

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3913985号
(P3913985)

(45) 発行日 平成19年5月9日(2007.5.9)

(24) 登録日 平成19年2月9日(2007.2.9)

(51) Int. Cl.

G06K 9/20 (2006.01)

F I

G06K 9/20 340L

請求項の数 11 (全 16 頁)

(21) 出願番号	特願2000-611236 (P2000-611236)	(73) 特許権者	000005223
(86) (22) 出願日	平成11年4月14日(1999.4.14)		富士通株式会社
(86) 国際出願番号	PCT/JP1999/001986		神奈川県川崎市中原区上小田中4丁目1番1号
(87) 国際公開番号	W02000/062243	(74) 代理人	100074099
(87) 国際公開日	平成12年10月19日(2000.10.19)		弁理士 大菅 義之
審査請求日	平成16年9月6日(2004.9.6)	(74) 代理人	100067987
前置審査			弁理士 久木元 彰
		(72) 発明者	藤本 克仁
			神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
		(72) 発明者	鎌田 洋
			神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

最終頁に続く

(54) 【発明の名称】 文書画像中の基本成分に基づく文字列抽出装置および方法

(57) 【特許請求の範囲】

【請求項 1】

入力された文書画像から基本成分の集合を抽出する基本成分抽出手段と、
 前記基本成分の集合に含まれる基本成分間の包含関係を用いて基本成分が文字成分に対応するか否かを判定し、文字成分の集合を抽出する文字成分抽出手段と、
 前記文字成分の集合を用いて文字列を抽出する文字列抽出手段とを備え、
前記文字成分抽出手段は、所定数以上の基本成分を含む基本成分、および、所定数以上の基本成分と重なり合っている基本成分を、文字成分ではないと判定することを特徴とする文字列抽出装置。

【請求項 2】

前記文字列抽出手段により抽出された文字列の認識を行って、文字列の確信度を求める文字列認識手段をさらに備え、前記文字成分抽出手段は、所定のしきい値以上の確信度を持つ文字列に含まれる文字成分を真の文字成分であると判定し、該しきい値より小さい確信度を持つ文字列に含まれる文字成分を文字成分ではないと判定して、新たな文字成分の集合を抽出し、該文字列抽出手段は、該新たな文字成分の集合を用いて、再度、文字列を抽出することを特徴とする請求項 1 記載の文字列抽出装置。

【請求項 3】

前記文字成分抽出手段と文字列抽出手段は、新たな文字成分の集合の抽出と文字列の再抽出を複数回繰り返すことを特徴とする請求項 2 記載の文字列抽出装置。

【請求項 4】

10

20

前記基本成分抽出手段は、前記文書画像として多階調画像が入力されたとき、該多階調画像において所定の範囲の階調レベルを持つ画素の連結成分を求め、該連結成分および該連結成分の外接図形のうち少なくとも一方を基本成分として抽出することを特徴とする請求項 2 または 3 記載の文字列抽出装置。

【請求項 5】

前記基本成分抽出手段は、前記文書画像としてカラー画像が入力されたとき、該カラー画像において所定の範囲の色情報を持つ画素の連結成分を求め、該連結成分および該連結成分の外接図形のうち少なくとも一方を基本成分として抽出することを特徴とする請求項 2 または 3 記載の文字列抽出装置。

【請求項 6】

前記基本成分抽出手段は、前記文書画像として多階調画像が入力されたとき、該多階調画像において所定の範囲の階調レベルを持つ画素の連結成分を求め、該連結成分の外接図形に含まれる画像を二値化して二値画像を生成し、得られた二値画像における画素連結成分および該画素連結成分の外接図形のうち少なくとも一方を基本成分として抽出することを特徴とする請求項 2 または 3 記載の文字列抽出装置。

【請求項 7】

前記基本成分抽出手段は、前記文書画像としてカラー画像が入力されたとき、該カラー画像において所定の範囲の色情報を持つ画素の連結成分を求め、該連結成分の外接図形に含まれる画像を二値化して二値画像を生成し、得られた二値画像における画素連結成分および該画素連結成分の外接図形のうち少なくとも一方を基本成分として抽出することを特徴とする請求項 2 または 3 記載の文字列抽出装置。

【請求項 8】

前記基本成分抽出手段は、前記文書画像として多階調画像およびカラー画像のうち的一方が入力されたとき、該文書画像のエッジ二値画像を生成し、得られたエッジ二値画像におけるエッジ連結成分を求め、該エッジ連結成分および該エッジ連結成分の外接図形のうち少なくとも一方を基本成分として抽出することを特徴とする請求項 2 または 3 記載の文字列抽出装置。

【請求項 9】

前記基本成分抽出手段は、前記文書画像として多階調画像およびカラー画像のうち的一方が入力されたとき、該文書画像のエッジ二値画像を生成し、得られたエッジ二値画像におけるエッジ連結成分を求め、該エッジ連結成分の外接図形に含まれる画像を二値化して二値画像を生成し、得られた二値画像における画素連結成分および該画素連結成分の外接図形のうち少なくとも一方を基本成分として抽出することを特徴とする請求項 2 または 3 記載の文字列抽出装置。

【請求項 10】

入力された文書画像に含まれる文字成分の集合に基づいて文字列を抽出するコンピュータのためのプログラムを記録した記録媒体であって、

前記文書画像から基本成分の集合を抽出するステップと、

前記文書画像に含まれる基本成分間の包含関係を用いて基本成分が文字成分に対応するか否かを判定するステップと、

所定数以上の基本成分を含む基本成分を、文字成分ではないと判定するステップと、

判定結果に基づいて前記文字成分の集合を抽出するステップと、

前記文字成分の集合を用いて文字列を抽出するステップと、

抽出された文字列の認識を行って文字列の確信度を求めるステップと、

所定のしきい値以上の確信度を持つ文字列に含まれる文字成分を真の文字成分であると判定し、該しきい値より小さい確信度を持つ文字列に含まれる文字成分を文字成分ではないと判定して、新たな文字成分の集合を抽出するステップと、

