。

【0022】

請求項 6 の情報処理装置によれば、同一行又は同一列に属するラインが複数存在する場合にも対応することができる。

【0023】

請求項 7 の情報処理装置によれば、予め定めた値ではなく、その電子文書内のラインに合わせて段落を統合することができる。

40

【0024】

請求項 8 の情報処理装置によれば、その電子文書内の画素塊の大きさ又は画素塊の間隔が一定でない場合であっても、ラインの形状を整えることができる。

【0025】

請求項 9 の情報処理装置によれば、同等の形状の画素塊に対しては、同等の画素塊を囲む矩形とすることができる。

【0026】

請求項 10 の情報処理装置によれば、その言語に用いられる文字の特徴に沿った画素塊を囲む矩形の大きさを算出することができる。

50

【 0 0 2 7 】

請求項 11 の情報処理プログラムによれば、電子文書内の画素塊を囲む矩形の位置及び大きさを揃えて、ラインの形状を整えることができる。

【 発明を実施するための最良の形態 】

【 0 0 2 8 】

まず、本実施の形態が対象とする電子ドキュメントについて説明する。

例えば、図 14 に示す例のように「美しい日本」という文字列が表示されている電子ドキュメント 1400 の「美しい日本」のテキストを PC 上で選択すると、図 15 に示す例のように「美しい日本」の部分が反転して（図 15 に示す選択テキスト 1501）、「美しい日本」が選択されたことをユーザに示すことができる。

10

あるいは、前述のようにテキストを選択した状態で、PC 上でコピー & ペーストを行うと、別のファイル上に「美しい日本」というテキスト情報をコピーすることが可能となる。図 16 に示す例のように、ワードプロセッサ等のような別のアプリケーションファイル（図 16 に示す電子ドキュメント 1600）上に、テキスト情報をペーストすることができる。

【 0 0 2 9 】

次に、このような電子ドキュメント内の文字形状を指定するため、PDF などのように、フォント情報を電子ドキュメント内に包含させることができるものがある。電子ドキュメントを表示あるいはプリントする場合に、電子ドキュメントを作成するユーザの意図通りの文字形状を復元するため、電子ドキュメント内にフォント情報（文字形状情報）を埋め込んでしまう。このように電子ドキュメント内にフォント情報を埋め込んでしまうことによって、同じフォント情報を持っていない電子ドキュメントの受け手（プリンタや PC）が、電子ドキュメントの作成者と全く同じ文字形状を復元することができるようになる。

20

【 0 0 3 0 】

前述したように、電子ドキュメント内の文字形状を指定するためにフォント情報（文字形状情報）を埋め込む際に、電子ドキュメントの受け手（プリンタや PC）のデバイス情報に合わせて文字部分の解像度を上げたり、編集や再利用が可能なようにアウトライン化するなどの処理が行われる。ここで文字のアウトライン化とは、文字の輪郭形状をベジエのような曲線で近似して表す方法である。

30

このように、文字形状を指定するためのフォント情報の文字部分に前述したような画像処理が施された場合に、文字部分の画像処理に応じた適切なフォント情報の更新が行われないと、電子ドキュメントをビューワで閲覧した場合のテキスト情報選択動作の挙動がオリジナルの電子ドキュメントと異なることがある。

例えば図 17 に示す例のように、「美しい日本」のテキストを選択したことを示す反転矩形形状（図 17 に示す選択テキスト 1701 から 1705）が、図 15 で示した例のように整った長方形の反転矩形形状にはならず、各文字で矩形が独立し、さらにはその大きさも異なる矩形形状となり、反転矩形形状の品質が低下する。これは文字部分の画像処理によりオリジナルのフォント情報に存在した“文字列として選択された場合の形状も考慮した”矩形情報が失われているか、あるいは適切に情報の修正が行われていないことに起因する。

40

したがって、反転矩形形状を整ったものにするためには、電子ドキュメント内に埋め込む文字矩形情報を適切に修正する必要がある。

本実施の形態によって出力される電子ドキュメントは、フォント情報が埋め込まれており、その文字列を選択した場合における反転矩形形状の品質劣化が抑制されたものである。

【 0 0 3 1 】

次に、本実施の形態の概要について説明する。

本実施の形態では、電子ドキュメントに埋め込まれるフォント情報内の矩形情報の修正を、文字毎の情報にのみ基づくのではなく、その電子ドキュメント全体から矩形情報の修

50

正に必要な情報を抽出又は算出（段落の抽出、その段落の統合処理等を含む）して、それらに基づいて文字毎の矩形の修正を行うものである。

【0032】

具体的には、横書きの電子ドキュメントの場合は以下（A1からA6）のようにする。

（A1）電子ドキュメント内の文字外接矩形情報（その電子ドキュメント内の座標値（絶対座標値又は相対座標値のいずれであってもよい）及び矩形サイズ（例えば、その矩形の高さ、幅の組））から行を抽出する。なお、文字外接矩形情報とは、電子ドキュメント内の文字を囲む矩形（外接矩形）に関する情報である。

（A2）行に関する特徴情報（例えば、行中の全ての文字外接矩形が収まるような最小値、行矩形サイズ、行座標値等）を求める。

（A3）複数の行からなる段落を行に関する特徴情報に基づいて抽出し、その段落に関する特徴を算出する。

（A4）その算出された段落に関する特徴に基づいて、複数の段落を統合する。

（A5）統合された段落に含まれる各行の行に関する特徴情報から矩形高さ及び矩形幅の修正値を決定する。

（A6）文字外接矩形情報、矩形高さ及び矩形幅の修正値に基づいて文字毎の矩形情報を修正する。

【0033】

また、縦書きの電子ドキュメントの場合は以下（B1からB6）のようにする。

（B1）電子ドキュメント内の文字外接矩形情報（その電子ドキュメント内の座標値（絶対座標値又は相対座標値のいずれであってもよい）及び矩形サイズ（例えば、その矩形の高さ、幅の組））から列を抽出する。なお、文字外接矩形情報とは、電子ドキュメント内の文字を囲む矩形（外接矩形）に関する情報である。

（B2）列に関する特徴情報（例えば、列中の全ての文字外接矩形が収まるような最小値、列矩形サイズ、列座標値等）を求める。

（B3）複数の列からなる段落を列に関する特徴情報に基づいて抽出し、その段落に関する特徴を算出する。

（B4）その算出された段落に関する特徴に基づいて、複数の段落を統合する。

（B5）統合された段落に含まれる各列の列に関する特徴情報から矩形高さ及び矩形幅の修正値を決定する。

（B6）文字外接矩形情報、矩形高さ及び矩形幅の修正値に基づいて文字毎の矩形情報を修正する。

【0034】

本実施の形態では、電子ドキュメント内の文字外接矩形情報から、行又は列を抽出し、その抽出した行又は列に基づいて、文字列を選択した場合にでも反転矩形形状が揃うように文字矩形情報の幅又は高さの修正を行い、文字列選択時の反転矩形形状の劣化を抑えるものである。

