

明 細 書

発明の名称：

多モデル深層学習による文書のデジタル化アーキテクチャ、文書画像処理プログラム

技術分野

[0001] 本発明は、電子文書生成装置、電子文書生成方法、及び電子文書生成プログラムに関し、特に紙文書を走査して電子文書を生成する電子文書生成装置、電子文書生成方法、及び電子文書生成プログラムに関するものである。

背景技術

[0002] デジタル情報技術が進展しペーパーレス化が普及してきているが、紙文書による情報の蓄積や伝達は依然として広く利用されている。膨大な紙文書を抱える企業などから、紙文書を効率良くデジタル文書に変換できる技術が望まれている。

[0003] 従来からのOCRテキスト認識技術では、一文字単位で文字認識を行っていたので文字認識の認識効率が良くない点が問題となっていた（例えば、特許文献1参照）。

先行技術文献

特許文献

[0004] 特許文献1：特開2010-244372号公報

発明の概要

発明が解決しようとする課題

[0005] そこで、本開示の電子文書生成装置、電子文書生成方法、及び電子文書生成プログラムは、文書画像に含まれる文字列を、従来の光学文字認識とは異なる手法によって、テキストデータに変換することを目的とする。

課題を解決するための手段

[0006] すなわち、第1の態様に係る電子文書生成装置は、文書を画像化した文書

画像を取得する文書画像取得部と、文書画像と当該文書画像に含まれる文字列との対応関係を学習した文字列学習モデルを用いて、文書画像取得部に取得された文書画像に含まれる文字列を文字認識し、当該文字列に係るテキストデータを生成する文字列認識部と、テキストデータを電子媒体のテキストとして出力する出力部とを備える。

[0007] 第2の態様は、第1の態様に係る電子文書生成装置において、文書画像に含まれる複数の要素と、当該複数の要素の各々の識別情報との対応関係を学習したレイアウト学習モデルを用いて、文書画像取得部に取得された文書画像に含まれる複数の要素の各々の文書画像内における範囲を特定し、複数の要素の各々の種類を認識し、複数の要素の各々の範囲に係る文書画像内における位置情報を取得するレイアウト認識部をさらに備え、文字列認識部は、レイアウト認識部により特定された範囲に含まれる文字列について、文字列学習モデルを用いて文字認識し、文字列に係るテキストデータを生成し、出力部は、複数の要素に係る範囲の各々の位置情報に、複数の要素に係るテキストデータの各々を電子媒体のテキストとして出力することとしてもよい。

[0008] 第3の態様は、第2の態様に係る電子文書生成装置において、要素の種類は、文字列、表、画像、印章、又は手書きのいずれかであることとしてもよい。

[0009] 第4の態様は、第3の態様に係る電子文書生成装置において、レイアウト認識部により認識された種類が表に該当する要素において、当該要素に含まれる表の中のセルの各々を切り出し、セルの各々の文書画像内における位置情報を取得する切出部をさらに備え、文字列認識部は、切出部に切り出されたセルの各々に含まれる文字列について、文字列学習モデルを用いて文字認識を行い、文字列に係るテキストデータを生成することとしてもよい。

[0010] 第5の態様は、第2ないし4の態様に係る電子文書生成装置において、複数の要素を含む文書画像であって、当該要素に当該要素の各々に該当する種類に関連付けられたアノテーションが付与されており、アノテーションが付与された複数の文書画像を蓄積してレイアウト学習用データを生成するレイ

アウト学習用データ生成部をさらに備え、レイアウト学習用データはレイアウト学習モデルの教師有り学習に用いられることとしてもよい。

[0011] 第6の態様は、第5の態様に係る電子文書生成装置において、文書画像に、アノテーションとともに文書画像に含まれる複数の要素に係る範囲の各々の文書画像内における位置情報が付与されることとしてもよい。

[0012] 第7の態様は、第5または6の態様に係る電子文書生成装置において、入力に基づいて、レイアウト認識部により認識された複数の要素の各々の種類、及び複数の要素の各々の範囲の文書画像内における位置情報の少なくともいずれかが修正され、この修正されたデータを追加することでレイアウト学習用データを更新するレイアウト学習用データ修正部をさらに備えることとしてもよい。

[0013] 第8の態様は、第7の態様に係る電子文書生成装置において、レイアウト学習用データ修正部により更新されたレイアウト学習用データを用いて、レイアウト学習モデルの再学習を行うレイアウト学習部をさらに備えることとしてもよい。

[0014] 第9の態様は、第2ないし8の態様に係る電子文書生成装置において、文字列学習モデルの教師有り学習に用いる文字列学習用データを生成する文字列学習用データ生成部をさらに備えることとしてもよい。

[0015] 第10の態様は、第9の態様に係る電子文書生成装置において、入力に基づいて、文字列認識部により生成されたテキストデータが修正され、この修正されたテキストデータを追加することで文字列学習用データを更新する文字列学習用データ修正部をさらに備えることとしてもよい。

[0016] 第11の態様は、第10の態様に係る電子文書生成装置において、文字列学習用データ修正部により更新された文字列学習用データを用いて、文字列学習モデルの再学習を行う文字列学習部をさらに備えることとしてもよい。

[0017] 第12の態様は、第2ないし11の態様に係る電子文書生成装置において、文字列認識部は、複数の文字列学習モデルを備え、複数の要素の各々に含まれる文字列の言語に適応した文字列学習モデルを用いることとしてもよい。

。

[0018] 第13の態様は、第2ないし12の態様に係る電子文書生成装置において、文書画像取得部が取得した文書画像について前処理を行う前処理部をさらに備え、前処理部は、背景除去部、傾き補正部、及び形状調整部を備え、背景除去部は、文書画像取得部が取得した文書画像の背景を除去し、傾き補正部は、文書画像取得部が取得した文書画像の傾きを補正し、形状調整部は、文書画像取得部が取得した文書画像の全体の形状及び大きさを調整することとしてもよい。

[0019] 第14の態様は、第2ないし13の態様に係る電子文書生成装置において、レイアウト学習モデルは、契約書用のレイアウト学習モデル、請求書用のレイアウト学習モデル、覚書用のレイアウト学習モデル、納品書用のレイアウト学習モデル、又は領収書用のレイアウト学習モデルのいずれかであることとしてもよい。

[0020] 第15の態様に係る電子文書生成方法は、電子文書生成装置に用いられるコンピュータが、文書を画像化した文書画像を取得する文書画像取得ステップと、文書画像と当該文書画像に含まれる文字列との対応関係を学習した文字列学習モデルを用いて、文書画像取得ステップにて取得された文書画像に含まれる文字列を文字認識し、当該文字列に係るテキストデータを生成する文字列認識ステップと、テキストデータを電子媒体のテキストとして出力する出力ステップとを実行する。

[0021] 第16の態様に係る電子文書生成プログラムは、電子文書生成装置に用いられるコンピュータに、文書を画像化した文書画像を取得する文書画像取得機能と、文書画像と当該文書画像に含まれる文字列との対応関係を学習した文字列学習モデルを用いて、文書画像取得機能にて取得された文書画像に含まれる文字列を文字認識し、当該文字列に係るテキストデータを生成する文字列認識機能と、テキストデータを電子媒体のテキストとして出力する出力機能とを発揮させる。

発明の効果

[0022] 本開示に係る電子文書生成装置は、文書を画像化した文書画像を取得する文書画像取得部と、文書画像と当該文書画像に含まれる文字列との対応関係を学習した文字列学習モデルを用いて、文書画像取得部に取得された文書画像に含まれる文字列を文字認識し、当該文字列に係るテキストデータを生成する文字列認識部と、テキストデータを電子媒体のテキストとして出力する出力部とを備え、機械学習されたモデルを用いて文書画像に含まれる文字列を文字認識するので、文書画像をテキストデータに変換する際の文字認識の認識効率を向上させることができる。

図面の簡単な説明

[0023] [図1]本実施形態に係る電子文書生成装置を含む電子文書生成システムの概略を示す図である。

[図2]電子文書生成装置の物理的構成を示すブロック図である。

[図3]電子文書生成装置が行う処理の概略を示す図である。

[図4]電子文書生成装置の機能的構成を示すブロック図である。

[図5]電子文書生成装置の入力データと出力データとを説明する図である。

[図6]前処理で行う背景除去を説明する図である。

[図7]前処理で行う傾き補正を説明する図である。

[図8]前処理で行う形状調整を説明する図である。

[図9]レイアウト認識処理で行う欠落解消の補正処理を説明する図である。

[図10]レイアウト認識処理で行う重なり解消の補正処理を説明する図である。

[図11]レイアウト認識処理で行うレイアウト認識を説明する図である。

[図12]レイアウト認識処理で行う表の認識を説明する図である。

[図13]セル画像の切り出しを説明する図である。

[図14]セル画像内の文字列を説明する図である。

[図15]文字列認識処理で行うテキストデータの配置を説明する図である。

[図16]文字列認識処理で行うノイズ除去を説明する図である。

[図17]アノテーションが付与されたレイアウト学習用データの例を示す図で

ある。

[図18]アノテーションが付与されたレイアウト学習用データの例を示す図である。

[図19]アノテーションが付与されたレイアウト学習用データの例を示す図である。

[図20]アノテーションが付与されたレイアウト学習用データの例を示す図である。

[図21]アノテーションが付与されたレイアウト学習用データの例を示す図である。

[図22]アノテーションが付与されたレイアウト学習用データの例を示す図である。

[図23]アノテーションが付与されたレイアウト学習用データの例を示す図である。

[図24]アノテーションが付与された文字列学習用データの例を示す図である。

[図25]電子文書生成プログラムのフローチャートである。

[図26]電子文書生成プログラムに係る一実施形態のフローチャート（1／3）である。

[図27]電子文書生成プログラムに係る一実施形態のフローチャート（2／3）である。

[図28]電子文書生成プログラムに係る一実施形態のフローチャート（3／3）である。

発明を実施するための形態

[0024] 図1乃至図24を参照して本開示に係る電子文書生成装置10の一実施形態について説明する。本実施形態では、電子文書生成装置10をインターネット及びLAN（Local Area Network）などの情報通信ネットワーク11に接続して使用する一例を示す。図1を参照して、電子文書生成装置10を含む電子文書生成システム100の概略を説明する。図1