前記新たな文字成分の集合を用いて、再度、文字列を抽出するステップとを含む処理を前記コンピュータに実行させるための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。

10

20

30

40

50

【請求項 11】

入力された文書画像から基本成分の集合を抽出し、
前記基本成分の集合に含まれる基本成分間の包含関係を用いて基本成分が文字成分に対応するか否かを判定し、

所定数以上の基本成分を含む基本成分を、文字成分ではないと判定し、

判定結果に基づいて文字成分の集合を抽出し、

前記文字成分の集合を用いて文字列を抽出し、

抽出された文字列の認識を行って文字列の確信度を求め、

所定のしきい値以上の確信度を持つ文字列に含まれる文字成分を真の文字成分であると判定し、該しきい値より小さい確信度を持つ文字列に含まれる文字成分を文字成分ではないと判定して、新たな文字成分の集合を抽出し、

10

前記新たな文字成分の集合を用いて、再度、文字列を抽出することを特徴とする文字列抽出方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、文書画像に含まれる文字、図形等の情報の基本成分に基づいて、文書画像から文字列を抽出する文字列抽出装置およびその方法に関する。

【背景技術】

【0002】

20

文書画像における文字列パターンは、1つ以上の文字パターンの並びに対応し、文字パターンは、任意の言語の文字、記号等のパターンに対応する。文字列抽出装置は、文書画像を入力とし、文書画像中の文字列パターンを抽出して、後段の文字コード化処理あるいは検索処理に提供する装置である。現在、二値文書画像を入力とする文字列抽出装置が製品として存在している。

【0003】

また、近年、情報共有のための文書管理システムが注目されており、構造を持った電子化文書、構造を持たない生の画像文書、紙に記録された文書のような様々な文書の一元管理の仕組みが求められている。

【0004】

30

そこで、構造を持たない画像文書や紙文書も含めた検索可能化技術として、文書画像から情報検索のためのテキスト情報を抽出する文字列抽出装置に対する期待が高まってきた。特に、写真を含むグレースケール文書やカラー文書が増加しているため、これらの文書から高精度に文字列を抽出する技術の必要性が増大している。

【0005】

このような要求に応えるため、汎用性があり、様々な情報が混在した文書を扱うことが可能な文字列抽出技術がいくつか提案されている。これらの技術では、文書構造の先見の知識を必要とせず、図と文章領域の混在、文章の横書きと縦書きの混在、図中の文字列の抽出等が視野に入れられている。それらのうち代表的なものを次に説明することにする。

【0006】

40

ただし、画像のぼかしを用いる方法、黒画素の投影分布を用いる方法、および局所領域の画像特徴を用いる方法は、段組の入り組んだ文字列や図中の文字列の抽出には適さないため、除外した。

【0007】

従来の文字列抽出技術は、文字の一部または全部を表す画像パターンである文字成分の集合を何らかの方法で抽出し、その文字成分の大きさの同質性や相互の近接性を用いて、文字成分の部分集合としての文字列を抽出するといったような、基本的考え方に基づいている。この場合、文字成分抽出の精度が文字列抽出の精度に大きな影響を与える。従来の文字成分抽出方法としては、以下のようなものがある。

【0008】

50

文字成分抽出の第1の考え方は、文字、図形等の一部または全部を表す画像パターンである基本成分の集合を文書画像から何らかの方法で抽出し、その基本成分を大きさ・形状により分類して、文字成分のみ抽出することである。

【0009】

例えば、公開特許公報の特開昭61-072374(文字認識装置)および特開昭61-026150(文書画像ファイル登録検索装置)では、文書画像中の文字の大きさがほぼ一定であるという前提を用いて、文字成分の抽出を行っている。

【0010】

また、特開昭62-165284(文字列抽出方式)および特開平09-167233(画像処理方法および画像処理装置)では、二値画像における黒画素連結成分の外接矩形を基本成分とし、大きさが一定値以下の基本成分を文字成分とみなして抽出している。

10

また、特開平06-111060(光学的文字読取装置)では、カラー画像の色ごとの連結成分を基本成分とし、大きさが一定値以下の基本成分を文字成分とみなして抽出し、カラー画像中の文字列抽出を可能としている。

【0011】

また、文字成分抽出の第2の考え方は、基本成分あるいは基本成分の集合から構成される文字列候補を文字認識あるいは文字列認識して得られた確信度を用いて、基本成分の集合を文字成分とそうでないものとに分類することである。

例えば、特開平05-028305(画像認識装置および認識方法)では、基本成分の近接性により文字列候補を生成し、文字認識結果の評価値により文字列らしいもののみを残して、基本成分の抽出ひいては文字列の抽出を行っている。

20

【0012】

また、上述した第1および第2の考え方を併せ持つ方法も提案されている。例えば、特開平07-168911(文書認識装置)では、黒画素連結成分の外接矩形を基本成分とし、基本成分を大きさ・縦横比により文字候補・図形候補・罫線候補・画像候補に分類している。そして、文字候補を文字認識して得られた確信度が低い場合には、それを画像候補に変更し、図形候補を文字認識して得られた確信度が高い場合には、それを文字候補に変更して、文字成分を抽出している。

【0013】

しかしながら、このような従来の文字列抽出技術では文字成分抽出の精度が充分ではなく、結果として文字列抽出自身の精度も充分ではないという問題がある。

30

第1の考え方では、文字と図形のように異なる種類の情報に対応する基本成分が同じ程度の大きさを持つ場合に、文字成分抽出に失敗し、結果的に文字列抽出にも失敗してしまう。このため、抽出精度が充分でなくなる。

【0014】

第2の考え方では、このような場合でも、文字認識あるいは文字列認識を行うことにより、文字成分とそうでないものを区別できる可能性が高くなる。しかし、現在の文字認識技術の水準では、文字認識結果の確信度自身の信頼性がそれほど高くはない。