【0035】

以下、図面に基づき本発明を実現するにあたっての好適な一実施の形態の例を説明する。

図1は、本実施の形態の構成例についての概念的なモジュール構成図を示している。

なお、モジュールとは、一般的に論理的に分離可能なソフトウェア（コンピュータ・プログラム）、ハードウェア等の部品を指す。したがって、本実施の形態におけるモジュールはコンピュータ・プログラムにおけるモジュールのことだけでなく、ハードウェア構成におけるモジュールも指す。それゆえ、本実施の形態は、コンピュータ・プログラム、システム及び方法の説明をも兼ねている。ただし、説明の都合上、「記憶する」、「記憶させる」、これらと同等の文言を用いるが、これらの文言は、実施の形態がコンピュータ・プログラムの場合は、記憶装置に記憶させる、又は記憶装置に記憶させるように制御するの意である。また、モジュールは機能にほぼ一対一に対応しているが、実装においては、1モジュールを1プログラムで構成してもよいし、複数モジュールを1プログラムで構成

10

20

30

40

50

してもよく、逆に 1 モジュールを複数プログラムで構成してもよい。また、複数モジュールは 1 コンピュータによって実行されてもよいし、分散又は並列環境におけるコンピュータによって 1 モジュールが複数コンピュータで実行されてもよい。なお、1 つのモジュールに他のモジュールが含まれていてもよい。また、以下、「接続」とは物理的な接続の他、論理的な接続（データの授受、指示、データ間の参照関係等）の場合にも用いる。

また、システム又は装置とは、複数のコンピュータ、ハードウェア、装置等がネットワーク（一対一対応の通信接続を含む）等の通信手段で接続されて構成されるほか、1 つのコンピュータ、ハードウェア、装置等によって実現される場合も含まれる。「装置」と「システム」とは、互いに同義の用語として用いる。「予め定められた」とは、対象として

10

【0036】

以下、行又は列をラインと称する。また、横書きの電子ドキュメントを対象とした場合を主に説明する。したがって、横書きの場合の行の高さ又は縦書きの場合の列の幅として、行の高さを主に例示して説明する。

また、画素塊とは、4 連結又は 8 連結で連続する画素領域を少なくとも含み、これらの画素領域の集合をも含む。これらの画素領域の集合とは、4 連結等で連続した画素領域が複数あり、その複数の画素領域は近傍にあるものをいう。ここで、近傍にあるものとは、例えば、互いの画素領域が距離的に近いもの、文章としての 1 行から 1 文字ずつ切り出すように縦又は横方向に射影し、空白地点で切り出した画像領域、又は予め定められた間隔で切り出した画像領域等がある。例えば、文字認識処理を行って、1 文字として認識された画像を 1 つの画素塊としてもよい。

20

なお、1 つの画素塊として、1 文字の画像となる場合が多い。本実施の形態では、画素塊のことを文字又は文字画像ともいう。

【0037】

本実施の形態は、図 1 に示すように、ライン認識処理モジュール 110、ライン特徴算出モジュール 120、段落認識処理モジュール 130、段落統合処理モジュール 140、文字情報補正処理モジュール 150 を有している。

【0038】

ライン認識処理モジュール 110 は、ライン特徴算出モジュール 120 と接続されており、文字情報データを用いて、その電子ドキュメント内の行又は列であるラインを抽出し、その抽出したラインに関する情報をライン特徴算出モジュール 120 へ渡す。

30

【0039】

ライン認識処理モジュール 110 について、より詳細に説明する。

ライン認識処理モジュール 110 は、文字情報データを受け付ける。ここでいう文字情報データとは、電子ドキュメント内の画素塊の矩形に関する情報を少なくとも含む。例えば、前述の文字外接矩形情報、フォント情報であってもよい。また、画素塊に対応している文字の認識順序に関する情報（文字認識装置によって認識順に順序付けられた番号）が含まれていてもよい。例えば、電子ドキュメント中における文字の座標（例えば、文字を囲む外接矩形の左上座標）、文字の大きさを表す外接矩形サイズ（外接矩形幅、高さ）、文字形状、文字コード、文字の順序情報、縦書き文字なのか横書き文字なのかを表す情報などである。本実施の形態では、これらの文字情報データを文字認識装置から受け取った場合について説明する。ただし、文字認識装置に限る必要はなく、文字の外接矩形を受け取って、同等の文字情報データを生成するようにしてもよい。

40

【0040】

次にライン認識処理モジュール 110 は、受け付けた文字情報データに基づいて、電子ドキュメント内のラインを抽出する。例えば、横書きの場合、外接矩形の高さ方向の位置（y 座標）を用いて、その外接矩形を含むラインである各行の高さを抽出する。縦書きの場合、外接矩形の幅方向の位置（x 座標）を用いて、その外接矩形を含むラインである各

50

列の幅を抽出する。より詳細な例について、図 2、図 3 に行抽出の手法例を示す。

【 0 0 4 1 】

図 2 は、ライン認識処理モジュール 1 1 0 が外接矩形の座標値に基づいて行を認識する手法の例を示している。

ライン認識処理モジュール 1 1 0 は、図 2 (a) の例に示すように、注目文字情報データの外接矩形 (注目外接矩形 2 1 2) の左上 y 座標 ($upper_y$) が、その一つ前の文字情報データの外接矩形 (注目外接矩形 2 1 1) の左下 y 座標 ($lower_y$) より小さいときは ($upper_y < lower_y$)、その注目文字情報データの外接矩形 (注目外接矩形 2 1 2) は、注目外接矩形 2 1 1 と同じ行であると認識する。なお、左上を原点 (0 , 0) として、x 座標は右方向へ、y 座標は下方向へ向かうと数値が増える座標系である。

また、図 2 (b) の例に示すように、注目文字情報データの外接矩形 (注目外接矩形 2 2 2) の左上 y 座標 ($upper_y$) が、その一つ前の文字情報データの外接矩形 (注目外接矩形 2 2 1) の左下 y 座標 ($lower_y$) より大きいときは ($lower_y < upper_y$)、異なる行であると認識する。

そして、同じライン内にあると認識された文字情報データの列をライン特徴算出モジュール 1 2 0 へ渡す。

なお、受け付けた文字情報データは、文字画像の外接矩形の出現順序 (例えば、横書きの場合は、左上から右へ走査し、次の行ではまた左から右へ走査した順番に並んでいる) となっているので、一つ前の文字情報データの外接矩形とは、出現順序で一つ前である。また、外接矩形の左上の座標を用いてソートしてもよい。