は、電子文書生成装置 10 を含む電子文書生成システム 100 の概略を示す図である。

[0025] 電子文書生成システム 100 は、電子文書生成装置 10、ユーザ端末 12、文字列学習モデル 13、レイアウト学習モデル 14、及び文書画像データベース 15などを備える。電子文書生成装置 10、ユーザ端末 12、文字列学習モデル 13、レイアウト学習モデル 14、及び文書画像データベース 15は、情報通信ネットワーク 11に接続され、おのおの相互に情報通信が可能である。

[0026] 電子文書生成システム 100 は、電子文書生成装置 10 を用いて文書画像に含まれた文字列を文字認識しテキストデータを生成するものである。電子文書生成装置 10 は、文書画像のレイアウトについてレイアウト学習モデルを用いて画像認識し、文書画像に含まれた文字列について文字列学習モデルを用いて文字認識するものである。

[0027] 電子文書生成装置 10 とは、例えば、パソコンなどに代表されるコンピュータの一種であり情報処理装置である。電子文書生成装置 10 は、さらに様々なコンピュータに含まれる演算処理装置及びマイコン等も含み、アプリケーションによって本開示に係る機能を実現することが可能な機器、及び装置などをも含むものとする。

[0028] 文字列学習モデル 13 は、文書画像に含まれる文字列の画像認識を行う学習モデルであり、電子文書生成装置 10 の文字認識に用いられる。文字列学習モデル 13 は、情報通信ネットワーク 11 を介して電子文書生成装置 10 に利用可能であればその保存場所は任意であり、例えばパソコン、サーバー装置、データベースなどの情報処理装置に保存される。本実施形態の説明の便宜上、文字列学習モデル 13 は、文字列学習モデル 13 が保存される情報処理装置を表すものとする。

[0029] 文字列学習モデル 13 は、既存の学習モデルにより構成されてもよいし、若しくは電子文書生成装置 10 の利用に適した学習モデルとして独自に構成されてもよい。文字列学習モデル 13 は、日本語、英語、中国語などの各種

言語に適した学習モデルをそれぞれ備えるものとし、図1では第1文字列学習モデル、第2文字列学習モデル、第3文字列学習モデルなどと記載するものとする。

[0030] なお、文字列学習モデル13は、情報通信ネットワーク11に接続されるものに限らず、電子文書生成装置10に含まれ当該装置10の直接的な制御のもとに利用されてもよい。また、文字列学習モデル13は、情報通信ネットワーク11に接続された複数の情報処理装置に分散して保存されてもよい。

[0031] レイアウト学習モデル14は、文書画像に含まれる複数の要素と当該複数の要素の各々の識別情報との対応関係を後述のレイアウト学習用データに基づいて学習し、文書画像のレイアウトの認識を行う学習モデルであり、電子文書生成装置10のレイアウト認識に用いられる。レイアウト学習モデル14は、文字列学習モデル13と同様に、情報通信ネットワーク11を介して電子文書生成装置10に利用可能であればその保存場所は任意であり、情報通信ネットワーク11に接続された情報処理装置に保存される。本実施形態の説明の便宜上、レイアウト学習モデル14は、レイアウト学習モデル14が保存される情報処理装置を表すものとする。

[0032] レイアウト学習モデル14は、契約書用のレイアウト学習モデル、請求書用のレイアウト学習モデル、覚書用のレイアウト学習モデル、納品書用のレイアウト学習モデル、及び領収書用のレイアウト学習モデルなどを含む。

[0033] 契約書用のレイアウト学習モデルは、契約書の文書画像のレイアウト認識を行う学習モデルであり、契約書用のレイアウト学習用データを用いて学習する。契約書用のレイアウト学習モデルは、契約書のどの位置に、どのような情報があるかを学習し、特に、箇条書きで記載され、表が無い場合が多く、手書きの署名欄が有るなど契約書に特有のレイアウトについて学習する。契約書用のレイアウト学習用データは、後述するアノテーションが付与された、例えば、200種類の契約書のフォームであって1フォームにつき少なくとも3、4枚の契約書の文書画像に基づいて生成される。

- [0034] 請求書用のレイアウト学習モデルは、請求書の文書画像のレイアウト認識を行う学習モデルであり、請求書用のレイアウト学習用データを用いて学習する。請求書用のレイアウト学習モデルは、請求書のどの位置に、どのような情報があるかを学習し、特に、表の占める範囲が大きい場合が多く、日本語で書かれていても英数字の字句も少なくないなど請求書に特有のレイアウトについて学習する。請求書用のレイアウト学習用データは、後述するアノテーションが付与された、例えば、200種類の請求書のフォームであって1フォームにつき少なくとも3、4枚の請求書の文書画像に基づいて生成される。
- [0035] 覚書用のレイアウト学習モデルは、覚書の文書画像のレイアウト認識を行う学習モデルであり、覚書のレイアウト学習用データを用いて学習する。覚書用のレイアウト学習モデルは、覚書のどの位置に、どのような情報があるかを学習し、特に、表が無い場合が多く、手書きの署名欄が有るなど覚書に特有のレイアウトについて学習する。覚書用のレイアウト学習用データは、後述するアノテーションが付与された、例えば、200種類の覚書のフォームであって1フォームにつき少なくとも3、4枚の覚書の文書画像に基づいて生成される。
- [0036] 納品書用のレイアウト学習モデルは、納品書の文書画像のレイアウト認識を行う学習モデルであり、納品書用のレイアウト学習用データを用いて学習する。納品書用のレイアウト学習モデルは、納品書のどの位置に、どのような情報があるかを学習し、特に、表の占める範囲が大きい場合が多く、商品の名称及び品番などの記載が少なくないなど納品書に特有のレイアウトについて学習する。納品書用のレイアウト学習用データは、後述するアノテーションが付与された、例えば、200種類の納品書のフォームであって1フォームにつき少なくとも3、4枚の納品書の文書画像に基づいて生成される。
- [0037] 領収書用のレイアウト学習モデルは、領収書の文書画像のレイアウト認識を行う学習モデルであり、領収書用のレイアウト学習用データを用いて学習する。領収書用のレイアウト学習モデルは、領収書のどの位置に、どのよう

な情報があるかを学習し、特に、金額の手書き欄、若しくは金額が記載された表が記載されている場合が多いなど領収書に特有のレイアウトについて学習する。領収書用のレイアウト学習用データは、後述するアノテーションが付与された、例えば、200種類の領収書のフォームであって1フォームにつき少なくとも3、4枚の領収書の文書画像に基づいて生成される。

[0038] なお、レイアウト学習モデル14は、情報通信ネットワーク11を介する電子文書生成装置10の利用に限定されるものではなく、電子文書生成装置10に含まれるものとしてもよい。また、レイアウト学習モデル14は、情報通信ネットワーク11に接続された複数の情報処理装置に分散して保存されてもよい。

[0039] 文書画像データベース15は、文書の画像を蓄積したデータベースである。電子文書生成装置10は、文書画像データベース15に記憶された文書画像を取得し、文字列学習モデルの学習に用いる文字列学習用データ、及びレイアウト学習モデルの学習に用いるレイアウト学習用データを生成する。

[0040] ユーザ端末12は、電子文書生成装置10の操作に用いられる。電子文書生成装置10によって生成された電子文書の中に誤って認識された文字があった場合、若しくは電子文書生成装置10によって生成された電子文書のレイアウトが誤って認識されたものであった場合に、ユーザ端末12のユーザからの修正入力に従って当該電子文書は修正され、電子文書生成装置10は当該修正を受け付けて、文字列学習モデル13及びレイアウト学習モデル14の少なくともいずれかの再学習を行う。

[0041] 次に図2を参照して、電子文書生成装置10の機械的構成について説明する。図2は、電子文書生成装置10の機械的構成を示すブロック図である。電子文書生成装置10は、入出力インターフェース20、通信インターフェース21、Read Only Memory (ROM) 22、Random Access Memory (RAM) 23、記憶部24、Central Processing Unit (CPU) 25、Graphics processing units (GPU) 28等を備えている。

[0042] 入出力インターフェース 20 は、電子文書生成装置 10 の外部装置に対してデータなどの送受信を行う。外部装置とは、電子文書生成装置 10 に対してデータなどの入出力を行う入力装置 26 及び出力装置 27 のことである。入力装置 26 とはキーボード、マウス、及びスキャナーなどのことであり、出力装置 27 とはモニター、プリンタ及びスピーカなどのことである。

[0043] 通信インターフェース 21 は、情報通信ネットワーク 11 を介して外部との通信を行う際に電子文書生成装置 10 のデータなどの入出力を行う機能を備える。

記憶部 24 は、記憶装置として利用でき、電子文書生成装置 10 が動作する上で必要となる各種アプリケーション及び当該アプリケーションによって利用される各種データなどが記録される。GPU 28 は、機械学習などを実行する上で行われる繰り返し演算を多用する場合に適しており、CPU 25 とともに用いる。

[0044] 電子文書生成装置 10 は、後述する電子文書生成プログラムを ROM 22 若しくは記憶部 24 に保存し、RAM 23 など構成されるメインメモリに当該電子文書生成プログラムを取り込む。CPU 25 は、電子文書生成プログラムを取り込んだメインメモリにアクセスして、電子文書生成プログラムを実行する。

[0045] 次に図 3 を参照して、電子文書生成装置 10 が行う処理の概略を説明する。図 3 は、電子文書生成装置 10 が行う処理の概要を示す図である。

電子文書生成装置 10 は、次に述べる処理 I ～ III をこの順に行う。

[0046] 処理 I では、文書画像の「背景除去」、「傾き補正」、及び「形状調整」を含む前処理 55 を行う。

前処理 55 とは、文字列を含む画像に対して、学習モデルを用いた文字認識を実行（認識）しやすくするための事前の処理を行うことをいい、処理 I、II、III で行う認識処理の認識精度を向上させることを目的とする。

[0047] 処理 II では、レイアウト認識処理 56 を行う。

レイアウト認識処理 56 では、先ず文書画像の「レイアウト認識」を行う

。レイアウト認識処理56とは、入力された画像内で、どの位置に、どのような情報があるのかを認識する処理である。

[0048] ここでいう情報とは、文字列、表、画像、印章、手書きなどのことをいう。電子文書生成装置10は、文書画像のレイアウトを認識して文書画像に表が含まれている場合は、「表の認識」を行い、表に含まれるセルについて「セルの画像の切り出し」を行う。

[0049] 処理111では、文字列認識処理57を行う。

文字列認識処理57とは、文字列を含む画像を、画像と画像に含まれる文字列との対応関係を学習した文字列学習モデル13を用いて、テキストデータに変換する処理のことである。文字列認識処理57は、「テキストデータの配置」及び「ノイズ除去」などの処理を含むものとしてもよい。