【0015】

このため、高い信頼度で文字成分であると判定するための確信度のしきい値は、高い信頼度で文字成分ではないと判定するための確信度のしきい値と大きく異なる値に設定せざるを得ない。その結果、2つのしきい値の中間的な確信度を持つ基本成分に対する判定が困難となり、強引に文字成分である/文字成分ではないと判定した場合に、文字成分抽出の精度が充分でなくなる。

40

【発明の開示】

【発明が解決しようとする課題】

【0016】

本発明の課題は、文書画像に含まれる基本成分を用いて、より正確に文字列を抽出する文字列抽出装置およびその方法を提供することである。

【課題を解決するための手段】

50

【 0 0 1 7 】

図 1 は、本発明の文字列抽出装置の原理図である。図 1 の文字列抽出装置は、基本成分抽出手段 1、文字成分抽出手段 2、および文字列抽出手段 3 を備える。基本成分抽出手段 1 は、入力された文書画像から基本成分の集合を抽出する。文字成分抽出手段 2 は、基本成分の集合に含まれる基本成分間の包含関係を用いて基本成分が文字成分に対応するか否かを判定し、文字成分の集合を抽出する。文字列抽出手段 3 は、文字成分の集合を用いて文字列を抽出する。

【 0 0 1 8 】

文書画像として、二値画像、多階調画像、カラー画像等が入力されると、基本成分抽出手段 1 は、画素連結成分、画素連結成分の外接矩形等を基本成分として抽出する。次に、文字成分抽出手段 2 は、抽出された基本成分間の包含関係を用いて、各基本成分が文字成分に対応するか否かを判定し、文字成分であると判定された基本成分の集合を文字成分の集合として抽出する。そして、文字列抽出手段 3 は、抽出された文字成分の集合から文字列に対応する文字成分を抽出する。

【 0 0 1 9 】

文字成分抽出手段 2 が用いる基本成分間の包含関係とは、文書画像内においてある基本成分が他の基本成分の内側に含まれるか、外側にあるか、他の基本成分と重なり合っているか等の 2 次元的な位置関係を表す。

【 0 0 2 0 】

文字成分抽出手段 2 は、例えば、所定数以上の基本成分を含む基本成分や、所定数以上の基本成分と重なり合っている基本成分や、何らかの方法により文字成分であると判定された基本成分を 1 つでも含むような基本成分は、文字成分ではないと判定する。

【 0 0 2 1 】

このように、基本成分の大きさ・形状、文字認識結果の確信度等だけでなく、基本成分間の包含関係も判定基準として用いることで、従来は正しく判定できなかった基本成分をも正しく判定することができる。したがって、文字成分抽出の精度ひいては文字列抽出の精度が向上する。

【 0 0 2 2 】

本発明の別の文字列抽出装置は、基本成分抽出手段 1、文字成分抽出手段 2、文字列抽出手段 3、および文字認識手段を備える。基本成分抽出手段 1 は、入力された文書画像から基本成分の集合を抽出する。文字認識手段は、基本成分抽出手段 1 により抽出された基本成分の文字認識を行う。文字成分抽出手段 2 は、基本成分の集合に含まれる基本成分間の包含関係を用いて基本成分が文字成分に対応するか否かを判定し、文字成分の集合を抽出する。このとき、文字成分抽出手段 2 は、所定数以上の基本成分を含む基本成分、および、所定数以上の基本成分と重なり合っている基本成分を、文字成分ではないと判定する。文字列抽出手段 3 は、文字成分の集合を用いて文字列を抽出する。

【発明を実施するための最良の形態】

【 0 0 2 3 】

以下、図面を参照しながら、本発明の実施の形態を詳細に説明する。

本実施形態においては、二値画像における黒画素連結成分の集合あるいはカラー画像における同一色画素の連結成分の集合のように、何らかの方法で得られた基本成分の集合に対して、基本成分間の包含関係を用いて文字成分であるか否かの判定を行う。そして、判定結果に基づいて文字成分の集合を抽出し、文字成分の集合から文字列を抽出する。

【 0 0 2 4 】

図 2 は、このような文字列抽出装置の構成図である。図 2 の文字列抽出装置は、基本成分抽出部 1 1、文字成分抽出部 1 2、および文字列抽出部 1 3 を備える。

基本成分抽出部 1 1 は、入力された文書画像から基本成分の集合を抽出する。文字成分抽出部 1 2 は、基本成分の集合を入力とし、基本成分間の包含関係を用いて各基本成分が文字成分であるか否かを判定し、文字成分を抽出する。また、文字列抽出部 1 3 は、例えば、文字成分の同質性あるいは近接性に基づいて、文字列に対応する文字成分の部分集合

10

20

30

40

50

を求める。

【 0 0 2 5 】

基本成分には、文字成分、図形成分等が含まれる。文字成分は、文字の一部または全部を表す画像パターンであり、図形成分は、図形、写真、表等の一部または全部を表す画像パターンである。

【 0 0 2 6 】

例えば、図 3 のような二値画像において、黒画素連結成分の外接矩形を基本成分として採用し、図形 2 1 の外接矩形 2 2 と、文字“あ”、“い”、“う”、“え”、“お”の外接矩形 2 3、2 4、2 5、2 6、2 7 が抽出されたとする。この場合、外接矩形 2 2 は図形成分に対応し、外接矩形 2 3、2 4、2 5、2 6、2 7 は文字成分に対応するが、基本成分抽出部 1 1 は、これらを区別せずに、ともに基本成分として抽出する。

10

【 0 0 2 7 】

また、基本成分間の包含関係とは、文書画像内においてある基本成分が他の基本成分の内側に含まれるか、外側にあるか、他の基本成分と重なり合っているか等の 2 次元的な位置関係を表す。