【 0 0 4 2 】

図 3 は、ライン認識処理モジュール 1 1 0 が外接矩形間の距離に基づいて行を認識する手法の例を示している。

ライン認識処理モジュール 1 1 0 は、図 3 (a) の例に示すように、注目文字情報データの外接矩形 (注目外接矩形 3 0 3) と、その一つ前の文字情報データの外接矩形 (外接矩形 3 0 2) との外接矩形間距離 3 1 1 (以下、現外接矩形間距離ともいう) が、現在処理している行において、既に同一行であると認識された各外接矩形間の距離の平均値 (以下、平均外接矩形間距離ともいう) を α 倍した値以下である場合は (つまり、現外接矩形間距離 $\leq$ 平均外接矩形間距離 $\times \alpha$ を満たす場合)、注目外接矩形 3 0 3 は外接矩形 3 0 2 と同じ行であると認識する。なお、 α は、ライン認識パラメータであり、予め定められた値である。例えば、文字情報データに応じて定められる。

また、図 3 (b) の例に示すように、注目文字情報データの外接矩形 (注目外接矩形 3 2 3) と、その一つ前の文字情報データの外接矩形 (外接矩形 3 2 2) との外接矩形間距離 3 3 1 が、現在処理している行における平均外接矩形間距離を α 倍した値より大である場合は (現外接矩形間距離 $>$ 平均外接矩形間距離 $\times \alpha$)、注目外接矩形 3 2 3 は外接矩形 3 2 2 とは異なる行であると認識する。

【 0 0 4 3 】

ライン特徴算出モジュール 1 2 0 は、ライン認識処理モジュール 1 1 0、段落認識処理モジュール 1 3 0 と接続されており、行高列幅算出モジュール 1 2 1、矩形間距離算出モジュール 1 2 2 を有している。ライン認識処理モジュール 1 1 0 により同じラインと認識された文字情報データを受け取り、そのラインに関する特徴を算出し、その算出したラインに関する情報を段落認識処理モジュール 1 3 0 へ渡す。行高列幅算出モジュール 1 2 1 は行の高さを算出し、矩形間距離算出モジュール 1 2 2 は矩形間の距離を算出する。

つまり、ライン認識処理モジュール 1 1 0 によって同じ行と認識された文字情報データ列から行高さ、行幅、行外接矩形座標、平均外接矩形間距離などのラインに関する特徴を算出する。

【 0 0 4 4 】

ライン特徴算出モジュール 1 2 0 は、同じ行に属する文字情報データの外接矩形を含む矩形を求める。例えば、図 4 の例に示すように、同一行内の外接矩形 4 0 1 から外接矩形

10

20

30

40

50

4 1 9 を囲む矩形である行外接矩形 4 5 0 を求める。そして、行外接矩形座標として、図 4 で示すように行外接矩形の左上 ($\min_x$, $\min_y$)、右下 ($\max_x$, $\max_y$) を求める。

また、行高列幅算出モジュール 1 2 1 は、行高さ (h) を先に求めた行外接矩形座標を用いて、 $h = \max_y - \min_y$ として求める。同様に、行幅 (w) を行外接矩形座標を用いて、 $w = \max_x - \min_x$ として求める。これらの行高さ、行幅は、各外接矩形のサイズ (高さ、幅) 又はその座標を用いて求める。

また、矩形間距離算出モジュール 1 2 2 は、平均文字外接矩形間距離を、同じ行に属する隣接する文字情報データの外接矩形間距離 g_0 , g_1 , ... , g_n の平均値として求める。また、最大外接矩形間距離 $\max_g$ を、 g_0 , g_1 , ... , g_n のうちの最大値として求める。なお、リストデータとして g_0 , g_1 , ... , g_n のそれぞれの値も保持するようにしてもよい。

【 0 0 4 5 】

段落認識処理モジュール 1 3 0 は、ライン特徴算出モジュール 1 2 0、段落統合処理モジュール 1 4 0 と接続されており、ライン認識処理モジュール 1 1 0 で認識された各行、それら各行についてライン特徴算出モジュール 1 2 0 で算出されたライン特徴量から、電子ドキュメント内の段落を抽出し、その段落情報を算出する。また、横書きの場合、段落の抽出は、ライン認識処理モジュール 1 1 0 によって抽出された各行の高さ及びそのラインの座標 (高さ方向の位置 (y 座標)) を用いて行うようにしてもよい。縦書きの場合、段落の抽出は、ライン認識処理モジュール 1 1 0 によって抽出された各列の幅及びそのラインの座標 (幅方向の位置 (x 座標)) を用いて行うようにしてもよい。また、ライン認識処理モジュール 1 1 0 によって抽出されたラインと、処理対象としている段落との位置関係に基づいて段落を抽出するようにしてもよい。また、その抽出した段落に関する情報として、その段落を囲む外接矩形の位置に関する情報を算出してもよく、又はその段落の順序に関する情報をその段落に含まれる文字の出現順序に関する情報から算出するようにしてもよい。また、横書きの場合、同一行に属するラインが複数存在する場合には、そのラインを順序付けするようにしてもよい。縦書きの場合、同一列に属するラインが複数存在する場合には、そのラインを順序付けするようにしてもよい。段落を囲む外接矩形の位置に関する情報として、例えば、段落外接矩形の左上角の座標値、段落外接矩形の幅、高さ等がある。また、段落認識処理モジュール 1 3 0 は、それが認識した段落に関する情報として、その段落に含まれるラインの高さ又は幅 (横書きの場合は各行の高さ、縦書きの場合は各列の幅) を用いて、その段落の代表値を算出するようにしてもよい。段落の代表値として、より具体的には、横書きの場合、同一段落として認識された段落内に含まれている行のうち、最も行高さが大きい値をいう。縦書きの場合、同一段落として認識された段落内に含まれている列のうち、最も列幅が大きい値をいう。

【 0 0 4 6 】

図 5 は、本実施の形態による段落の認識処理例を示すフローチャートである。つまり、段落認識処理モジュール 1 3 0 が行う処理例を示すものである。

ステップ S 5 0 2 では、まず初めにライン認識処理モジュール 1 1 0 で認識された行に関して、行外接矩形の y 座標値である $\min_y$ 値で昇順にソートする。

ステップ S 5 0 4 では、ステップ S 5 0 2 でソートされた行を全て探索 (ステップ S 5 0 6 からステップ S 5 1 4 までの処理) したかどうかを判定する。全て探索されていればステップ S 5 1 6 に、探索が終了していなければステップ S 5 0 6 に処理を移す。