[0050] 文字列認識処理57では、文字列の画像をテキストデータに変換するとともに、「テキストデータの配置」及び「ノイズ除去」を行う。

「テキストデータの配置」とは、切出した文字列の画像にスペースが含まれている場合には文字列とともにスペースも一緒に認識されるので、テキストデータはスペースとともに配置されることを指す。

[0051] 「ノイズ除去」とは、切出した文字列の画像にノイズが含まれている場合に、ノイズについては電子文書生成装置10によって認識されないのでテキストデータから受動的に除去されることをいう。ここでいうノイズとは、切り出された文字列の画像に含まれる、文字を構成しない画素のことをいう。

[0052] 次に図4を参照して、電子文書生成装置10の機能的構成について説明する。図4は、電子文書生成装置10の機能的構成を示すブロック図である。電子文書生成装置10は、後述する電子文書生成プログラムを実行することで、CPU25に、文書画像取得部31、前処理部32、背景除去部32a、傾き補正部32b、形状調整部32c、レイアウト認識部33、切出部34、文字列認識部35、出力部36、レイアウト学習用データ生成部40、レイアウト学習用データ修正部41、レイアウト学習部42、文字列学習用データ生成部43、文字列学習用データ修正部44、文字列学習部45等を

備える。

- [0053] 文書画像取得部 31 (図 4 参照) は、文書を画像化した文書画像を取得する。

文書画像取得部 31 は、文書画像データベース 15 から文書画像を取得してもよい。或いは、文書画像取得部 31 は、入力装置 26 のスキャナーから文書画像を入手してもよい。

- [0054] 図 5 を参照して、文書画像取得部 31 に取得された文書画像、及び電子文書生成装置 10 から出力される電子文書について説明する。

図 5 は電子文書生成装置 10 の入力データと出力データとを説明する図であり、図 5 (a) は入力データとして文書画像取得部 31 に取得された文書画像を示す。当該文書画像には、ホチキス跡 50、手書き 51、印章 52、及び画像 53 などのノイズが存在する。

- [0055] これらのノイズは、人及びパソコンなどの情報処理装置が当該文書の内容を理解する上で邪魔若しくは不要となる。ノイズの他の例としては、ファイリングのために開けた穴、紙面に残った折り目などがある。折り目は線として認識される恐れがあり、電子文書に反映されないように除去の必要性が高いものである。

- [0056] 電子文書生成装置 10 は、取得した文書画像のレイアウトを維持しつつ、当該文書画像内の文字列をテキストデータに変換して電子文書を出力する (図 5 (b) 参照)。電子文書生成装置 10 は、ノイズとして認識したホチキス跡 50、手書き 51、印章 52、及び画像 53 について能動的処理によりノイズ除去 54 し、その他、文字列及びノイズとして認識されなかった文書画像内の画素については、電子文書に残留させない受動的処理により除去する。

- [0057] 図 5 (b) の文書画像内の表は、当該文書画像内の配置を維持しつつ、電子文書内のオブジェクトデータとしてテキストデータとともに出力される。電子文書生成装置 10 は、出力する電子文書に含める要素を任意に選択できるものである。

[0058] 例えば、ホチキス跡 5 0、手書き 5 1、印章 5 2、及び画像 5 3 などは通常使用では除去されるが、印章 5 2、及び画像 5 3 をイメージデータとして電子文書に含めて出力することもできる。

[0059] 前処理部 3 2（図 4 参照）は、文書画像取得部 3 1 が取得した文書画像について前処理 5 5 を行う。

前処理 5 5 は、後述するレイアウト認識部 3 3 及び文字列認識部 3 5 による、学習モデルを用いる画像認識の認識精度を向上させるために行われる。

[0060] 前処理部 3 2 は、背景除去部 3 2 a、傾き補正部 3 2 b、及び形状調整部 3 2 c を備える。

背景除去部 3 2 a（図 4 参照）は、文書画像取得部 3 1 が取得した文書画像の背景を除去する。

[0061] 図 6 を参照して、背景除去部 3 2 a により行われる処理について説明する。図 6 は、前処理 5 5 で行う背景除去を説明する図である。図 6（a）は背景除去される前の文書画像 5 8 a を示し、図 6（b）は背景除去された後の文書画像 5 8 b を示す。

[0062] 背景除去部 3 2 a は、文書画像の背景色を白色にすることによって文書画像の背景を除去する。具体的には、背景除去部 3 2 a は、取得した文書画像の背景色を検出し、当該背景色が白色か否かを判断する。背景色が白色ではないと判断された場合、背景除去部 3 2 a は、文書画像の背景以外の情報を抽出し、背景色を白色にした後に抽出した情報を重ね合わせる。

[0063] 背景除去部 3 2 a によれば、背景を削除することでレイアウト認識部 3 3 及び文字列認識部 3 5 による画像認識の誤動作の原因となるノイズを除去することができ、認識精度を向上させることができる。

[0064] 傾き補正部 3 2 b（図 4 参照）は、文書画像取得部 3 1 が取得した文書画像の傾きを補正する。

図 7 を参照して、傾き補正部 3 2 b により行われる処理について説明する。図 7 は、前処理 5 5 で行う傾き補正を説明する図である。図 7（a）は傾き補正される前の文書画像 5 9 a を示し、図 7（b）は傾き補正された後の

文書画像 5 9 b を示す。

[0065] 傾き補正部 3 2 b は、文書画像の中に傾いた文字列がある場合に当該文字列の傾きを補正し、当該文字列を書字方向に対して平行若しくは垂直にする。傾き補正部 3 2 b は、文書画像が縦書きの場合は、傾いた文字列を縦書きの方向に対して平行になるように補正し、文書画像が横書きの場合は、傾いた文字列を横書きの方向に対して平行になるように補正する。

[0066] 具体的には、傾き補正部 3 2 b は、文書画像の文字列を抽出し、抽出した文字列の中に傾斜した文字列があるか否かを判断する。抽出した文字列の中に傾斜した文字列があると判断した場合に、傾き補正部 3 2 b は、傾斜した当該文字列の書字方向に対する傾斜角を検出し、傾斜した当該文字列に対して傾斜角がゼロになるように回転処理を施す。

[0067] 傾き補正部 3 2 b によれば、文字列の傾きを補正することで、文字列認識部 3 5 による画像認識の認識精度を向上させることができる。さらに、レイアウト認識部 3 3 によるレイアウトの認識エラーを低減することができる。

[0068] 形状調整部 3 2 c (図 4 参照) は、文書画像取得部 3 1 が取得した文書画像の全体の形状及び大きさを調整する。

図 8 を参照して、形状調整部 3 2 c により行われる処理について説明する。図 8 は、前処理で行う形状調整を説明する図である。図 8 (a) は形状調整される前の文書画像 6 0 a を示し、図 8 (b) は形状調整された後の文書画像 6 0 b を示す。

[0069] 形状調整部 3 2 c は、文書画像取得部 3 1 が取得した文書画像の全体の形状が実際の文書と比べて異なる場合、当該文書画像の全体の形状を実際の文書の全体の形状に基づいて調整を行う。具体的には、文書画像取得部 3 1 が取得した文書画像の全体の縦横比が実際の文書の全体の縦横比と異なる場合は、当該文書画像の全体の縦横比が実際の文書の全体の縦横比と等しくなるように形状調整部 3 2 c が調整する。

[0070] また、文書画像取得部 3 1 が取得した文書画像の大きさが大きすぎる場合、若しくは小さすぎる場合に、その後の処理が正常に行われない可能性がある

るので、形状調整部 3 2 c は、その後の処理が正常に行われるように文書画像取得部 3 1 が取得した文書画像の大きさを調整する。

[0071] 形状調整部 3 2 c によれば、文書画像取得部 3 1 が取得した文書画像の形状及び大きさを調整することで、その後に行われるレイアウト認識部 3 3 による実際の文書に則したレイアウトの認識精度を向上させることができ、さらに文字列認識部 3 5 による画像認識の認識精度を向上させることができる。

[0072] レイアウト認識部 3 3 (図 4 参照) は、文書画像 6 1 に含まれる複数の要素と、当該複数の要素の各々の識別情報との対応関係を学習したレイアウト学習モデル 1 4 を用いて、文書画像取得部 3 1 に取得された文書画像 6 1 に含まれる複数の要素の各々の文書画像 6 1 内における範囲を特定し、複数の要素の各々の種類を認識し、複数の要素の各々の範囲に係る文書画像 6 1 内における位置情報を取得する。

[0073] 要素の種類は、文字列 4 8、表 4 9、画像 5 3、印章 5 2、又は手書き 5 1 のいずれかであることとしてもよい。なお、要素の種類はこれに限らず、ホチキス跡 5 0、パンチ穴跡、破損(破れ)跡、複写用カーボン汚れなどを用いてもよい。

[0074] 要素の種類は、文書の種類(例えば、契約書、請求書、覚書、納品書、又は領収書など)に適したものをを用いてもよい。例えば、領収書の裏面に複写用のカーボンが添布してあり、表面にカーボンが移り汚れとなる場合は、要素の種類に複写用カーボンによる汚れを用いて能動的に当該複写用カーボンによる汚れを除去してもよい。

[0075] レイアウト学習モデル 1 4 は、契約書用のレイアウト学習モデル、請求書用のレイアウト学習モデル、覚書用のレイアウト学習モデル、納品書用のレイアウト学習モデル、又は領収書用のレイアウト学習モデルのいずれかであることとしてもよい。

[0076] 要素の種類は、文書の種類に応じて必要なものと不要なものに分類してもよい。

この場合、レイアウト認識部33は、文書画像61に含まれる複数の要素のうち、認識した要素が不要なものに該当する場合は当該要素の位置情報は取得されず、認識した要素が必要なものに該当する場合は当該要素の位置情報を取得することとしてもよい。または、レイアウト認識部33は、文書画像61に含まれる複数の要素のうち、必要な要素のみを認識し、当該要素の位置情報を取得することとしてもよい。

[0077] レイアウト認識部33は、要素の各々の種類を認識して、当該要素の各々の範囲に係る文書画像の位置情報を取得した後、要素同士が重なり合う、若しくは要素同士が離れすぎている場合には、実際の文書に基づいて、当該要素の各々の範囲及び取得した位置情報を補正する。