【 0 0 2 8 】

例えば、図 4 に示すように、基本成分 2 8 が多くの他の基本成分を内側に含んでいる場合、基本成分 2 8 は図形成分に対応する可能性が高い。そこで、このような場合、文字成分抽出部 1 2 は、基本成分 2 8 を文字成分ではないと判定する。

【 0 0 2 9 】

20

また、図 5 に示すように、基本成分 2 9 が多くの他の基本成分 3 0、3 1、3 2、3 3 と重なり合っている場合、基本成分 2 9 は図形成分に対応する可能性が高い。そこで、このような場合、文字成分抽出部 1 2 は、基本成分 2 9 を文字成分ではないと判定する。

【 0 0 3 0 】

また、図 3 に示したように、基本成分が文字成分を含む場合、その基本成分は図形成分に対応する可能性が高い。そこで、文字成分抽出部 1 2 は、あらかじめ何らかの方法で文字成分と判定された基本成分を 1 つ以上含む基本成分を、文字成分ではないと判定する。

【 0 0 3 1 】

図 6 は、このような文字成分抽出処理のフローチャートである。文字成分抽出部 1 2 は、まず、基本成分の集合から 1 つの基本成分を取り出し（ステップ S 1）、それが所定数以上の基本成分を含むか否かをチェックする（ステップ S 2）。そして、その基本成分が所定数以上の基本成分を含んでいれば、それを文字成分の候補から除外する（ステップ S 7）。

30

【 0 0 3 2 】

基本成分が所定数以上の基本成分を含んでいなければ、次に、それが所定数以上の基本成分と重なり合っているか否かをチェックする（ステップ S 3）。そして、その基本成分が所定数以上の基本成分と重なり合っていれば、それを文字成分の候補から除外する（ステップ S 7）。

【 0 0 3 3 】

基本成分が所定数以上の基本成分と重なり合っていなければ、次に、それが 1 つ以上の文字成分を含むか否かをチェックする（ステップ S 4）。そして、その基本成分が文字成分を含んでいれば、それを文字成分の候補から除外する（ステップ S 7）。

40

【 0 0 3 4 】

基本成分が文字成分の候補から除外された場合、あるいはステップ S 4 において基本成分が文字成分を含んでいない場合、判定対象となる次の基本成分があるか否かをチェックする（ステップ S 5）。次の基本成分があれば、その基本成分についてステップ S 1 以降の処理を繰り返す。

【 0 0 3 5 】

そして、次の基本成分がなくなると、文字成分の候補として残された基本成分の集合から、任意の方法により文字成分の集合を抽出して（ステップ S 6）、処理を終了する。

50

ステップS 4の判定を行うために、例えば、各基本成分をあらかじめ文字認識しておくことが考えられる。この場合、入力された基本成分は、高い信頼度で文字成分であると判定されたものとそうでないものとに分類される。

【0036】

図7は、このような文字成分抽出処理を行う構成を示している。文字成分抽出部12は、入力された基本成分の集合を文字認識部41に渡し、認識結果を受け取る。認識結果には、認識された文字の種類を表す情報と、認識処理の確信度とが含まれる。

【0037】

文字成分抽出部12は、所定のしきい値以上の確信度を持つ基本成分を文字成分であると判定し、判定結果を用いてステップS 4の判定を行う。このとき、ステップS 4の判定精度を高めるため、確信度のしきい値として比較的厳しい値を用いることにする。

10

【0038】

ところで、文字成分を含む基本成分が文字成分に対応する場合も有り得る。例えば、図8に示す文字“話”の画像の場合、“話”の外接矩形51は、“話”を構成する6つの部分の外接矩形52、53、54、55、56、57を含んでいる。このうち、外接矩形52～55は文字“一”に対応し、外接矩形56は文字“口”に対応し、外接矩形57は文字“舌”に対応する。この場合、ステップS 4の判定によれば、外接矩形51は文字成分ではないと判定されてしまう。

【0039】

そこで、基本成分が“一”、“口”等の単純な文字に対応する文字成分を含む場合、その基本成分を文字成分の候補から除外しない等の例外処理を付加することが望ましい。基本成分が単純な文字に対応するかどうかは、文字認識等によりチェックすることができる。あるいは、基本成分に含まれる文字成分の大きさに対する基本成分の大きさの比率を求め、それが所定値以上の場合に、基本成分を文字成分ではないと判定してもよい。

20

【0040】

また、あらかじめ文字成分の判定を行う方法としては、公開特許公報の特開昭61-026149（文書画像ファイル登録検索装置）に開示されている方法を用いることもできる。この方法によれば、黒画素数と黒ラン（連続して一列に配置された黒画素）の連結数を用いて、文字ストローク領域が識別される。

【0041】

ステップS 6の文字成分抽出においては、基本成分の大きさ・形状、あるいは文字認識結果の確信度等を判定基準として文字成分であるか否かが判定され、文字成分と判定されたものが次の文字列抽出処理に出力される。

30

【0042】

従来の文字成分抽出処理ではステップS 6の処理のみが行われていたが、本実施形態では、基本成分間の包含関係も判定基準として活用している。このため、従来は正しく判定できなかった基本成分をも正しく判定することができ、文字成分抽出の精度ひいては文字列抽出の精度を向上させることができる。

【0043】

こうして文字成分が抽出されると、文字列抽出部13は、例えば、先願の特願平10-146926（文書画像認識装置および文書画像認識プログラムの記憶媒体）に開示された方法を用いて、文字成分の集合から文字列を抽出する。この方法では、文字成分の大きさ・間隔が類似していることを表す同質性や、文字成分の大きさに比較して文字成分間の距離が小さいことを表す近接性に基づいて、文字列としての信頼度が評価される。さらに、文字成分の色等の他の性質に関する同質性を用いて、文字列としての信頼度を評価してもよい。