ステップ S 5 0 6 では、注目する行 (以降は、現探索行ともいう) をソート順に選択する。

ステップ S 5 0 8 では、現探索行に関して段落に登録されているかどうかを判定する。現探索行が段落に登録されているならば処理をステップ S 5 0 4 に戻し、登録されていないならばステップ S 5 1 0 に処理を移す。

【 0 0 4 7 】

ステップ S 5 1 0 では、現探索行が現段落における最初の登録行かどうかを判定する。

現探索行が現段落における最初の登録行であれば処理をステップ S 5 1 4 に移し、最初の登録行でなければステップ S 5 1 2 に処理を移す。

ステップ S 5 1 2 では、現段落に対して現探索行が登録できるかどうかを判定する。現探索行が現段落に登録可能ならば処理をステップ S 5 1 4 に移し、登録できないならば処理をステップ S 5 0 4 に戻す。なお、ステップ S 5 1 2 における現探索行の登録可否処理の詳細は、図 7 を用いて後で詳しく説明する。

【 0 0 4 8 】

ステップ S 5 1 4 では、現段落に対して、最初の登録行であるか、又は登録可能な行であると、それぞれステップ S 5 1 0 又はステップ S 5 1 2 で判定されている現探索行を現段落に登録し、段落情報を算出又は更新する。その後、処理をステップ S 5 0 4 に移す。

ここで、図 6 に段落情報の具体的な例を示す。段落情報として、例えば、その段落の位置情報（例えば、左上座標及び右下座標）、段落順序値（その段落を読む際の順序）を含む。段落認識処理モジュール 1 3 0 は、図 6 の例に示すように、段落に登録されている行情報（登録行情報）を用いて、段落に登録された全ての行の行外接矩形（登録行 0 6 0 0 から登録行 8 6 0 8）を含む矩形を段落外接矩形 6 1 0 として、その左上座標（ min_x , min_y ）及び右下座標（ max_x , max_y ）を算出する。また図 6 には図示していないが、同一段落に登録された各行のうち、最も行高さが大きい値 max_h を算出し、段落代表値とする。同一段落に登録された文字情報データ中で最も小さい文字認識順序の値 min_order を算出し、段落順序値とする。

【 0 0 4 9 】

次に、段落情報の更新について説明する。段落認識処理モジュール 1 3 0 は、本ステップにおいて、現段落に新たな行を登録する場合は、先述の段落外接矩形座標及び段落順序値を更新する。図 6 に示す具体例では、新たに処理対象とする行を登録行 8 6 0 8 とすると、その登録行 8 6 0 8 の行外接矩形の幅は、現段落外接矩形座標の幅（ min_x , max_x ）内に収まっているので、 min_x 及び max_x は更新せず、 max_y だけ更新する（図 6 では、更新前 max_y から更新後 max_y へ更新する）。さらに、現段落の段落代表値と新たに登録される登録行 8 6 0 8 の行高さを比較して、現段落の段落代表値より登録行 8 6 0 8 の行高さが大きい場合には段落代表値 max_h も更新する。つまり、登録行 8 6 0 8 の行高さを段落代表値 max_h とし、段落代表値 max_h をその段落内で最も大きい行高さとする。さらに、現段落順序値と新たに登録される登録行 8 6 0 8 中の全ての文字情報データの文字認識順序値を比較して、現段落順序値よりも小さい値がある場合には、段落順序値 min_order をその小さい値（文字認識順序値）に更新する。

【 0 0 5 0 】

ステップ S 5 1 6 では、ステップ S 5 0 4 でソート順に行の探索を終了しているので、現段落に登録すべき行は全て登録しているとして現段落の抽出処理を終了する。

ステップ S 5 1 8 では、全ての行が段落登録されたかを判定する。全ての行がいずれかの段落に登録されていれば段落抽出処理を終了する（ステップ S 5 9 9）。いずれの段落にも登録されていない行がある場合には、処理をステップ S 5 0 4 に戻し、次の段落抽出処理を行う。

【 0 0 5 1 】

次に、図 5 の例に示したフローチャートのステップ S 5 1 2 において、段落認識処理モジュール 1 3 0 が処理する現探索行の登録可否処理例の詳細について、図 7 の例に示すフローチャートを用いて説明する。

ステップ S 7 0 2 において、現探索行が現段落の段落外接矩形に対して、右又は左にずれているかを判定する。つまり、現探索行の左端が現段落の右端より右にあるか否か、又は現探索行の右端が現段落の左端より左にあるか否かを判定する。例えば、図 8 (a) の例に示すように、現探索行 8 1 2 が現段落 8 1 0 より右にずれているかどうか、又は図 8 (b) の例に示すように、現探索行 8 3 2 が現段落 8 3 0 よりも左にずれているかどうかを判定する。現探索行が図 8 の例のように右あるいは左にずれている場合には、現探索行

は現段落に登録せず、図5の例に示したステップS504に処理を戻す。それ以外の場合は、処理をステップS704に移す。

【0052】

ステップS704において、現探索行と現段落に登録された行の文字サイズ（行高さを含む）に基づいて、その現探索行に登録すべきか否かを判定する。つまり、現探索行が現段落内の登録行より文字サイズが大きい小さいかを判定する。例えば、ステップS704における文字サイズの判定は、図9の例に示すように行高さをを用いて行う。つまり、現段落920, 950に既に登録された各行（行900から行908、行930から行938）の平均行高さと現探索行910, 940の行高さを比較して、図9(a)の例に示すように、現探索行910の行高さが平均行高さより予め定められた量より大きい場合、又は図9(b)の例に示すように、現探索行940の行高さが平均行高さより予め定められた量より小さい場合は、現探索行910, 940は現段落920, 950に登録せず、図5の例に示したステップS504に処理を戻す。それ以外の場合は、処理をステップS706に移す。

10

【0053】

ステップS706において、現探索行が現段落の段落外接矩形に対して、下にずれたかどうかを判定する。つまり、図6の例に示した現段落の段落外接矩形610のmax_y（図6では更新後max_y）と図4の例に示した現探索行の行外接矩形450のmin_yとを比較して、max_y > min_yなら処理をステップS708に移し、max_y < min_yなら図5の例に示したステップS514に処理を移して、現探索行を現段落に登録し、段落情報を更新する。

20

【0054】

ステップS708において、ステップS704と同様に現段落に登録された各行の平均行高さと現探索行の行高さを比較して、現探索行の行高さが平均行高さより予め定められた量より大きい場合、又は現探索行の行高さが平均行高さより予め定められた量より小さい場合は、現探索行は現段落に登録せず、図5の例に示したステップS504に処理を戻す。それ以外の場合は、処理をステップS710に移す。