[0078] 図9を参照して、レイアウト認識部33が認識した認識範囲に欠落が生じた場合にレイアウト認識部33が行う欠落解消の補正処理の一例について説明する。欠落とは、レイアウト認識部33が要素として認識すべき範囲についてその一部が認識されず、要素の範囲の一部が不足することをいう。図9はレイアウト認識処理で行う欠落解消の補正処理を説明する図であり、図9(a)は補正前の様子を示し、図9(b)は補正後の様子を示す。

[0079] レイアウト認識部33は、文書画像取得部31によって取得された文書画像に含まれる文字列の画像70を文字列として認識する際に、その認識範囲に欠落が有るか否かの判定を行い、欠落が有る場合には欠落部分を追加する補正処理を行う。

[0080] 図9(a)は、レイアウト認識部33が、文字列の画像70について認識範囲72aの文字列として認識した様子を示している。認識範囲72aは、文字列の画像70の左端部分に欠落を有している。レイアウト認識部33は、認識範囲72aの周囲の所定範囲以内に黒線が有るか否かの判定を行い、黒線が有る場合には黒線を含む範囲72bを認識範囲72aに追加する補正を行う(図9(b)参照)。

[0081] なお、レイアウト認識部33が行う有無の判定は黒線に限定されるものではなく、文字と同色の線又は予め設定された色の線が認識範囲72aの周囲

の所定範囲以内に有るか否かの判定を行うこととしてもよい。なぜなら、レイアウト認識処理で行う欠落解消の補正処理は、その後に行われる文字認識処理の認識精度を向上させるのが主な目的だからである。

[0082] 当該補正処理によれば、レイアウト認識部33が認識した要素の範囲に欠落が生じ場合でも、その欠落した範囲を追加することで正常な認識範囲へと補正することができ、当該要素に含まれる文字列について文字列認識部35は正常に文字認識を行うことができる。

[0083] 図10を参照して、レイアウト認識部33が認識した認識範囲が、他の要素に重なってしまった場合にレイアウト認識部33が行う補正の一例について説明する。図10はレイアウト認識処理で行う重なり解消の補正処理を説明する図であり、図10(a)は補正前の様子を示し、図10(b)は補正後の様子を示す。

[0084] レイアウト認識部33は、文書画像取得部31によって取得された文書画像に含まれる文字列の画像73を文字列として認識する際に、その認識範囲75aが他の要素（例えば、表74）に重なっているか否かの判定を行い、重なりが生じている場合には重なりを解消する補正処理を行う。

[0085] 図10(a)は、レイアウト認識部33が、文字列の画像73について認識範囲75aの文字列として認識した様子を示している。認識範囲75aは、文字列の画像73の右隣の表74に空白（スペース）を超えて重なっている。レイアウト認識部33は、認識範囲75aの内部に所定の大きさの空白（スペース）が有るか否かの判定を行い、当該空白（スペース）が有る場合には、当該空白（スペース）及び当該空白（スペース）より右側の部分に係る認識範囲75aを削除して認識範囲75bとする補正を行う（図10(b)参照）。

[0086] 要素と他の要素との間には必ず所定の大きさの空白（スペース）があるので、レイアウト認識部33は、認識範囲の内部に所定の大きさの空白（スペース）が有る場合に当該認識範囲は他の要素と重なっていると断定するものである。レイアウト認識処理で行う重なり解消の補正処理によれば、レイア

ウト認識部33はレイアウトの認識精度を向上させることができる。

[0087] 図11を参照して、レイアウト認識部33により行われる処理について説明する。図11は、レイアウト認識処理56で行うレイアウト認識を説明する図であり、図11(a)はレイアウト認識される前の文書画像61の状態を示し、図11(b)はレイアウト認識された後の文書画像62の状態を示す。

[0088] レイアウト認識部33は、文書画像61に含まれる要素(文字列48、表49、印章52、画像53)の文書画像61内における範囲について、レイアウト学習モデル14を用いた画像認識により特定する。

[0089] 図11(b)において、説明の便宜上、特定された文字列48の範囲を実線で囲い、特定された表49、印章52、及び画像53の範囲を破線で囲う。要素の境界は電子文書生成装置10が認識できれば良いので、人に対して可視化されていなくても良い。

[0090] レイアウト認識部33は、特定された文書画像61内における範囲において、レイアウト学習モデル14を用いた画像認識により該当する要素の種類を認識し、当該要素の種類とともに当該範囲の文書画像62内に係る位置情報を取得する。位置情報は、文書画像62内の所定点を原点とした平面直交座標によって表されてもよい。

[0091] レイアウト学習モデル14は、文書画像61の種類に合わせて予め設定されており、レイアウト認識部33は、予め設定されたレイアウト学習モデル14を用いて文書画像61のレイアウトを認識する。

[0092] すなわち、文書画像取得部31により取得された文書画像61が契約書だった場合は契約書用のレイアウト学習モデル14を用いて画像認識を行い、請求書だった場合は請求書用のレイアウト学習モデル14を用いて画像認識を行い、覚書だった場合は覚書用のレイアウト学習モデル14を用いて画像認識を行い、納品書だった場合は納品書用のレイアウト学習モデル14を用いて画像認識を行い、領収書だった場合は領収書用のレイアウト学習モデル14を用いて画像認識を行う。

- [0093] レイアウト認識部33は、文書画像取得部31により取得された文書画像61の種類に合わせてレイアウト学習モデル14を使い分けるので、文書画像61のレイアウト認識の認識精度を向上させることができる。
- [0094] 切出部34（図4参照）は、レイアウト認識部33により認識された種類が表に該当する要素において、当該要素に含まれる表の中のセルの各々を切り出し、セルの各々の文書画像内における位置情報を取得する。
- [0095] 図12を参照して、レイアウト認識部33による表49の認識について説明する。図12はレイアウト認識処理56で行う表の認識を説明する図であり、図12（a）はレイアウト認識部33に認識される前の表63を示し、図12（b）はレイアウト認識部33に認識された後の表64を示す。図12（b）では、説明の便宜上、縦線65として認識された線を一点鎖線として表し、横線66として認識された線を破線として表すことにする。
- [0096] レイアウト認識部33は、表64を構成する全ての縦線65及び横線66の各々の長さや位置を認識する。レイアウト認識部33は、表64を構成する全ての縦線65及び横線66の長さや位置を認識することで、表64に含まれる全てのセルについて認識する。すなわち、レイアウト認識部33は、隣接する2本の縦線65及び隣接する2本の横線66により構成される四角形をセルとして認識する。
- [0097] さらに、レイアウト認識部33は表64を構成する線の線種についても認識する。認識された線種は、取得した文書画像に基づいて電子文書を再現する際に、当該電子文書に含まれる表を構成する線のオブジェクトに反映される。従って、例えば、文書画像62内の表の線が破線であった場合、文書画像62に基づいて再現された電子文書に含まれる表の線は破線のオブジェクトとして表現される。
- [0098] 切出部34は、レイアウト認識部33により把握された表64に含まれる全てのセルについてセル単体毎の画像に切り出す。
- 図13を参照して、切出部34によるセル画素の切り出しについて説明する。図13は、セル画像の切り出しを説明する図である。切出部34により

切り出されたセル 6 7 は、複数の文字列を含む場合もある。

- [0099] 切出部 3 4 は、表 6 4 に含まれる全てのセルについてセル単体毎の画像と当該セルの表 6 4 における位置情報について取得する。位置情報は、表 6 4 内の所定点を原点とした平面直交座標によって表されてもよいし、若しくは表 6 4 における（行、列）によって表されてもよい。
- [0100] 切出部 3 4 は、レイアウト認識部 3 3 により認識された表を構成する全ての縦線及び横線を再生し、全てのセルの位置情報を生成する。
- [0101] 図 1 4 を参照して、複数の文字列を含むセル 6 7 について説明する。図 1 4 は、セル画像内の文字列を説明する図である。
- [0102] 切出部 3 4 は、切り出されたセル 6 7 に複数行の文字列が含まれている場合は、さらに全ての文字列について文字列単体毎の画像を切り出す。図 1 4 に示すセル 6 7 は、2 行の文字列を含んでおり、切出部 3 4 は文字列の画像 6 7 a 及び文字列の画像 6 7 b を切出す。
- [0103] 文字列認識部 3 5（図 4 参照）は、文書画像と当該文書画像に含まれる文字列との対応関係を学習した文字列学習モデル 1 3 を用いて、文書画像取得部 3 1 に取得された文書画像に含まれる文字列を文字認識し、当該文字列に係るテキストデータを生成する。
- [0104] 文字列認識部 3 5 は、レイアウト認識部 3 3 により認識された範囲に含まれる文字列について、文字列学習モデル 1 3 を用いて文字認識し、文字列に係るテキストデータを生成してもよい。
- [0105] 文字列認識部 3 5 は、切出部 3 4 に切り出されたセルの各々に含まれる文字列について、文字列学習モデル 1 3 を用いて文字認識を行い、文字列に係るテキストデータを生成してもよい。
- [0106] 文字列認識部 3 5 は、複数の文字列学習モデル 1 3 を備え、複数の要素の各々に含まれる文字列の言語に適した文字列学習モデル 1 3 を用いてもよい。

文字列認識部 3 5 は、英語で書かれた文書画像を文字認識する場合に、英語の文字列の認識に適した文字列学習モデルを用いることで、認識精度を向

上させることができる。

[0107] 図 1 5、1 6 を用いて文字列認識部 3 5 が行う文字認識について説明する。

図 1 5 は、文字列認識処理 5 7 で行うテキストデータの配置を説明する図であり、図 1 5 (a) は文字認識が行われる前の文字列の画像 6 7 a であり、図 1 5 (b) は文字認識が行われた後の文字列 6 8 a、すなわちテキストデータ 6 8 a である。

[0108] 図 1 6 は、文字列認識処理 5 7 で行うノイズ除去を説明する図であり、図 1 6 (a) は文字認識が行われる前の文字列の画像 7 1 a であり、図 1 6 (b) は文字認識が行われた後の文字列 7 1 b、すなわちテキストデータ 7 1 b である。

[0109] 図 1 5 (a) に示す文字列の画像 6 7 a は、1 行の文字列の他に手書きチェック跡が残る。当該文字列は、単語と単語との間に空白を含んでいる。文字列認識部 3 5 は、文字列の画像 6 7 a 全体について、文字列学習モデル 1 3 を用いて文字認識し、テキストデータを生成する。

[0110] 文字列認識部 3 5 は、文字列の画像 6 7 a の中の、2 個の字句である「L / C NO :」、「I L C 1 8 H 0 0 0 2 1 9」、及び 2 個の字句の間の空白について文字認識して、2 個の字句に対応するテキストデータ、及び 2 個の字句の間の空白に対応するテキストデータを生成する (6 8 a : 図 1 5 (b) 参照)。

