世界知的所有権機関
国際事務局

PCT

特許協力条約に基づいて公開された国際出願

(51) 国際特許分類6 G06K 9/20	A1	(11) 国際公開番号	WO00/62243
		(43) 国際公開日	2000年10月19日(19.10.00)

(21) 国際出願番号 PCT/JP99/01986

(22) 国際出願日 1999年4月14日(14.04.99)

(71) 出願人（米国を除くすべての指定国について）
富士通株式会社(FUJITSU LIMITED)[JP/JP]
〒211-8588 神奈川県川崎市中原区上小田中4丁目1番1号
Kanagawa, (JP)
(72) 発明者；および
(75) 発明者／出願人（米国についてのみ）
藤本克仁(FUJIMOTO, Katsuhito)[JP/JP]
鎌田　洋(KAMADA, Hiroshi)[JP/JP]
〒211-8588 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 Kanagawa, (JP)
(74) 代理人
大菅義之(OSUGA, Yoshiyuki)
〒102-0084 東京都千代田区二番町8番地20
二番町ビル3F Tokyo, (JP)

(81) 指定国 JP, US

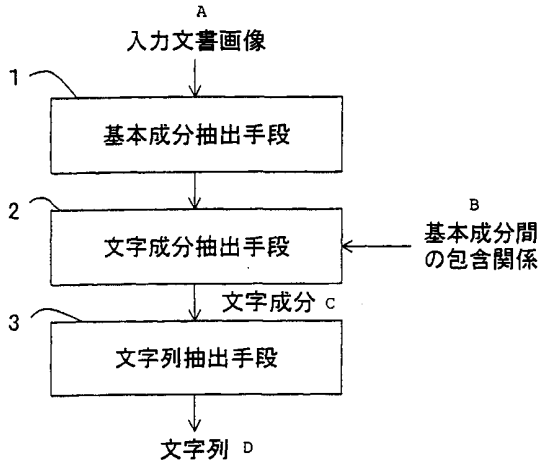
添付公開書類
国際調査報告書

(54)Title: CHARACTER STRING EXTRACTING DEVICE AND METHOD BASED ON BASIC COMPONENT IN DOCUMENT IMAGE

(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) Abstract

A character string extracting device extracts a set of basic components from a document image such as a binary image, a multi-level gradation image, or a color image, judges whether or not each basic component is a character component according to the inclusion relation between the basic components, extracts a set of character components based on the judgment result, and extracts a character string from the set of character components.



1 ... BASIC COMPONENT EXTRACTING MEANS
2 ... CHARACTER COMPONENT EXTRACTING MEANS
3 ... CHARACTER STRING EXTRACTING MEANS
A ... INPUT DOCUMENT IMAGE
B ... INCLUSION RELATION BETWEEN BASIC COMPONENTS
C ... CHARACTER COMPONENT
D ... CHARACTER STRING

(57)要約

文字列抽出装置は、二値画像、多階調画像、カラー画像等の文書画像から基本成分の集合を抽出し、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。

明 細 書

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

技術分野

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

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。

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。

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。

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、文書構造の先見的知識を必要とせず、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。

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。

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姓を用いて、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。

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。

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。

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。

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。

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。

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。

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。

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。

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。

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。

発明の開示

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。

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。

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。

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。

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。

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。

図面の簡単な説明

図１は、本発明の文字列抽出装置の原理図である。

図２は、文字列抽出装置の構成図である。

図３は、文字成分と図形成分を示す図である。

図４は、基本成分を含む基本成分を示す図である。

図５は、互いに重なり合う基本成分を示す図である。

図６は、文字成分抽出処理のフローチャートである。

図７は、文字認識に基づく文字成分抽出を示す図である。

図８は、文字成分を含む文字成分を示す図である。

図９は、文字列認識に基づく文字列抽出を示す図である。

図１０は、文字列抽出処理のフローチャートである。

図１１は、第１の基本成分抽出部を示す図である。

図１２は、第２の基本成分抽出部を示す図である。

図１３は、第３の基本成分抽出部を示す図である。

図１４は、第４の基本成分抽出部を示す図である。

図１５は、第５の基本成分抽出部を示す図である。

図１６は、第６の基本成分抽出部を示す図である。

図１７は、情報処理装置の構成図である。

図１８は、記録媒体を示す図である。

発明を実施するための最良の形態

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。

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１１、文字成分抽出部１２、および文字列抽出部１３を備える。

基本成分抽出部１１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１２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１３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を求める。

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。

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２１の外接矩形２２と、文字“あ”、“い”、“う”、“え”、“お”の外接矩形２３、２４、２５、２６、２７が抽出されたとする。この場合、外接矩形２２は図形成分に対応し、外接矩形２３、２４、２５、２

６、２７は文字成分に対応するが、基本成分抽出部１１は、これらを区別せずに、ともに基本成分として抽出する。

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。

例えば、図４に示すように、基本成分２８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２８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１２は、基本成分２８を文字成分ではないと判定する。

また、図５に示すように、基本成分２９が多くの他の基本成分３０、３１、３２、３３と重なり合っている場合、基本成分２９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１２は、基本成分２９を文字成分ではないと判定する。

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１２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。

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１２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Ｓ１）、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Ｓ２）。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Ｓ７）。

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Ｓ３）。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Ｓ７）。

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Ｓ４）。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Ｓ７）。

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Ｓ４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Ｓ５）。次の基本成分があれば、その基本成分についてステップＳ１以降の処理を繰り返す。

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Ｓ６）、処理を終了する。

ステップＳ４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。

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。

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Ｓ４の判定を行う。このとき、ステップＳ４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。