【0055】

ステップS710において、現探索行と現段落の行間と現段落に既に登録された各行の行間とを比較する。つまり、現段落に既に登録された各行の行間の平均値と現探索行と現段落の段落外接矩形との距離（min_y - max_y）を比較して、その差分が予め定められた量より大きい場合は行間が広がったと判定して現探索行は現段落に登録せず、図5の例に示したステップS504に処理を戻す。前記差分が予め定められた量より小さい場合は、行間は一定であると判定して、ステップS712に処理を移す。

30

【0056】

ステップS712において、現探索行の一つ前の同一行に登録行が複数あるかどうかを判定し、同一行に複数登録行がある場合は、行外接矩形のx座標値であるmin_x値で昇順にソートする。ここで、同一行とは、行外接矩形のy座標が現探索行のそれと予め定められた範囲内にある行であって、ライン認識処理モジュール110では別々の行であると認識されたが、段落認識処理モジュール130による現段落の生成過程において、現探索行よりも前に登録された行（複数の場合もあり得る）のことをいう。ここで、y座標が予め定められた範囲内にあるとは、その段落において1行が存在するy座標の範囲内にあることをいう。同一行に複数登録行がない場合はそのまま図5の例に示したステップS514に処理を移して、現探索行を現段落に登録し、段落情報を更新する。図10の例は、同一行上に3つの登録行（登録行1010、登録行1011、登録行1012）がある場合を表しており、この図10の例では前記3つの登録行の行外接矩形の各x座標値である、「min_x」：登録行1010、「min_x」：登録行1011、「min_x」：登録行1012を用いて昇順でソートを行う。ソート処理終了後、図5の例に示したステップS514に処理を移して、現探索行を現段落に登録し、段落情報を更新する。

40

【0057】

50

段落統合処理モジュール 140 は、段落認識処理モジュール 130、文字情報補正処理モジュール 150 と接続されており、段落認識処理モジュール 130 によって抽出された段落を統合して、その段落に関する情報を算出する。そして、その算出した段落に関する情報を文字情報補正処理モジュール 150 に渡す。

より具体的に説明すると、段落統合処理モジュール 140 は、段落認識処理モジュール 130 で認識された段落を、各段落の段落代表値 (max_h) を用いて統合する。

【0058】

図 11 は、段落統合処理モジュール 140 が行う段落の統合処理例を示すフローチャートである。

ステップ S1102 において、段落認識処理モジュール 130 で認識された全ての段落の段落代表値 max_h の差分値を算出し、その差分値が最小となる 2 つの段落を抽出する (このときの差分値を以下では「差分最小値」ともいう)。

ステップ S1104 において、ステップ S1102 で算出された差分最小値を予め定めた閾値と比較する。前記差分最小値が予め定めた閾値より大きい場合 (ステップ S1104 で NO) は、これ以上統合すべき段落はないと判断して、段落認識処理モジュール 130 における段落統合処理を終了する (ステップ S1199)。前記差分最小値がある所定の閾値より小さい場合 (ステップ S1104 で YES) は、ステップ S1106 に処理を移す。

【0059】

ステップ S1106 において、ステップ S1102 で差分最小値であるとして抽出された 2 つの段落を統合する。ここでいう段落の統合とは、段落代表値が近い段落同士であることを示すために、例えば 2 つの段落の段落情報に同じ識別番号などを付与あるいは追加するという意味である。

ステップ S1108 において、ステップ S1106 において統合された段落の段落代表値 max_h を、統合元の 2 つの段落の段落代表値の大きい方で設定し、処理をステップ S1102 に戻す。つまり、統合後の段落の段落代表値 max_h を、元の段落の段落代表値 max_h のうち大きい値とする。

このように段落統合処理モジュール 140 は、前述したように、ステップ S1102 で算出する差分最小値がステップ S1104 において予め定めた閾値より大きくなるまでステップ S1102 からステップ S1108 の統合処理を繰り返して段落を統合する。

【0060】

文字情報補正処理モジュール 150 は、段落統合処理モジュール 140 と接続されており、段落統合処理モジュール 140 で統合された段落中のラインである行の高さ及びラインを構成する画素塊間の間隔に基づいて、その統合された段落内の画素塊を囲む矩形の位置及び大きさを算出して、ライン認識処理モジュール 110 が受け付けた文字情報データを補正する。例えば、段落統合処理モジュール 140 によって統合された段落内で、ラインである行の高さ又は列の幅を統一し、文字間に隙間が生じないように、その統合された段落内の画素塊を囲む矩形の位置及び大きさを算出するようにしてもよい。また、電子ドキュメント内に同等の形状 (つまり、文字画像として同等である場合又は外接矩形として同等である場合をいう。文字画像として同等であるとは、その文字画像の特徴を抽出し、特徴空間内で予め定めた閾値内の距離にあることをいう。外接矩形として同等であるとは、外接矩形の高さ及び幅が、他の外接矩形の幅及び高さとは予め定めた閾値以下である場合をいう) の文字がある場合には、その文字を囲む矩形の位置及び大きさを同等の値とするようにしてもよい。また、電子ドキュメント内の文字の言語に基づいて、外接矩形の大きさを算出するようにしてもよい。

【0061】

文字情報補正処理モジュール 150 は、段落統合処理モジュール 140 で統合された段落の段落代表値 max_h 及びライン特徴算出モジュール 120 で算出された最大外接矩形間距離 max_g に基づいて、矩形補正值を算出して、各行毎に分類された文字情報データを補正する。図 12 は文字情報補正処理モジュール 150 における文字情報の補正の

具体的な一例を示すものである。

文字情報補正処理モジュール 150 では、図 12 の例に示す各補正値を以下のように算出する。

補正矩形高さ H には、補正対象の文字情報データが属する統合段落の段落代表値 max_h を設定する。

補正矩形幅 W は、補正対象の文字情報データが属する行の最大外接矩形間距離 max_g を用いて、以下の式 (1) で算出する。

$$W = max_g + w \quad \dots \dots \dots \text{式 (1)}$$

ここで w は、補正前 (ライン認識処理モジュール 110 が受け付けた元の文字情報データのものの) の外接矩形幅である。

【0062】

補正矩形 1220 の左上 x 座標値 (new_x , new_y) は、以下の式 (2) で算出する。

$$\begin{aligned} new_x &= x - max_g / 2 \\ new_y &= min_y - (H - h) / 2 \quad \dots \dots \dots \text{式 (2)} \end{aligned}$$

ここで x は補正前の外接矩形 1210 の左上 x 座標値、 max_g は補正対象の文字情報データが属する行の最大外接矩形間距離の代表値、 min_y は補正対象の文字情報データが属する行の y 座標の最小値、H は補正矩形 1220 の高さ、h は補正前の外接矩形高さである。