従って、文字列認識部 3 5 は、字句と字句の間にあるスペースについても認識してテキストデータに変換するので、画像 6 7 a と同様に 2 つの字句を離して配置することができる。

[0111] 文字列認識部 3 5 は、文字列の画像 6 7 a について文字認識する際に、手書きチェック跡については文字認識されずテキストデータに含まれないので、出力される電子文書から手書きチェック跡は削除される (6 8 a : 図 1 5 (b))。従って、文字列認識部 3 5 の文字認識の対象とならない手書チェック跡などのノイズは受動的に電子文書から除去される。

- [0112] 図16(a)に示す文字列の画像71aは、1行の文字列に重なった印章の一部がノイズとして残っている。文字列認識部35は、文字列の画像71a全体について、文字列学習モデル13を用いて文字認識し、テキストデータを生成する。
- [0113] 文字列認識部35は、文字列の画像71a全体について文字認識し、文字列「authorized to act on behalf of the」について、当該文字列に対応するテキストデータについて生成する(71b:図16(b)参照)。
- [0114] 文字列の画像71aに含まれるノイズは、文字列認識部35の文字認識の対象とはならないので、受動的に電子文書から除去される(71b:図16(b)参照)。
- [0115] 図15、16を用いて、文書画像と、その文章画像に対する文字認識を施した後の文字列のテキストデータについて説明したが、文字列学習モデル13は、図15(a)と図15(b)とを対応付けたデータや、図16(a)と図16(b)とを対応付けたデータを、教師データとして、多数学習することで、このような深層学習を用いた画像からの文字認識を実現することができる。
- [0116] 文字列認識部35は、文字列学習モデル13を用いて、画像67a、71aに含まれる文字列について文字認識する際に当該文字列に含まれる文字の大きさ及び書体などの属性データを取得してもよい。この文字の属性データは、後述の出力部36によって出力されるテキストデータの属性データとして反映される。
- [0117] 出力部36(図4参照)は、テキストデータを電子媒体のテキストとして出力する。
- 出力部36は、複数の要素に係る範囲の各々の位置情報に、複数の要素に係るテキストデータの各々を電子媒体のテキストとして出力してもよい。
- [0118] 電子媒体とは、電子的に記録媒体に保存されたデータに限らず、記録媒体に保存された状態でなくパソコンなどの情報処理装置がその内容を扱うこと

が出来るデータそのものも含むものとする。

要素の位置情報は、文書画像 6 2 内の所定点を原点とした平面直交座標によって表されてもよい。

[0119] 出力部 3 6 によれば、複数の要素に係るテキストデータを当該要素に係る位置情報に基づいて出力するので、取得した文書画像 6 1 のレイアウトを維持しつつノイズを除去して、当該文書画像 6 1 内の文字列をテキストデータに変換して電子文書を出力することができる。

[0120] 出力部 3 6 は、文字列認識部 3 5 によって取得された文字の属性データについて、テキストデータに反映させて電子文書に出力してもよい。当該出力部 3 6 によれば、電子文書生成装置 1 0 は、文書画像 6 1 に含まれる文字の大きさ及び書体などの属性データについて、出力する電子文書に含まれるテキストデータの属性データとして再現することができる。

[0121] レイアウト学習用データ生成部 4 0（図 4 参照）は、複数の要素を含む文書画像であって、当該要素に当該要素の各々に該当する種類に関連付けられたアノテーションが付与されており、アノテーションが付与された複数の文書画像を蓄積してレイアウト学習用データを生成する。

[0122] レイアウト学習用データはレイアウト学習モデル 1 4 の教師有り学習に用いられる。

レイアウト学習用データに蓄積される文書画像に、アノテーションとともに文書画像に含まれる複数の要素に係る範囲の各々の文書画像内における位置情報が付与されてもよい。

[0123] 図 1 7 から図 2 3 を参照して、アノテーションが付与されたレイアウト学習用データについて説明する。図 1 7 から図 2 3 は、アノテーションが付与されたレイアウト学習用データの例を示す図である。

[0124] レイアウト学習用データ生成部 4 0 は、文書画像データベース 1 5 から文書画像を取得し、当該文書画像にアノテーションを付与してレイアウト学習用データを生成する。なお、レイアウト学習用データを生成する際に、レイアウト学習用データ生成部 4 0 を用いることなく、ユーザが手動でレイアウト

ト学習用データを生成することもできる。ユーザが手動でレイアウト学習用データを生成する場合は、文書画像データベース15から取得した文書画像にユーザ端末12を用いてアノテーションを付与することができる。

[0125] 図17、18を参照して、請求書用のレイアウト学習モデル14の学習に用いられるレイアウト学習用データについて説明する。文書画像に含まれる要素として、文字列、表、画像、印章、外枠、ノイズについて、電子文書生成装置10が識別し分類できるように、それぞれの要素に注釈記号を付与する。

[0126] 文字列に係る要素には、文字列の注釈記号76を付与し、当該文字列を矩形の枠線で囲い、当該枠線に「Text」のタグを目印として付ける。矩形の枠線で囲われた部分は、文書画像内における当該文字列に係る要素の占める範囲としてレイアウト学習モデル14に学習される。

[0127] 表に係る要素には、表の注釈記号77を付与し、当該表の外枠に矩形の枠線を重ね合わせ、当該枠線に「Border Table」のタグを目印として付ける。矩形の枠線で囲われた部分は、文書画像内における当該表に係る要素の占める範囲としてレイアウト学習モデル14に学習される。

[0128] 画像に係る要素には、画像の注釈記号78を付与し、当該画像の境界線に注釈記号を示す枠線を重ね合わせ、当該枠線に「Image」のタグを目印として付ける。画像には、ロゴ、マーク、写真、及びイラストなどが含まれるものとする。枠線で囲われた部分は、文書画像内における当該画像に係る要素の占める範囲としてレイアウト学習モデル14に学習される。

[0129] 印章に係る要素には、印章の注釈記号79を付与し、当該印章の境界線に注釈記号を示す枠線を重ね合わせ、当該枠線に「Hun」のタグを目印として付ける。枠線で覆われた部分は、文書画像内における当該印章に係る要素の占める範囲としてレイアウト学習モデル14に学習される。

[0130] 外枠に係る要素には、外枠の注釈記号80を付与し、当該外枠の境界線に枠線を重ね合わせ、当該枠線に「Border」のタグを目印として付ける。枠線を構成する4本の線分について、その長さや位置について、レイアウト

ト学習モデル 14 に学習される。

[0131] ノイズに係る要素には、ノイズの注釈記号 81 を付与し、当該ノイズを矩形の枠で囲い、当該枠に「Noise」のタグを目印として付ける。枠で覆われた部分は、文書画像内における当該ノイズに係る要素の占める範囲としてレイアウト学習モデル 14 に学習される。

[0132] 図 19 を参照して、表の認識の学習に用いられるレイアウト学習データについて説明する。文書画像データベース 15 から取得した文書画像に含まれる表について、表を構成する全ての縦線に縦線の注釈記号 83 である一点鎖線を重ね合わせ、表を構成する全ての横線に横線の注釈記号 84 である破線を重ね合わせる。

[0133] レイアウト学習モデル 14 は、全ての一点鎖線及び破線を認識することで、表の大きさ、表が占める範囲、位置、及び表に含まれる全セルの情報について学習することができる。セルの情報とは、表に含まれるセルの数、及びセル各々の表における位置のことであり、表における位置は表の（行、列）で表される。

[0134] 図 20、21 を参照して、表のセルの中の文字列の認識の学習に用いられるレイアウト学習データについて説明する。図 20 は、セルの各々に 1 行ずつ文字列が含まれているレイアウト学習データである。図 21 は、1 行の文字列を含むセル、2 行の文字列を含むセル、3 行の文字列を含むセルに係る表を認識するレイアウト学習データである。

[0135] 図 20、図 21 が示す通り、1 つのセルに含まれる文字列の行数に影響を受けることなく、文字列の各々に文字列の注釈記号 76 を付与し、当該文字列を矩形の枠線で囲い、当該枠線に「Text」のタグを目印として付ける。

[0136] レイアウト学習モデル 14 は、文字列の注釈記号 76 の範囲、及び表内における文字列の位置を学習する。電子文書生成装置 10 は、表を構成する全ての縦線及び横線に係るオブジェクトデータとともに、文字列に係るテキストデータを電子文書に出力することで表を再現することができる。

- [0137] 図22を参照して、表のセルの中の文字列の認識の学習に用いられるレイアウト学習データについて説明する。図22の表のセルに含まれる文字列の各々に文字列の注釈記号76を付与し、当該文字列を矩形の枠線で囲い、当該枠線に「Text」のタグを目印として付ける。
- [0138] レイアウト学習モデル14は、文字列の注釈記号76の範囲、及び文字列の文書内における位置情報について学習をする。電子文書生成装置10は、文字列に係るテキストデータを文書内における位置に置くことで表を電子文書内に再現することができる。電子文書生成装置10は、表を構成する縦線及び横線を電子文書内に再現することなく、テキストデータの出力のみで電子文書内に表を再現することができる。
- [0139] 図23を参照して、印章の認識の学習に用いられるレイアウト学習モデルについて説明する。図23に係るレイアウト学習モデルにより、レイアウト学習モデル14は、印章に係る要素を用いることなく、文字列に係る要素、及び文字列の下部に位置する空白により印章の範囲と位置を学習することができる。
- [0140] レイアウト学習用データ修正部41（図4参照）は、入力に基づいて、レイアウト認識部33により取得された複数の要素の各々の種類、及び複数の要素の各々の範囲の文書画像内における位置情報の少なくともいずれかが修正され、この修正されたデータを追加することでレイアウト学習用データを更新する。
- [0141] レイアウト認識部33に画像認識される前の文書画像61とレイアウト認識部33に画像認識された後の文書画像62との間に齟齬が生じる場合がある。例えば、文字列の一部が認識されない場合、画像として認識されるべき要素が印章として認識されてしまう場合、表の位置にズレが生じた場合などがある。
- [0142] このような場合に、レイアウト認識部33に画像認識された後の文書画像62について、レイアウト認識部33に画像認識される前の文書画像61に合わせるように修正を行い、この修正されたデータをレイアウト学習用デー