40

【0044】

このように、大きさ、間隔、色等が互いに類似している文字成分や、互いに近接している文字成分は、同じ文字列に属する文字を表しているものとみなされ、これらの文字成分の集合は1つの文字列として出力される。

50

【 0 0 4 5 】

また、文字成分抽出部 1 2 と文字列抽出部 1 3 が相互に作用することで、文字列抽出の精度をさらに向上させることも可能である。この場合、文字列抽出装置は、文字列認識の確信度が高い文字列に含まれる文字成分のみを真の文字成分であるとみなし、それ以外の文字列に含まれる文字成分を文字成分ではないと判定し直して、再度、文字列抽出を行う。

【 0 0 4 6 】

図 9 は、このような文字列抽出処理を行う構成を示している。文字成分抽出部 1 2 は、入力された基本成分の集合を文字認識部 4 1 に渡し、上述したような認識結果を受け取る。そして、所定のしきい値以上の確信度を持つ基本成分を文字成分であると判定し、判定結果を用いてステップ S 4 の判定を行う。

10

【 0 0 4 7 】

文字列抽出部 1 3 は、図 1 0 に示すような文字列抽出処理を行う。文字列抽出部 1 3 は、まず、文字成分抽出部 1 2 から文字成分の集合を受け取り、文字成分の同質性や近接性に基づいて文字列を抽出する（ステップ S 1 1）。そして、文字列抽出の結果得られた文字列集合を文字列認識部 6 1 に渡し、認識結果を受け取る（ステップ S 1 2）。

【 0 0 4 8 】

このとき、文字列認識部 6 1 は、文字認識、文字列認識等を行って、認識された文字列を表す情報と認識処理の確信度とを、認識結果として文字列抽出部 1 3 に返す。

次に、文字列抽出部 1 3 は、所定のしきい値以上の確信度を持つ文字列を抽出し、それらに含まれる文字成分を選択して、文字成分抽出部 1 2 に渡す（ステップ S 1 3）。

20

【 0 0 4 9 】

文字成分抽出部 1 2 は、例えば、文字列抽出部 1 3 から受け取った文字成分を真の文字成分であると判定し、それ以外の文字成分を文字成分ではないと判定する。そして、新たな文字成分の集合を生成して、再度、文字列抽出部 1 3 に出力する。これを受けて、文字列抽出部 1 3 は、新たな文字成分の集合から文字列を抽出し（ステップ S 1 4）、処理を終了する。

【 0 0 5 0 】

このように、文字成分抽出部 1 2 と文字列抽出部 1 3 が相互に作用して文字列抽出を繰り返すことにより、文字成分抽出および文字列抽出の精度が向上する。例えば、1 回目の文字列抽出において順位の低かった候補を、2 回目の文字列抽出において文字列と判定することができるようになる。ここでは、文字列抽出を 2 回行っているが、同様にして、これを 3 回以上繰り返してもよい。

30

【 0 0 5 1 】

次に、図 1 1 から図 1 6 までを参照しながら、図 2 の基本成分抽出部 1 1 の構成について説明する。

図 1 1 の基本成分抽出部は、連結成分抽出部 7 1 を含む。文書画像として二値画像が入力されたとき、連結成分抽出部 7 1 は、入力画像から黒画素連結成分を抽出し、それを基本成分として出力する。また、グレースケール文書のような多階調画像が入力されたとき、入力画像から画素レベルがほぼ同一である画素の連結成分を抽出し、それを基本成分として出力する。また、カラー文書のようなカラー画像が入力されたとき、入力画像から色がほぼ同一である画素の連結成分を抽出し、それを基本成分として出力する。

40

【 0 0 5 2 】

多階調画像の場合は、例えば、あらかじめ画素の階調レベルを複数の範囲に分類しておき、1 つの範囲に属する階調レベルを持つ隣接画素を連結して、画素連結成分を生成すればよい。また、カラー画像の場合は、例えば、あらかじめ画素の色情報（R G B の値）を複数の範囲に分類しておき、1 つの範囲に属する色情報を持つ隣接画素を連結して、画素連結成分を生成すればよい。

【 0 0 5 3 】

このような基本成分抽出部によれば、二値画像だけでなく、多階調画像あるいはカラー

50

画像からも基本成分を抽出することができ、これらの文書画像から文字列を抽出することができる。

【 0 0 5 4 】

図 1 2 の基本成分抽出部は、連結成分抽出部 7 1 と外接矩形生成部 7 2 を含む。連結成分抽出部 7 1 の処理については、図 1 1 の場合と同様である。外接矩形生成部 7 2 は、入力された画素連結成分に外接する矩形を生成し、それを基本成分として出力する。

【 0 0 5 5 】

外接矩形の形状は画素連結成分の形状より簡略化されているため、基本成分間の包含関係のチェックがより簡単になり、文字成分抽出、文字列抽出等の後処理が高速化される。外接矩形の代わりに、他の多角形、円、楕円等の任意の外接図形を用いることもできる。

10

【 0 0 5 6 】

図 1 3 の基本成分抽出部は、連結成分抽出部 7 1、外接矩形生成部 7 2、および二値画像生成部 7 3 を含む。連結成分抽出部 7 1 と外接矩形生成部 7 2 の処理については、図 1 2 の場合と同様である。二値画像生成部 7 3 は、入力された外接矩形に含まれる多階調画像あるいはカラー画像を二値化して二値画像を生成し、連結成分抽出部 7 1 に出力する。

【 0 0 5 7 】

二値画像生成部 7 3 は、例えば、先願の特願平 1 0 - 3 5 3 0 4 5 (カラー文書画像認識装置) に開示された方法を用いて、多階調画像あるいはカラー画像から二値画像を生成する。この方法では、各画素の明度成分が所定のしきい値で二値化され、描画領域に対応する値と背景領域に対応する値のいずれか一方を持つ画素から構成される二値画像が生成