【0063】

補正矩形 1220 から外接矩形 1210 への相対移動量 $shift_x$ 、 $shift_y$ は、以下の式 (3) で算出する。

$$\begin{aligned} shift_x &= max_g / 2 \\ shift_y &= y - new_y \quad \dots \dots \dots \text{式 (3)} \end{aligned}$$

ここで y は補正前の外接矩形 1210 の左上 y 座標値である。

【0064】

前述したように文字情報補正処理モジュール 150 では、入力された文字情報データの外接矩形情報から補正文字矩形を生成し、文字同士の矩形高さが揃い、また矩形同士の隙間が空かないような補正を行う。

【0065】

また文字情報補正処理モジュール 150 は、電子ドキュメント内に同等の形状の文字がある場合 (文字画像として同等である場合又は外接矩形として同等である場合) には、補正文字矩形の位置及び大きさを同等の値とするようにしてもよい。そして、それら補正文字矩形のデータを補正文字情報データ内で共有することによって、電子ドキュメント全体のデータ量を削減するようにしてもよい。

【0066】

また文字情報補正処理モジュール 150 は、前述した補正に加えて、電子ドキュメント内の文字の言語に基づいて、補正後の文字矩形の大きさを算出するようにしてもよい。例えば、対象とする電子ドキュメントが日本語の場合には、補正後の文字矩形が正方形となるように、補正矩形幅 W を補正矩形高さ H と等しくなるよう設定するようにしてもよい。なお、対象とする電子ドキュメント内の文字の言語の判断は、その電子ドキュメント内に含まれている言語に関するヘッダ、文字コード、画像である場合は文字認識処理の結果等を用いて行う。

【0067】

また、同等の文字情報データの場合には補正後のデータも同等として設定し、それら補正データを文字情報データに対応付けて共有することで電子ドキュメント全体のデータ量を削減するようにしてもよい。つまり、ライン認識処理モジュール 110 が受け付けた文字情報データ内で同じものを抽出し、その抽出した同等の文字情報データに対して、本実施の形態によるその文字情報データの処理結果である補正後のデータを 1 つ対応付けるようにしてもよい。

10

20

30

40

50

【 0 0 6 8 】

図 1 3 を参照して、本実施の形態のハードウェア構成例について説明する。図 1 3 に示す構成は、例えばパーソナルコンピュータ（ P C ）などによって構成されるものであり、スキャナ等のデータ読み取り部 1 3 1 7 と、プリンタなどのデータ出力部 1 3 1 8 を備えたハードウェア構成例を示している。

【 0 0 6 9 】

C P U (C e n t r a l P r o c e s s i n g U n i t) 1 3 0 1 は、前述の実施の形態において説明した各種のモジュール、すなわち、ライン認識処理モジュール 1 1 0 、ライン特徴算出モジュール 1 2 0 、段落認識処理モジュール 1 3 0 、段落統合処理モジュール 1 4 0 、文字情報補正処理モジュール 1 5 0 等の各モジュールの実行シーケンスを記述したコンピュータ・プログラムにしたがった処理を実行する制御部である。

10

【 0 0 7 0 】

R O M (R e a d O n l y M e m o r y) 1 3 0 2 は、C P U 1 3 0 1 が使用するプログラムや演算パラメータ等を格納する。R A M (R a n d o m A c c e s s M e m o r y) 1 3 0 3 は、C P U 1 3 0 1 の実行において使用するプログラムや、その実行において適宜変化するパラメータ等を格納する。これらは C P U バスなどから構成されるホストバス 1 3 0 4 により相互に接続されている。

【 0 0 7 1 】

ホストバス 1 3 0 4 は、ブリッジ 1 3 0 5 を介して、P C I (P e r i p h e r a l C o m p o n e n t I n t e r c o n n e c t / I n t e r f a c e) バスなどの外部バス 1 3 0 6 に接続されている。

20

【 0 0 7 2 】

キーボード 1 3 0 8 、マウス等のポインティングデバイス 1 3 0 9 は、操作者により操作される入力デバイスである。ディスプレイ 1 3 1 0 は、液晶表示装置又は C R T (C a t h o d e R a y T u b e) などがあり、各種情報をテキストやイメージ情報として表示する。

【 0 0 7 3 】

H D D (H a r d D i s k D r i v e) 1 3 1 1 は、ハードディスクを内蔵し、ハードディスクを駆動し、C P U 1 3 0 1 によって実行するプログラムや情報を記録又は再生させる。ハードディスクには、文字情報データや文字情報補正処理モジュール 1 5 0 の処理結果データなどが格納される。さらに、その他の各種のデータ処理プログラム等、各種コンピュータ・プログラムが格納される。

30

【 0 0 7 4 】

ドライブ 1 3 1 2 は、装着されている磁気ディスク、光ディスク、光磁気ディスク、又は半導体メモリ等のリムーバブル記録媒体 1 3 1 3 に記録されているデータ又はプログラムを読み出して、そのデータ又はプログラムを、インタフェース 1 3 0 7 、外部バス 1 3 0 6 、ブリッジ 1 3 0 5 、及びホストバス 1 3 0 4 を介して接続されている R A M 1 3 0 3 に供給する。リムーバブル記録媒体 1 3 1 3 も、ハードディスクと同様のデータ記録領域として利用可能である。

【 0 0 7 5 】

接続ポート 1 3 1 4 は、外部接続機器 1 3 1 5 を接続するポートであり、U S B 、 I E E E 1 3 9 4 等の接続部を持つ。接続ポート 1 3 1 4 は、インタフェース 1 3 0 7 、及び外部バス 1 3 0 6 、ブリッジ 1 3 0 5 、ホストバス 1 3 0 4 等を介して C P U 1 3 0 1 等に接続されている。通信部 1 3 1 6 は、ネットワークに接続され、外部とのデータ通信処理を実行する。データ読み取り部 1 3 1 7 は、例えばスキャナであり、ドキュメントの読み取り処理を実行する。データ出力部 1 3 1 8 は、例えばプリンタであり、ドキュメントデータの出力処理を実行する。

40

【 0 0 7 6 】

なお、図 1 3 に示すハードウェア構成は、1 つの構成例を示すものであり、本実施の形態は、図 1 3 に示す構成に限らず、本実施の形態において説明したモジュールを実行可能

50

な構成であればよい。例えば、一部のモジュールを専用のハードウェア（例えば特定用途向け集積回路（Application Specific Integrated Circuit：ASIC）等）で構成してもよく、一部のモジュールは外部のシステム内にあり通信回線で接続しているような形態でもよく、さらに図13に示すシステムが複数互いに通信回線によって接続されていて互いに協調動作するようにしてもよい。また、複写機、ファックス、スキャナ、プリンタ、複合機（スキャナ、プリンタ、複写機、ファックス等のいずれか2つ以上の機能を有している画像処理装置）などに組み込まれていてもよい。