タに追加することで、レイアウト学習用データは更新される。

- [0143] レイアウト学習部42（図4参照）は、レイアウト学習用データ修正部41により更新されたレイアウト学習用データを用いて、レイアウト学習モデル14の再学習を行う。

レイアウト学習モデル14は再学習されることで、文書画像のレイアウトの認識精度を向上させることができる。

- [0144] 文字列学習用データ生成部43（図4参照）は、文字列学習モデル13の教師有り学習に用いる文字列学習用データを生成する。

文字列学習用データ修正部44（図4参照）は、入力に基づいて、文字列認識部35により生成されたテキストデータが修正され、この修正されたテキストデータを追加することで文字列学習用データを更新する。

- [0145] 文字列学習部45（図4参照）は、文字列学習用データ修正部44により更新された文字列学習用データを用いて、文字列学習モデル13の再学習を行う。

- [0146] 文字列学習用データ生成部43は、文書画像データベース15から文書画像を取得し、当該文書画像にアノテーションを付与して文字列学習用データを生成する。なお、文字列学習用データを生成する際に、文字列学習用データ生成部43を用いることなく、ユーザが手動で文字列学習用データを生成することもできる。ユーザが手動で文字列学習用データを生成する場合は、文書画像データベース15から取得した文書画像にユーザ端末12を用いてアノテーションを付与することができる。

- [0147] 図24を参照して、文字列学習モデル13の学習に用いられる文字列学習データについて説明する。図21は、アノテーションが付与された文字列学習データの例を示す図である。図24は、文字列学習用データ生成部43の出力画面であり、ユーザ端末12若しくは電子文書生成装置10の出力装置27に表示される。

- [0148] 文字列学習用データ生成部43は、文書画像データベース15から取得した文書画像に含まれる文字列について、当該文字列各々に対応するテキスト

データをテキストデータの注釈 85 として付与する。

[0149] なお、アノテーションの付与はテキストデータに替えて、対応する文字コードをテキストデータの注釈 85 として付与してもよい。文字列学習用データ生成部 43 は、文書画像に含まれる文字列が空白を含む場合、当該文字列に対応するテキストデータは同じ様に空白を含むようにして文字列学習用データを生成する。

[0150] 次に、図 25 を参照して、本実施形態に係る電子文書生成装置 10 によって実行される電子文書生成方法を電子文書生成プログラムとともに説明する。図 25 は、電子文書生成プログラムのフローチャートである。電子文書生成方法は、電子文書生成プログラムに基づいて、電子文書生成装置 10 の CPU 25 により実行される。

[0151] 電子文書生成プログラムは、電子文書生成装置 10 の CPU 25 に対して、文書画像取得機能、前処理機能、レイアウト認識機能、切出機能、文字認識機能、出力機能などの各種機能を実現させる。これらの機能は図 25 に示される順に実行されるが、適宜、順番を入れ替えて実行することもできる。なお、各機能は前述の電子文書生成装置 10 の説明と重複するため、その詳細な説明は省略する。

[0152] 文書画像取得機能は、文書を画像化した文書画像を取得する（S31：文書画像取得ステップ）。

文書画像の形式は、一例として PDF、JPG、及び GIF などがあり、この他電子文書生成装置 10 が画像として処理できるデータ形式のものは含み得る。

[0153] 前処理機能は、文書画像取得機能が取得した文書画像について前処理を行う（S32：前処理ステップ）。

前処理機能は背景除去機能、傾き補正機能、及び形状調整機能を備え、背景除去機能は文書画像取得機能が取得した文書画像の背景を除去し、傾き補正機能は文書画像取得機能が取得した文書画像の傾きを補正し、形状調整機能は文書画像取得機能が取得した文書画像の全体の形状及び大きさを調整す

る。

[0154] レイアウト認識機能は、文書画像に含まれる複数の要素と、当該複数の要素の各々の識別情報との対応関係を学習したレイアウト学習モデル14を用いて、文書画像取得機能に取得された文書画像に含まれる複数の要素の各々の文書画像内における範囲を特定し、複数の要素の各々の種類を認識し、及び複数の要素の各々の範囲に係る文書画像内における位置情報を取得する（S33：レイアウト認識ステップ）。

[0155] 要素の種類は、文書の種類に応じて必要なものと不要なものに分類してもよい。

この場合、レイアウト認識機能は、文書画像取得機能に取得された文書画像に含まれる複数の要素のうち、認識した要素が不要なものに該当する場合は当該要素の位置情報は取得されず、認識した要素が必要なものに該当する場合は当該要素の位置情報を取得することとしてもよい。または、レイアウト認識機能は、文書画像61に含まれる複数の要素のうち、必要な要素のみを認識し、当該要素の位置情報を取得することとしてもよい。

[0156] レイアウト認識機能は、要素の各々の種類を認識して、当該要素の各々の範囲に係る文書画像の位置情報を取得した後、要素同士が重なり合う、若しくは要素同士が離れすぎている場合には、実際の文書に基づいて、当該要素の各々の範囲及び取得した位置情報を補正する。

[0157] レイアウト認識機能は、表を構成する全ての縦線及び横線の各々の長さや位置を認識する。レイアウト認識機能は、表を構成する全ての縦線及び横線の長さや位置を把握することで、表に含まれる全てのセルについて把握する。すなわち、レイアウト認識機能は、隣接する2本の縦線及び隣接する2本の横線により構成される四角形をセルとして認識する。

[0158] さらに、レイアウト認識機能は表を構成する線の線種についても認識する。認識された線種は、取得した文書画像に基づいて電子文書を再現する際に、当該電子文書に含まれる表を構成する線のオブジェクトに反映される。従って、例えば、文書画像内の表の線が破線であった場合、文書画像に基づい

て再現された電子文書に含まれる表の線は破線のオブジェクトとして表現される。

[0159] 切出機能は、レイアウト認識機能により認識された種類が表に該当する要素において、当該要素に含まれる表の中のセルの各々を切り出し、セルの各々の文書画像内における位置情報を取得する（S 3 4：切出ステップ）。

切出機能は、レイアウト認識機能により認識された表を構成する全ての縦線及び横線を再生し、全てのセルの位置情報を生成する。

[0160] 切出機能により切り出されたセルは、複数の文字列を含む場合もある。切出機能は、切り出されたセルに複数行の文字列が含まれている場合は、さらに全ての文字列について文字列単体毎の画像を切り出す。

[0161] レイアウト認識機能によって認識された文字列の画像、及び切出し機能により切り出された文字列の画像は、1行ごとに文字認識機能に送り出される。

[0162] 文字認識機能は、文書画像と当該文書画像に含まれる文字列との対応関係を学習した文字列学習モデルを用いて、文書画像取得機能に取得された文書画像に含まれる文字列を文字認識し、当該文字列に係るテキストデータを生成する（S 3 5：文字認識ステップ）。

[0163] 出力機能は、テキストデータを電子媒体のテキストとして出力する（S 3 6：出力ステップ）。

出力機能は、レイアウト認識機能により取得された文字列の位置情報、及び切出部により取得されたセルの文書画像内における位置情報に基づいてテキストデータを出力し、電子媒体のテキストとして再生する。

[0164] 次に、上記した電子文書生成プログラムについて、図26から図28を参照して、領収書の文書画像を電子文書に変換する一実施形態について説明する。図26から図28は、電子文書生成プログラムに係る一実施形態のフローチャートである。図26から図28に示すフローチャートは、これらを結合することで、1つの電子文書生成プログラムのフローチャートを示すこととなる。

[0165] ステップS 1 0 2において、文書画像取得部3 1は、文書画像データベース1 5より文書画像若しくはPDFを取得する。

ステップS 1 0 3において、文書画像取得部3 1が取得したデータがPDFか否かの判定を行う。PDFではない場合（No：S 1 0 3）、即ち、文書画像取得部3 1が取得したデータが文書画像であった場合、ステップS 1 0 6に移行する。

[0166] 文書画像取得部3 1が取得したデータがPDFだった場合（Yes：S 1 0 3）、ステップS 1 0 4に移行し、当該PDFを文書画像に変換した後に当該文書画像を取得する（S 1 0 5）。

[0167] ステップS 1 0 6において、前処理部3 2は、取得した文書画像について前処理を行う。前処理部3 2は、背景除去部3 2 a、傾き補正部3 2 b、及び形状調整部3 2 cを備える。

[0168] 背景除去部3 2 aは、取得した文書画像の背景を除去する。傾き補正部3 2 bは、取得した文書画像に含まれる文字列が傾いている場合に傾き補正を行い文字列の傾きを補正する。形状調整部3 2 cは、取得した文書画像の全体の形状及び大きさの調整を行う。

[0169] ステップS 1 0 7において、レイアウト認識部3 3は、前処理部3 2により行われた前処理を経た文書画像を取得する。

取得された前処理後の文書画像は、後述のステップS 1 1 5、ステップS 1 2 0、及びステップS 1 3 6の文書画像切り出し処理に送られる。

[0170] ステップS 1 0 8及びステップS 1 0 9において、レイアウト認識部3 3は、文書画像のレイアウト認識を行い、文書画像に含まれる複数の要素について、要素ごとにその範囲を特定し、要素ごとに種類と位置情報とを取得する。

要素の種類は、文字列、表、画像、印章、手書きである。

[0171] ステップS 1 1 0において、レイアウト認識部3 3は、取得した要素の最小境界ボックスの位置情報の調整処理を行う。

最小境界ボックスとは、要素を囲う矩形のうち面積が最小のものをいい、

当該要素が占める範囲を意味する。レイアウト認識部33は、文書画像と取得した要素とを照合し、文書画像と取得した要素の位置情報との間に齟齬があった場合は取得した要素の最小境界ボックスの位置情報の調整を行う。

[0172] ステップS111において、レイアウト認識部33は、ステップS110にて実施された最小境界ボックスの調整処理後のレイアウト情報を取得する。当該レイアウト情報には、要素の種類、及び位置情報が含まれる。

[0173] ステップS112において、レイアウト認識部33は、後述するステップS130の処理により送られてくる内部記憶された要素のレイアウト情報を参照して、文書画像の中に他の要素が残っているか否かを判定する。

[0174] ステップS130の処理により送られてくる内部記憶された要素のレイアウト情報の中に全ての要素のレイアウト情報が含まれている場合、レイアウト認識部33は、文書画像の中に他の要素が残っていないと判定して（No：S112）、ステップS131に移行して、ステップS112からステップS130のループの終了処理を行いステップS132に移行する。