20

【 0 0 5 8 】

連結成分抽出部 7 1 は、入力された二値画像から同じ値を持つ画素の連結成分を抽出し、それを基本成分として出力する。あるいは、外接矩形生成部 7 2 が、得られた画素連結成分の外接矩形を生成し、それを基本成分として出力することもできる。

【 0 0 5 9 】

このような基本成分抽出部によれば、多階調画像あるいはカラー画像の基本成分に対応する部分が二値化された後に基本成分が抽出されるため、より精密に基本成分を抽出することができる。

【 0 0 6 0 】

30

図 1 4 の基本成分抽出部は、二値画像生成部 7 4 と連結成分抽出部 7 5 を含む。二値画像生成部 7 4 は、入力された多階調画像あるいはカラー画像からエッジ二値画像を生成し、連結成分抽出部 7 5 は、エッジ二値画像から同じ値を持つ画素の連結成分 (エッジ連結成分) を抽出し、それを基本成分として出力する。

【 0 0 6 1 】

二値画像生成部 7 4 は、例えば、上述の特願平 1 0 - 3 5 3 0 4 5 に開示された方法を用いて、多階調画像あるいはカラー画像からエッジ二値画像を生成する。この方法では、エッジ抽出処理によりエッジ強度画像あるいはエッジ方向画像が生成され、得られた画像が所定のしきい値で二値化されて、エッジ二値画像が生成される。エッジ抽出処理には、ソーベルフィルタ、ラプラシアンフィルタ等が用いられる。

40

【 0 0 6 2 】

このような基本成分抽出部によれば、エッジ抽出により多階調画像あるいはカラー画像の描画領域の輪郭が抽出されるため、より精密に基本成分を抽出することができる。

図 1 5 の基本成分抽出部は、二値画像生成部 7 4、連結成分抽出部 7 5、および外接矩形生成部 7 2 を含む。二値画像生成部 7 4 と連結成分抽出部 7 5 の処理については、図 1 4 の場合と同様である。外接矩形生成部 7 2 は、入力されたエッジ連結成分に外接する矩形を生成し、それを基本成分として出力する。

【 0 0 6 3 】

図 1 6 の基本成分抽出部は、二値画像生成部 7 4、連結成分抽出部 7 5、外接矩形生成部 7 2、および二値画像生成部 7 3 を含む。二値画像生成部 7 4、連結成分抽出部 7 5、

50

および外接矩形生成部 7 2 の処理については、図 1 5 の場合と同様である。二値画像生成部 7 3 は、入力された外接矩形に含まれる多階調画像あるいはカラー画像を二値化して二値画像を生成し、連結成分抽出部 7 5 に出力する。

【 0 0 6 4 】

連結成分抽出部 7 1 は、入力された二値画像から同じ値を持つ画素の連結成分を抽出し、それを基本成分として出力する。あるいは、外接矩形生成部 7 2 が、得られた画素連結成分の外接矩形を生成し、それを基本成分として出力することもできる。

【 0 0 6 5 】

このように、図 1 1 から図 1 6 に示した基本成分抽出部によれば、文書画像として多階調画像あるいはカラー画像が入力された場合でも、基本成分を抽出することができ、それ

10

【 0 0 6 6 】

特に、図 1 6 の基本成分抽出部と図 7 の構成を組み合わせれば、二値化により得られる精密な基本成分に対して文字認識が行われ、認識結果に基づいて基本成分間の包含関係がチェックされるため、より高精度な文字成分抽出および文字列抽出が可能となる。また、図 1 6 の基本成分抽出部と図 9 の構成を組み合わせれば、文字成分抽出部と文字列抽出部の相互作用により、さらに高精度な文字成分抽出および文字列抽出が可能となる。

【 0 0 6 7 】

ところで、上述した文字列抽出装置は、図 1 7 に示すような情報処理装置（コンピュータ）を用いて構成することができる。図 1 7 の情報処理装置は、C P U（中央処理装置）

20

【 0 0 6 8 】

メモリ 8 2 は、例えば、R O M（read only memory）、R A M（random access memory）等を含み、処理に用いられるプログラムとデータを格納する。C P U 8 1 は、メモリ 8 2 を利用してプログラムを実行することにより、必要な処理を行う。

【 0 0 6 9 】

この場合、図 2 の基本成分抽出部 1 1、文字成分抽出部 1 2、文字列抽出部 1 3、図 7 の文字認識部 4 1、および図 9 の文字列認識部 6 1 は、メモリ 8 2 に格納されたプログラ

30

【 0 0 7 0 】

入力装置 8 3 は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、ユーザからの指示や情報の入力に用いられる。出力装置 8 4 は、例えば、ディスプレイ、プリンタ、スピーカ等であり、ユーザへの問い合わせや処理結果の出力に用いられる。

【 0 0 7 1 】

外部記憶装置 8 5 は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク（magneto-optical disk）装置等である。情報処理装置は、この外部記憶装置 8 5 に、上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ 8 2 にロードして

40

【 0 0 7 2 】

媒体駆動装置 8 6 は、可搬記録媒体 8 9 を駆動し、その記録内容にアクセスする。可搬記録媒体 8 9 としては、メモ리카ード、フロッピーディスク、C D - R O M（compact disk read only memory）、光ディスク、光磁気ディスク等、任意のコンピュータ読み取り可能な記録媒体が用いられる。ユーザは、この可搬記録媒体 8 9 に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ 8 2 にロードして使用することができる。

【 0 0 7 3 】

ネットワーク接続装置 8 7 は、L A N（local area network）等の任意のネットワーク

50

(回線)を介して外部の装置と通信し、通信に伴うデータ変換を行う。情報処理装置は、必要に応じて、ネットワーク接続装置 87 を介して上述のプログラムとデータを外部の装置から受け取り、それらをメモリ 82 にロードして使用することができる。