【0077】

前記実施の形態においては、主に横書き電子ドキュメントの場合の行の高さを用いることを示したが、縦書きの場合は列の幅を同様に用いる。

なお、数式を用いて説明したが、数式には、その数式と同等のものを含めてもよい。同等のものとは、その数式そのものの他に、最終的な結果に影響を及ぼさない程度の数式の変形、又は数式をアルゴリズム的な解法で解くこと等が含まれる。

【0078】

なお、説明したプログラムについては、記録媒体に格納して提供してもよく、また、そのプログラムを通信手段によって提供してもよい。その場合、例えば、前記説明したプログラムについて、「プログラムを記録したコンピュータ読み取り可能な記録媒体」の発明として捉えてもよい。

「プログラムを記録したコンピュータ読み取り可能な記録媒体」とは、プログラムのインストール、実行、プログラムの流通などのために用いられる、プログラムが記録されたコンピュータで読み取り可能な記録媒体をいう。

なお、記録媒体としては、例えば、デジタル・バーサタイル・ディスク（DVD）であって、DVDフォーラムで策定された規格である「DVD-R、DVD-RW、DVD-RAM等」、DVD+RWで策定された規格である「DVD+R、DVD+RW等」、コンパクトディスク（CD）であって、読出し専用メモリ（CD-ROM）、CDレコーダブル（CD-R）、CDリライタブル（CD-RW）等、ブルーレイ・ディスク（Blue-ray Disk）、光磁気ディスク（MO）、フレキシブルディスク（FD）、磁気テープ、ハードディスク、読出し専用メモリ（ROM）、電氣的消去及び書換可能な読出し専用メモリ（EEPROM）、フラッシュ・メモリ、ランダム・アクセス・メモリ（RAM）等が含まれる。

そして、前記のプログラム又はその一部は、前記記録媒体に記録して保存や流通等させてもよい。また、通信によって、例えば、ローカル・エリア・ネットワーク（LAN）、メトロポリタン・エリア・ネットワーク（MAN）、ワイド・エリア・ネットワーク（WAN）、インターネット、イントラネット、エクストラネット等に用いられる有線ネットワーク、あるいは無線通信ネットワーク、さらにこれらの組み合わせ等の伝送媒体を用いて伝送させてもよく、また、搬送波に乗せて搬送させてもよい。

さらに、前記のプログラムは、他のプログラムの一部分であってもよく、あるいは別個のプログラムと共に記録媒体に記録されていてもよい。また、複数の記録媒体に分割して記録されていてもよい。また、圧縮や暗号化など、復元可能であればどのような態様で記録されていてもよい。

【図面の簡単な説明】

【0079】

【図1】本実施の形態の構成例についての概念的なモジュール構成図である。

【図2】ライン認識処理モジュールによるラインの抽出処理例を示す説明図である。

【図3】ライン認識処理モジュールによるラインの抽出処理例を示す説明図である。

【図4】ライン特徴算出モジュールによるライン特徴の抽出処理例を示す説明図である。

【図5】本実施の形態による段落の認識処理例を示すフローチャートである。

【図6】段落情報の更新に関する処理例を示す説明図である。

【図7】本実施の形態による段落登録の可否処理例を示すフローチャートである。

10

20

30

40

50

【図 8】左右のずれによって段落登録が行われない場合の例を示す説明図である。

【図 9】文字サイズによって段落登録が行われない場合の例を示す説明図である。

【図 10】同じ行に複数行がある状態例を示す説明図である。

【図 11】段落統合処理モジュールによる段落の統合処理例を示すフローチャートである。

【図 12】文字情報の補正の処理例を示す説明図である。

【図 13】本実施の形態を実現するコンピュータのハードウェア構成例を示すブロック図である。

【図 14】電子ドキュメントのテキストが表示されている例を示す説明図である。

【図 15】テキストを選択した状態の電子ドキュメントの表示例を示す説明図である。

10

【図 16】別のアプリケーションにテキストを複写した場合の別の電子ドキュメントの表示例を示す説明図である。

【図 17】画像処理後のフォントを埋め込んだ電子ドキュメントにおいて、テキストを選択した状態の表示例を示す説明図である。

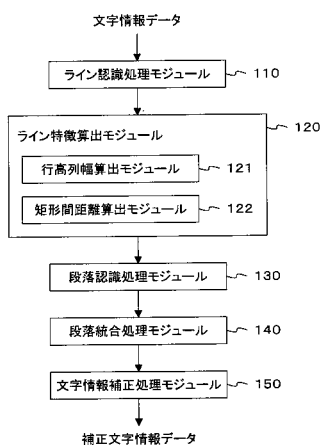
【符号の説明】

【0080】

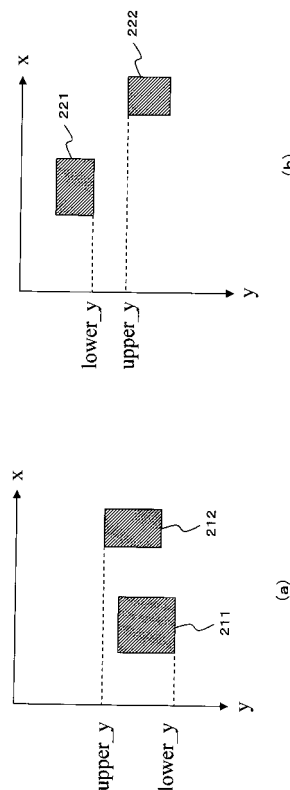
- 110 ... ライン認識処理モジュール
- 120 ... ライン特徴算出モジュール
- 121 ... 行高列幅算出モジュール
- 122 ... 矩形間距離算出モジュール
- 130 ... 段落認識処理モジュール
- 140 ... 段落統合処理モジュール
- 150 ... 文字情報補正処理モジュール