[0175] 一方、ステップS130の処理により送られてくる内部記憶された要素のレイアウト情報の中に全ての要素のレイアウト情報が含まれていない場合、レイアウト認識部33は、文書画像の中に他の要素が残っていると判定して（Yes：S112）、ステップS113に移行する。

[0176] ステップS113において、レイアウト認識部33は、文書画像の中に残る要素が表であるか否かを判定する。

文書画像に表が残っていない場合（No：S113）、後述のステップS130へ表以外のレイアウト情報を送る。

[0177] 文書画像に表が残っている場合（Yes：S113）、ステップS114に移行する。なお、文書画像は領収書に係るものであるため表を含む場合が多い。従って、文書画像が表を含まないと判断された場合、レイアウト認識部33は処理を中断して、電子文書が領収書に係るものか否かを確認するようにしてもよい。

[0178] ステップS114において、レイアウト認識部33は、文書画像内の表を

構成する全ての縦線及び横線の大きさ及び位置情報を取得する。表を構成する全ての縦線及び横線の大きさ及び位置情報を取得すれば、表に含まれる全てのセルについて、そのセルの大きさ及び位置を取得することができる。

[0179] ステップS 1 1 5において、切出部3 4は、ステップS 1 0 7の処理により取得された前処理後の文書画像の中から表の画像の切り出し処理を行う。

ステップS 1 1 6において、切出部3 4は、ステップS 1 1 5にて切り出された表の画像を取得する。

[0180] ステップS 1 1 7及びステップS 1 1 8において、切出部3 4は、ステップS 1 1 6にて取得された表の画像からセルを抽出する処理を行い（ステップS 1 1 7）、セルの情報を取得する（ステップS 1 1 8）。

[0181] セルの情報とは、表におけるセルの位置情報に相当する行、列、及び座標のことである。

ステップS 1 1 8にて取得されたセルの情報は、後述するステップS 1 2 7に送られる。

[0182] ステップS 1 1 9において、切出部3 4は、ステップS 1 2 7の処理により送られてきた内部記憶された表のレイアウト情報を参照して、表の中に他のセルが残っているか否かを判定する。

[0183] ステップS 1 2 7の処理により送られてきた内部記憶されたセルのレイアウト情報の中に全てのセルのレイアウト情報が含まれている場合、切出部3 4は、表の中に他のセルが残っていないと判定して（N o : S 1 1 9）、ステップS 1 2 8に移行して、ステップS 1 1 9からステップS 1 2 7のループの終了処理を行いステップS 1 3 0に移行する。

[0184] 一方、ステップS 1 2 7の処理により送られてくる内部記憶されたセルのレイアウト情報の中に全てのセルのレイアウト情報が含まれていない場合、切出部3 4は、表の中に他のセルが残っていると判定して（Y e s : S 1 1 9）、ステップS 1 2 0に移行する。

[0185] ステップS 1 2 0において、切出部3 4は、ステップS 1 0 7の処理により取得された前処理後の文書画像からセルの画像を切り出す処理を行う。

ステップS 1 2 1において、切出部3 4は、ステップS 1 2 0の処理により切り出されたセルの画像を取得する。

[0186] ステップS 1 2 2において、文字列認識部3 5は、ステップS 1 2 1の処理により取得されたセルの画像について文字列認識の処理を行う。

ステップS 1 2 3において、文字列認識部3 5は、文字列認識の処理が行われた文字列の位置情報を取得する。

[0187] ステップS 1 2 4において、文字列認識部3 5は、ステップS 1 2 3の処理により取得した文字列の最小境界ボックスの位置情報の調整処理を行う。

文字列認識部3 5は、文書画像と取得した文字列の位置情報とを照合し、文書画像と取得した文字列の位置情報との間に齟齬があった場合は取得した文字列の最小境界ボックスの位置情報の調整を行う。

[0188] ステップS 1 2 5において、文字列認識部3 5は、ステップS 1 2 4にて実施された文字列の最小境界ボックスの位置情報の調整処理後の位置情報を取得する。

[0189] ステップS 1 2 6及びステップS 1 2 7において、文字列認識部3 5は、ステップS 1 1 8の処理によって取得されたセルの情報とステップS 1 2 5の処理によって取得された調整後の文字列の位置情報とを併合し（ステップS 1 2 7）、表のレイアウト情報として内部記憶装置に内部記憶する（ステップS 1 2 6）。内部記憶装置とは、図2に示すRAM 2 3若しくは記憶部2 4の何れか、若しくはその両方のことをいう。

[0190] ステップS 1 1 9からステップS 1 2 7の処理は、表に含まれる全てのセルについて行われる。ステップS 1 1 9からステップS 1 2 7の処理が表に含まれる最後のセルについて行われた後にステップS 1 2 8のループの終了処理が行われて、文字列認識部3 5はステップS 1 3 0に移行する。

[0191] ステップS 1 2 9及びステップS 1 3 0において、出力部3 6は、ステップS 1 2 6の処理により取得された表のレイアウト情報とステップS 1 1 3の処理により取得された表以外のレイアウト情報とを併合し（ステップS 1 3 0）、全ての要素のレイアウト情報として内部記憶装置に内部記憶する（

ステップS 1 2 9)。

[0192] ステップS 1 1 2からステップS 1 3 0の処理は、文書画像に含まれる全ての要素について行われる。ステップS 1 1 2からステップS 1 3 0の処理が文書画像に含まれる最後の要素について行われた後にステップS 1 3 1のループの終了処理が行われて、文字列認識部3 5はステップS 1 3 2に移行する。

[0193] ステップS 1 3 2において、文字列認識部3 5は、文書画像に他の要素が残っているか否かの判定を行う。

文字列認識部3 5は、後述のステップS 1 4 0の処理により送られてきた内部記憶された要素のレイアウト情報を参照して、文書画像の中に他の要素が残っているか否かを判定する。

[0194] ステップS 1 4 0の処理により送られてきた内部記憶された要素のレイアウト情報の中に全ての要素のレイアウト情報が含まれている場合、文字列認識部3 5は、文書画像の中に他の要素が残っていないと判定して（No : S 1 3 2）、ステップS 1 4 1に移行して、ステップS 1 3 2からステップS 1 4 0のループの終了処理を行いステップS 1 4 2に移行する。

[0195] 一方、ステップS 1 4 0の処理により送られてくる内部記憶された要素のレイアウト情報の中に全ての要素のレイアウト情報が含まれていない場合、文字列認識部3 5は、文書画像の中に他の要素が残っていると判定して（Yes : S 1 3 2）、ステップS 1 3 3に移行する。

[0196] ステップS 1 3 3において、文字列認識部3 5は、文書画像の中に残っている要素が文字列であるか否かを判定する。

文字列認識部3 5が、文書画像に残っている要素が文字列であると判定した場合（Yes : S 1 3 3）、ステップS 1 3 5に移行する。

[0197] 文字列認識部3 5が、文書画像に残っている要素が文字列ではないと判定した場合（No : S 1 3 3）、ステップS 1 3 2に移行するループの続行処理が行われる（ステップS 1 3 4）。

ステップS 1 3 5において、文字列認識部3 5は、文字列の位置情報を取

得する。

- [0198] ステップS 1 3 6 及びステップS 1 3 7 において、文字列認識部 3 5 は、ステップS 1 0 7 の処理によって取得された前処理後の文書画像の中から文字列の画像を切り出し（ステップS 1 3 6 ）、文字列の画像を取得する。
- [0199] ステップS 1 3 8 及びステップS 1 3 9 において、文字列認識部 3 5 は、ステップS 1 3 7 の処理によって取得された文字列の画像について文字列認識の処理を行い（ステップS 1 3 8 ）、文字列認識の処理により予測されたテキストデータを生成する（ステップS 1 3 9 ）。
- [0200] ステップS 1 4 0 において、文字列認識部 3 5 は、ステップS 1 3 5 の処理によって取得された文字列の位置情報とステップS 1 3 9 の処理によって生成されたテキストデータとを併合して要素のレイアウト情報を生成する。生成された要素のレイアウト情報は、ステップS 1 2 9 に送られる。ステップS 1 2 9 では、送られてきた要素のレイアウト情報を内部記憶装置に内部記憶する。内部記憶装置とは、図 2 に示す R A M 2 3 若しくは記憶部 2 4 の何れか、若しくはその両方のことをいう。
- [0201] ステップS 1 3 2 からステップS 1 4 0 までの処理は、ステップS 1 4 0 の処理によって送られてきた要素のレイアウト情報の中に全ての要素のレイアウト情報が含まれているとステップS 1 3 2 の処理によって判定されるまで行われる。
- [0202] ステップS 1 4 1 において、ステップS 1 4 0 の処理によって送られてきた要素のレイアウト情報の中に全ての要素のレイアウト情報が含まれているとステップS 1 3 2 の処理によって判定されたことを受けて、ステップS 1 3 2 からステップS 1 4 0 までのループの終了処理が行われて、ステップS 1 4 2 に移行される。
- [0203] ステップS 1 4 2 において、電子文書生成装置 1 0 は後処理を行う。後処理では、全ての要素のテキストデータ、画像、及び位置情報について、J S O N (j a v a s c r i p t o b j e c t n o t a t i o n) への出力、及びT S V (T a b - S e p a r a t e d V a l u e s) への変換など

が行われる。

なお、上記した各機能部の処理は、電子文書生成装置 10 の CPU 25 により実行される処理である。

[0204] ステップ S 143 において、出力部 36 は、後処理を経た全ての要素の情報について、最終形態として単純なテキストファイル、HTML (HyperText Markup Language)、市販されている文字編集ソフトで編集可能なファイル形式、及び編集可能な PDF などの電子文書として出力する。

[0205] 上記した実施形態によれば、電子文書生成装置 10 は、文書画像のレイアウトについてレイアウト学習モデル 14 を用いて認識し、その上で文字列学習モデル 13 を用いて文書画像の文字認識を行う。すなわち、電子文書生成装置 10 は、文書画像に含まれる複数の要素の種類を特定し、要素の種類に適した文字認識を行うので、文字認識の認識精度を向上させることができる。

[0206] さらに、上記した実施形態によれば、電子文書生成装置 10 は、従来からの OCR テキスト認識技術で行っていた一文字単位の文字認識と比較して、文字列学習モデル 13 を用いて文書画像の文字認識を文字列ごとに行うので、文字認識の際の認識効率を向上させることができる。