【0074】

図 18 は、図 17 の情報処理装置にプログラムとデータを供給することのできるコンピュータ読み取り可能な記録媒体を示している。可搬記録媒体 89 や外部のデータベース 90 に保存されたプログラムとデータは、メモリ 82 にロードされる。そして、CPU 81 は、そのデータを用いてそのプログラムを実行し、必要な処理を行う。

【産業上の利用可能性】

【0075】

10

本発明によれば、文書画像の基本成分が文字成分であるか否かの判定において、従来は正しく判定できなかった基本成分をも正しく判定することができ、文字成分抽出の精度ひいては文字列抽出の精度を向上させることができる。

【図面の簡単な説明】

【0076】

【図 1】本発明の文字列抽出装置の原理図である。

【図 2】文字列抽出装置の構成図である。

【図 3】文字成分と図形成分を示す図である。

【図 4】基本成分を含む基本成分を示す図である。

【図 5】互いに重なり合う基本成分を示す図である。

20

【図 6】文字成分抽出処理のフローチャートである。

【図 7】文字認識に基づく文字成分抽出を示す図である。

【図 8】文字成分を含む文字成分を示す図である。

【図 9】文字列認識に基づく文字列抽出を示す図である。

【図 10】文字列抽出処理のフローチャートである。

【図 11】第 1 の基本成分抽出部を示す図である。

【図 12】第 2 の基本成分抽出部を示す図である。

【図 13】第 3 の基本成分抽出部を示す図である。

【図 14】第 4 の基本成分抽出部を示す図である。

【図 15】第 5 の基本成分抽出部を示す図である。

30

【図 16】第 6 の基本成分抽出部を示す図である。

【図 17】情報処理装置の構成図である。

【図 18】記録媒体を示す図である。

【図 1】

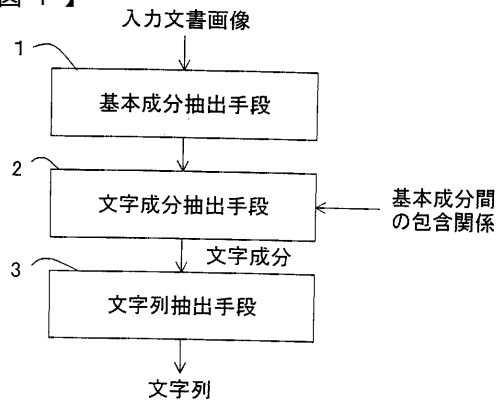


図 1

【図 2】

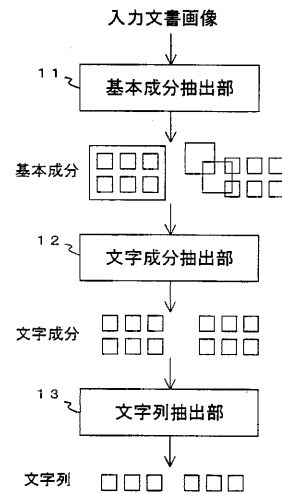


図 2

【図 3】

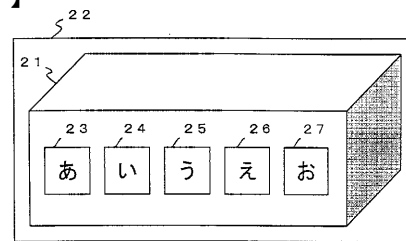


図 3

【図 4】

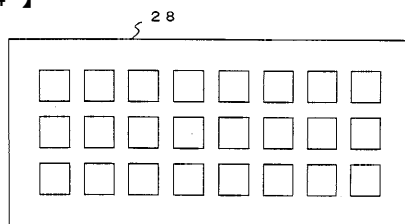