20

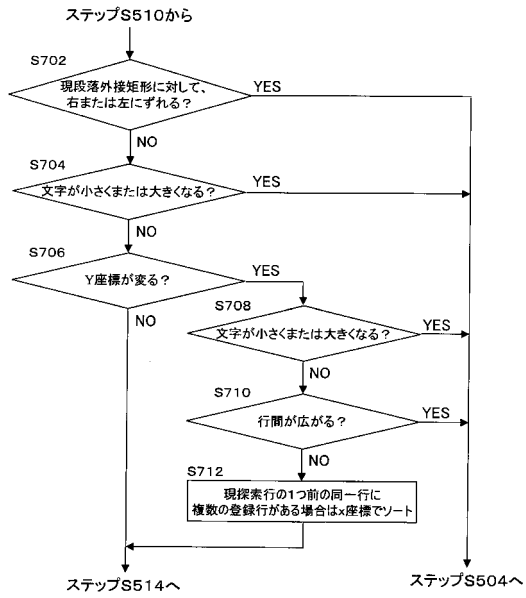
【図 1】



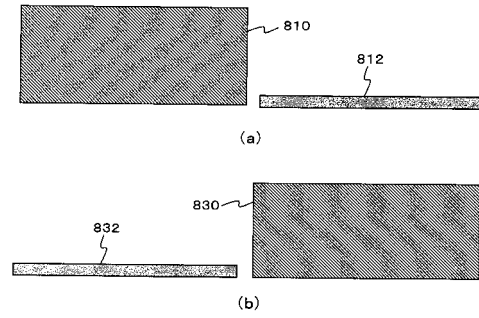
【図 2】



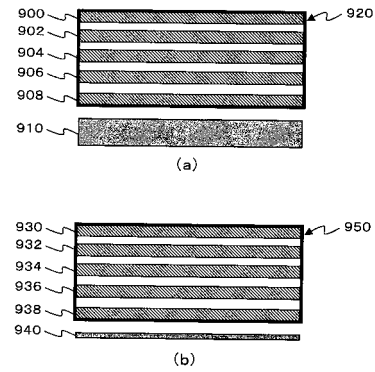
【図 7】



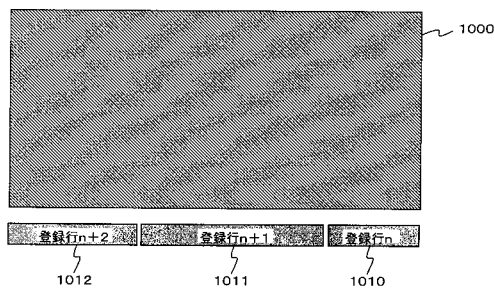
【図 8】



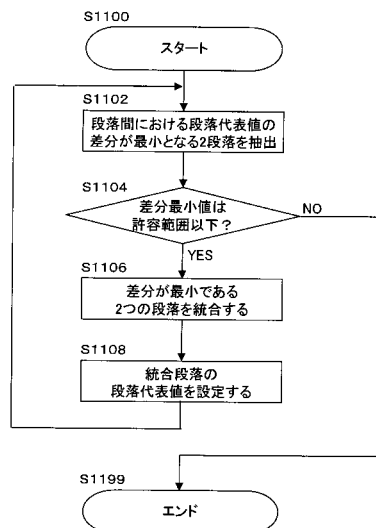
【図 9】



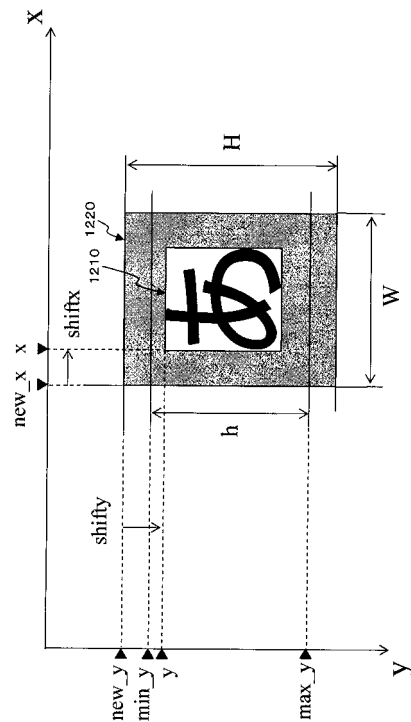
【図 10】



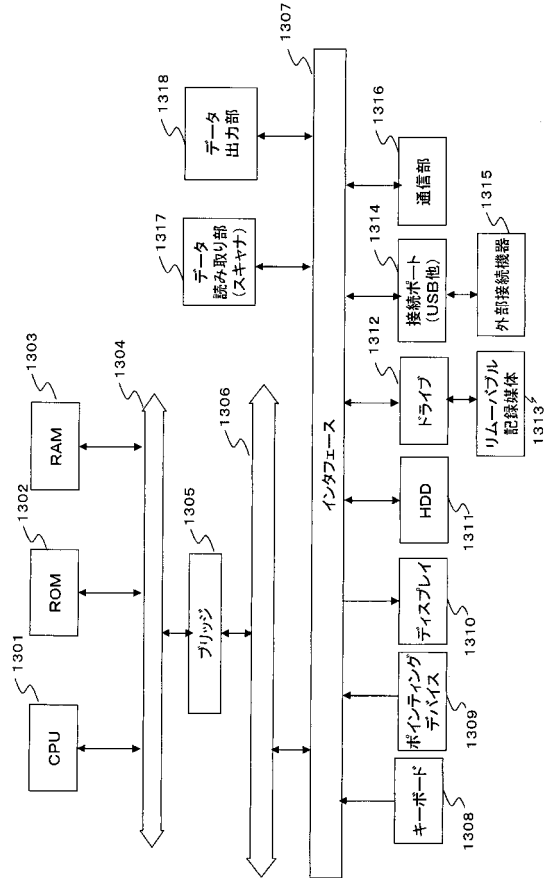
【図 11】



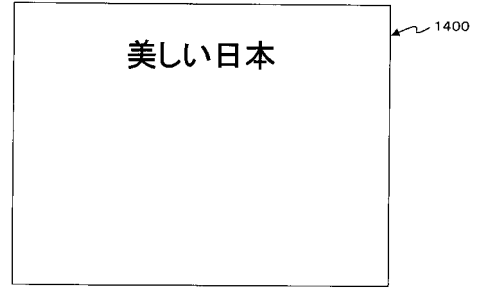
【図 12】



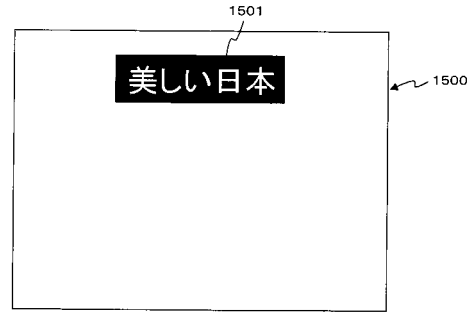
【図 13】



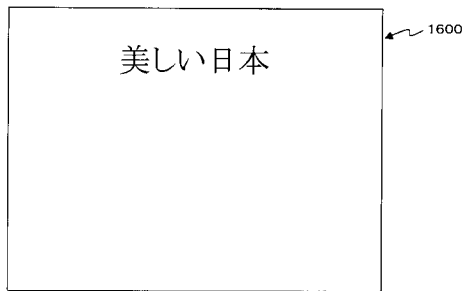
【図 14】



【図 15】



【図 16】



【図 17】