[0207] さらに、上記した実施形態によれば、電子文書生成装置 10 は、文字認識を行う際に、一文字単位の文字認識ではなく、文字列ごとに文字認識を行うので、文字に重なって存在するノイズなどの影響を抑制して文字認識を行うことができ、一文字単位で行う文字認識と比較して文字認識の認識精度を向上させることができる。

[0208] さらに、上記した実施形態によれば、従来の OCR テキスト認識技術を用いた文字認識では誤認識するような文字でも、文字列学習モデル 13 を用いた文字認識では正しく認識することができる。例えば、文字の上に印章が重ねられた場合、その文字について従来の OCR テキスト認識技術では誤認識する可能性があったが、文字列学習モデル 13 を用いた文字認識では正しく

認識することができる。

[0209] さらに、上記した実施形態によれば、電子文書生成装置 10 は、種類が表に該当する要素については、セル単体に係る画像に含まれる文字列ごとに文字認識を行うので、表に含まれる文字列の文字認識の認識精度を向上させることができる。

[0210] さらに、上記した実施形態によれば、アノテーションが付与された文字列学習データ及びレイアウト学習データにより、文字列学習モデル 13 及びレイアウト学習モデル 14 を学習するので、レイアウト認識部 33 及び文字列認識部 35 の認識精度を向上させることができる。

[0211] さらに、上記した実施形態によれば、文書画像の中に表が有る場合は表を構成する全ての縦線及び横線について先ず認識し、当該表に含まれる全てのセルについて認識する。その後で、全てのセルについて、表の内部における位置情報に影響を受けることが無くセルごとの画像に文字列認識を行うので、セルの内部の文字列の文字認識の認識精度を向上させることができる。

[0212] 本開示は上記した実施形態に係る電子文書生成装置 10 に限定されるものではなく、特許請求の範囲に記載した本開示の要旨を逸脱しない限りにおいて、その他種々の変形例、若しくは応用例により実施可能である。

符号の説明

- [0213] 10 電子文書生成装置
11 情報通信ネットワーク
12 ユーザ端末
13 文字列学習モデル
14 レイアウト学習モデル
15 文書画像データベース
20 入出力インターフェース
21 通信インターフェース
22 ROM
23 RAM

- 2 4 記憶部
- 2 5 C P U
- 2 6 入力装置
- 2 7 出力装置
- 2 8 G P U
- 3 1 文書画像取得部
- 3 2 前処理部
 - 3 2 a 背景除去部
 - 3 2 b 傾き補正部
 - 3 2 c 形状調整部
- 3 3 レイアウト認識部
- 3 4 切出部
- 3 5 文字列認識部
- 3 6 出力部
- 4 0 レイアウト学習用データ生成部
- 4 1 レイアウト学習用データ修正部
- 4 2 レイアウト学習部
- 4 3 文字列学習用データ生成部
- 4 4 文字列学習用データ修正部
- 4 5 文字列学習部
- 4 7 傾き補正前の文書画像
- 4 8 文字列
- 4 9 表
- 5 0 ホッチキス跡
- 5 1 手書き
- 5 2 印章
- 5 3 画像
- 5 4 ノイズ除去

5 5 前処理

5 6 レイアウト認識処理

5 7 文字列認識処理

5 8 a、5 9 a、6 0 a 文書画像

5 8 b、5 9 b、6 0 b 文書画像

6 1、6 2 文書画像

6 3、6 4 表

6 5 縦線

6 6 横線

6 7 セル画像

6 9、7 0、7 3 文字列の画像

7 1 a 文字列の画像

7 1 b テキストデータ

7 2 認識範囲

7 3 表

7 5 認識範囲

7 6 文字列の注釈記号

7 7 表の注釈記号

7 8 画像の注釈記号

7 9 印章の注釈記号

8 0 外枠の注釈記号

8 1 ノイズの注釈記号

8 2 手書きの注釈記号

8 3 縦線の注釈記号

8 4 横線の注釈記号

8 5 テキストデータの注釈

1 0 0 電子文書生成システム

S 3 1 文書画像取得ステップ

S 3 2 前処理ステップ

S 3 3 レイアウト認識ステップ

S 3 4 切出ステップ

S 3 5 文字認識ステップ

S 3 6 出力ステップ

請求の範囲

- [請求項1] 文書を画像化した文書画像を取得する文書画像取得部と、
文書画像と当該文書画像に含まれる文字列との対応関係を学習した文字列学習モデルを用いて、
前記文書画像取得部に取得された前記文書画像に含まれる文字列を文字認識し、当該文字列に係るテキストデータを生成する文字列認識部と、
前記テキストデータを電子媒体のテキストとして出力する出力部と、
、
を備えることを特徴とする電子文書生成装置。
- [請求項2] 文書画像に含まれる複数の要素と、当該複数の要素の各々の識別情報との対応関係を学習したレイアウト学習モデルを用いて、
前記文書画像取得部に取得された前記文書画像に含まれる複数の要素の各々の前記文書画像内における範囲を特定し、前記複数の要素の各々の種類を認識し、前記複数の要素の各々の前記範囲に係る前記文書画像内における位置情報を取得するレイアウト認識部をさらに備え、
、
前記文字列認識部は、前記レイアウト認識部により認識された前記範囲に含まれる文字列について、前記文字列学習モデルを用いて文字認識し、前記文字列に係るテキストデータを生成し、
前記出力部は、前記複数の要素に係る前記範囲の各々の前記位置情報に、前記複数の要素に係る前記テキストデータの各々を電子媒体のテキストとして出力する、
ことを特徴とする請求項1に記載の電子文書生成装置。
- [請求項3] 前記要素の前記種類は、文字列、表、画像、印章、又は手書きのいずれかである、
ことを特徴とする請求項2に記載の電子文書生成装置。
- [請求項4] 前記レイアウト認識部により認識された前記種類が前記表に該当す

る前記要素において、当該要素に含まれる前記表の中のセルの各々を切り出し、前記セルの各々の前記文書画像内における位置情報を取得する切出部をさらに備え、

前記文字列認識部は、前記切出部に切り出された前記セルの各々に含まれる文字列について、前記文字列学習モデルを用いて文字認識を行い、前記文字列に係るテキストデータを生成する、

ことを特徴とする請求項3に記載の電子文書生成装置。

[請求項5] 複数の前記要素を含む文書画像であって、当該要素に当該要素の各々に該当する前記種類に関連付けられたアノテーションが付与されており、

前記アノテーションが付与された複数の前記文書画像を蓄積してレイアウト学習用データを生成するレイアウト学習用データ生成部をさらに備え、

前記レイアウト学習用データは前記レイアウト学習モデルの教師有り学習に用いられる、

ことを特徴とする請求項2ないし4のいずれか1項に記載の電子文書生成装置。

[請求項6] 前記文書画像に、前記アノテーションとともに前記文書画像に含まれる前記複数の要素に係る範囲の各々の前記文書画像内における位置情報が付与されることを特徴とする請求項5に記載の電子文書生成装置。

[請求項7] 入力に基づいて、前記レイアウト認識部により認識された前記複数の要素の各々の種類、及び前記複数の要素の各々の範囲の前記文書画像内における位置情報の少なくともいずれかが修正され、この修正されたデータを追加することで前記レイアウト学習用データを更新するレイアウト学習用データ修正部をさらに備える、

ことを特徴とする請求項5または6に記載の電子文書生成装置。

[請求項8] 前記レイアウト学習用データ修正部により更新された前記レイアウト

ト学習用データを用いて、前記レイアウト学習モデルの再学習を行うレイアウト学習部をさらに備える、

ことを特徴とする請求項 7 に記載の電子文書生成装置。

[請求項9]

前記文字列学習モデルの教師有り学習に用いる文字列学習用データを生成する文字列学習用データ生成部をさらに備える、

ことを特徴とする請求項 2 ないし 8 のいずれか 1 項に記載の電子文書生成装置。

[請求項10]

入力に基づいて、前記文字列認識部により生成されたテキストデータが修正され、この修正されたテキストデータを追加することで前記文字列学習用データを更新する文字列学習用データ修正部をさらに備える、

ことを特徴とする請求項 9 に記載の電子文書生成装置。

[請求項11]

前記文字列学習用データ修正部により更新された前記文字列学習用データを用いて、前記文字列学習モデルの再学習を行う文字列学習部をさらに備える、

ことを特徴とする請求項 10 に記載の電子文書生成装置。

[請求項12]

前記文字列認識部は、複数の前記文字列学習モデルを備え、前記複数の要素の各々に含まれる文字列の言語に適応した前記文字列学習モデルを用いる、

ことを特徴とする請求項 2 ないし 11 のいずれか 1 項に記載の電子文書生成装置。

[請求項13]

前記文書画像取得部が取得した文書画像について前処理を行う前処理部をさらに備え、

前記前処理部は、背景除去部、傾き補正部、及び形状調整部を備え、

前記背景除去部は、前記文書画像取得部が取得した前記文書画像の背景を除去し、

前記傾き補正部は、前記文書画像取得部が取得した前記文書画像

の傾きを補正し、

前記形状調整部は、前記文書画像取得部が取得した前記文書画像の全体の形状及び大きさを調整する

ことを特徴とする請求項2ないし12のいずれか1項に記載の電子文書生成装置。

[請求項14] 前記レイアウト学習モデルは、契約書用のレイアウト学習モデル、請求書用のレイアウト学習モデル、覚書用のレイアウト学習モデル、納品書用のレイアウト学習モデル、又は領収書用のレイアウト学習モデルのいずれかであることを特徴とする請求項2ないし13のいずれか1項に記載の電子文書生成装置。

[請求項15] 電子文書生成装置に用いられるコンピュータが、
文書を画像化した文書画像を取得する文書画像取得ステップと、
文書画像と当該文書画像に含まれる文字列との対応関係を学習した文字列学習モデルを用いて、
前記文書画像取得ステップにて取得された前記文書画像に含まれる文字列を文字認識し、当該文字列に係るテキストデータを生成する文字列認識ステップと、
前記テキストデータを電子媒体のテキストとして出力する出力ステップと、
を実行することを特徴とする電子文書生成方法。

[請求項16] 電子文書生成装置に用いられるコンピュータに、
文書を画像化した文書画像を取得する文書画像取得機能と、
文書画像と当該文書画像に含まれる文字列との対応関係を学習した文字列学習モデルを用いて、
前記文書画像取得機能にて取得された前記文書画像に含まれる文字列を文字認識し、当該文字列に係るテキストデータを生成する文字列認識機能と、
前記テキストデータを電子媒体のテキストとして出力する出力機能

と、

を発揮させることを特徴とする電子文書生成プログラム。