図 4

【図 5】

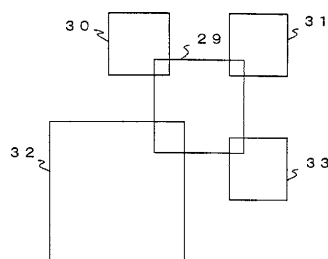


図 5

【図 6】

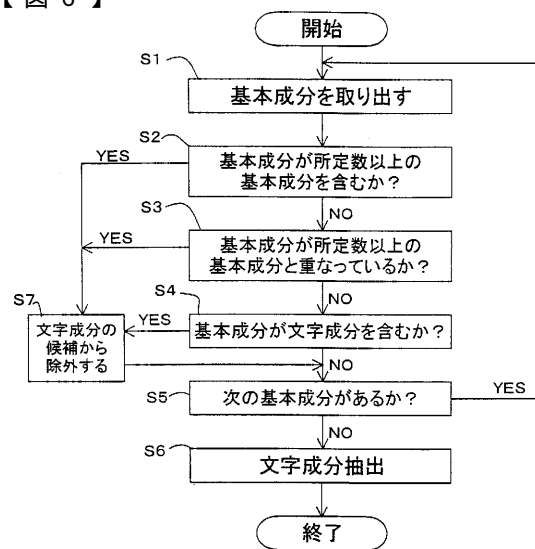


図 6

【図 7】

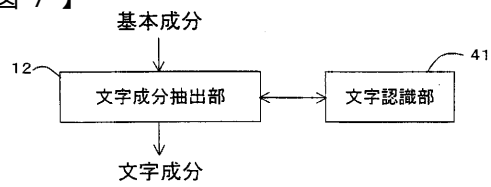


図 7

【図 8】

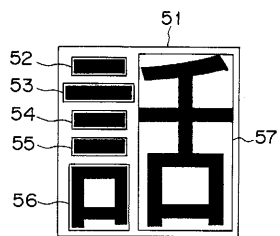


図 8

【図 9】

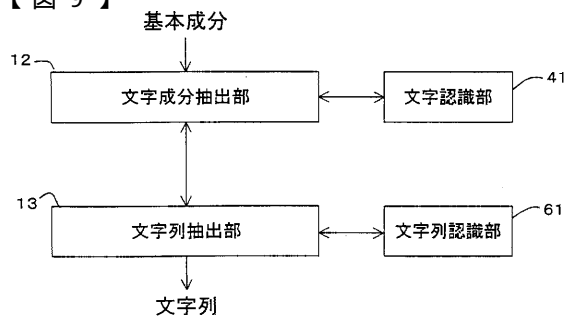


図 9

【図 10】

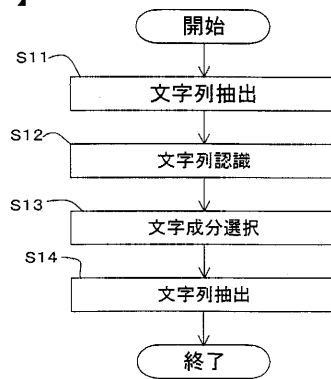


図 10

【図 11】

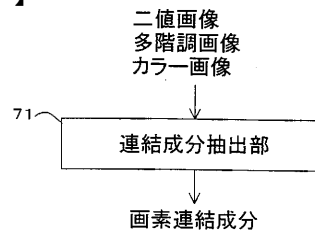


図 11

【図 1 2】

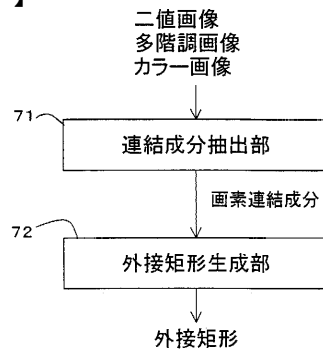


図12

【図 1 3】

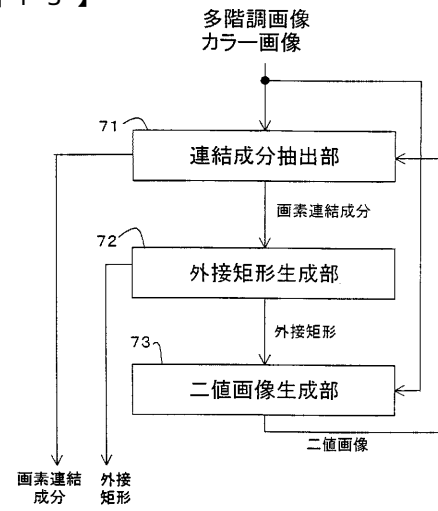


図13

【図 1 4】

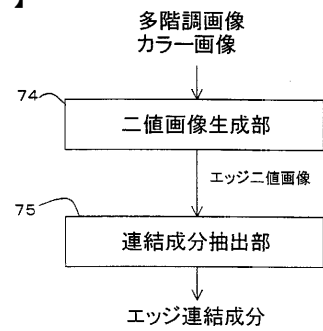


図14

【図 1 5】

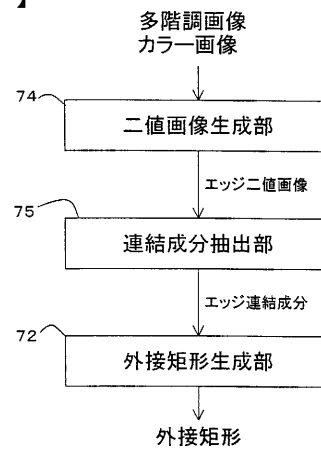


図15

【図 16】

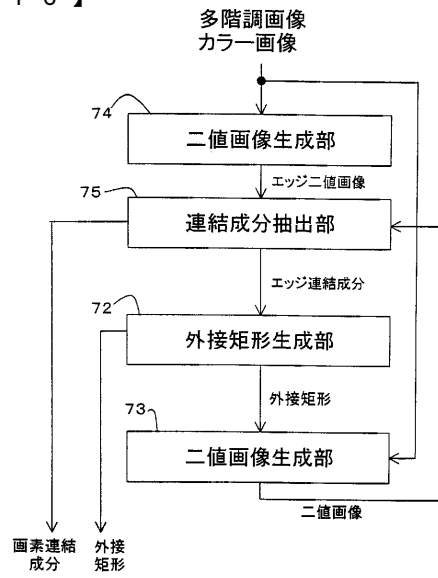


図16

【図 17】

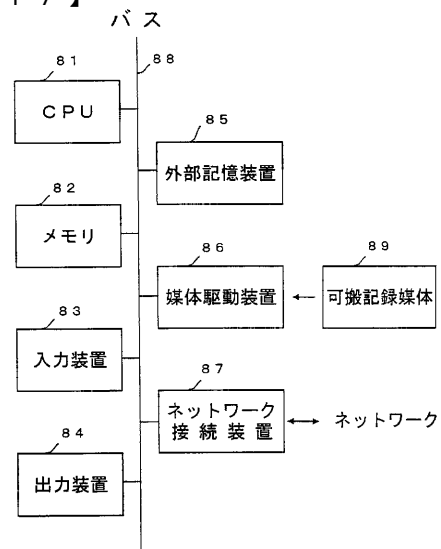


図 17

【図 18】

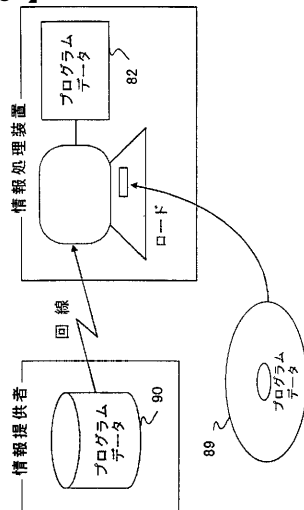


図 18

フロントページの続き

審査官 西山 昇

- (56)参考文献 特開平09 - 016713 (JP, A)
特開平07 - 168911 (JP, A)
特開平05 - 166002 (JP, A)
特開平09 - 167233 (JP, A)
特開平08 - 055188 (JP, A)
特開平09 - 081743 (JP, A)
特開平05 - 028305 (JP, A)