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Ｓ４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。

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。

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。

ステップＳ６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。

従来の文字成分抽出処理ではステップＳ６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。

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。

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。

また、文字成分抽出部１２と文字列抽出部１３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。

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１２は、入力された基本成分の集合を文字認識部４１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Ｓ４の判定を行う。

文字列抽出部１３は、図１０に示すような文字列抽出処理を行う。文字列抽出部１３は、まず、文字成分抽出部１２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Ｓ１１）。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６１に渡し、認識結果を受け取る（ステップＳ１２）。

このとき、文字列認識部６１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１３に返す。

次に、文字列抽出部１３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１２に渡す（ステップＳ１３）。

文字成分抽出部１２は、例えば、文字列抽出部１３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１３に出力する。これを受けて、文字列抽出部１３は、新たな文字成分の集合から文字列を抽出し（ステップＳ１４）、処理を終了する。

このように、文字成分抽出部１２と文字列抽出部１３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。

次に、図１１から図１６までを参照しながら、図２の基本成分抽出部１１の構成について説明する。

図１１の基本成分抽出部は、連結成分抽出部７１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７１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。

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（ＲＧＢ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。

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。

図１２の基本成分抽出部は、連結成分抽出部７１と外接矩形生成部７２を含む。連結成分抽出部７１の処理については、図１１の場合と同様である。外接矩形生成部７２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。

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。

図１３の基本成分抽出部は、連結成分抽出部７１、外接矩形生成部７２、および二値画像生成部７３を含む。連結成分抽出部７１と外接矩形生成部７２の処理については、図１２の場合と同様である。二値画像生成部７３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７１に出力する。

二値画像生成部７３は、例えば、先願の特願平１０－３５３０４５（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される。

連結成分抽出部７１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７２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。

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。

図１４の基本成分抽出部は、二値画像生成部７４と連結成分抽出部７５を含む。二値画像生成部７４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７５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。

二値画像生成部７４は、例えば、上述の特願平１０－３５３０４５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。

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１５の基本成分抽出部は、二値画像生成部７４、連結成分抽出部７５、および外接矩形生成部７２を含む。二値画像生成部７４と連結成分抽出部７５の処理については、図１４の場合と同様である。外接矩形生成部７２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。

図１６の基本成分抽出部は、二値画像生成部７４、連結成分抽出部７５、外接矩形生成部７２、および二値画像生成部７３を含む。二値画像生成部７４、連結成分抽出部７５、および外接矩形生成部７２の処理については、図１５の

場合と同様である。二値画像生成部７３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７５に出力する。

連結成分抽出部７１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７２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。

このように、図１１から図１６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、それに基づいて文字成分および文字列を抽出することができる。

特に、図１６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１６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。

ところで、上述した文字列抽出装置は、図１７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１７の情報処理装置は、ＣＰＵ（中央処理装置）８１、メモリ８８、入力装置８３、出力装置８４、外部記憶装置８５、媒体駆動装置８６、およびネットワーク接続装置８７を備え、それらはバス８８により互いに接続されている。

メモリ８２は、例えば、ＲＯＭ（read only memory）、ＲＡＭ（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。ＣＰＵ８１は、メモリ８２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。

この場合、図２の基本成分抽出部１１、文字成分抽出部１２、文字列抽出部１３、図７の文字認識部４１、および図９の文字列認識部６１は、メモリ８２に格納されたプログラムに対応するソフトウエアコンポーネントとして実装される。

入力装置８３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８４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。

外部記憶装置８５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８５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８２にロードして使用することができる。

媒体駆動装置８６は、可搬記録媒体８９を駆動し、その記録内容にアクセスする。可搬記録媒体８９としては、メモリカード、フロッピーディスク、ＣＤ－ＲＯＭ（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８９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８２にロードして使用することができる。

ネットワーク接続装置８７は、ＬＡＮ（local area network）等の任意のネットワーク（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。

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、ＣＰＵ８１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。

5

産業上の利用可能性

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。

請 求 の 範 囲

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

を備えることを特徴とする文字列抽出装置。

2． 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を、文字成分ではないと判定することを特徴とする請求項1記載の文字列抽出装置。

3． 前記文字成分抽出手段は、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請求項1記載の文字列抽出装置。

4． 前記文字成分抽出手段は、文字成分であると判定された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ことを特徴とする請求項1記載の文字列抽出装置。

5． 前記基本成分抽出手段により抽出された基本成分の文字認識を行う文字認識手段をさらに備え、前記文字成分抽出手段は、該文字認識の結果に基づいて文字成分であると判定された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ことを特徴とする請求項4記載の文字列抽出装置。

6． 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う文字列認識手段をさらに備え、前記文字成分抽出手段は、該認識により得られた高い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、他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。

７． 前記文字成分抽出手段と文字列抽出手段が相互に作用して、文字列抽出を複数回繰り返すことを特徴とする請求項１記載の文字列抽出装置。

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１記載の文字列抽出装置。

９． 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１記載の文字列抽出装置。

１０． 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１記載の文字列抽出装置。

１１． 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１記載の文字列抽出装置。

１２． 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の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１記載の文字列抽出装置。

１３．　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の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１記載の文字列抽出装置。

１４．　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に含まれる文字成分の同質性および近接性のうち少なくとも一方に基づいて文字成分の部分集合を求め、該文字成分の部分集合を文字列として抽出する文字列抽出手段と

を備えることを特徴とする文字列抽出装置。

１５．　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、

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

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。

１６．　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、

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する

ことを特徴とする文字列抽出方法。

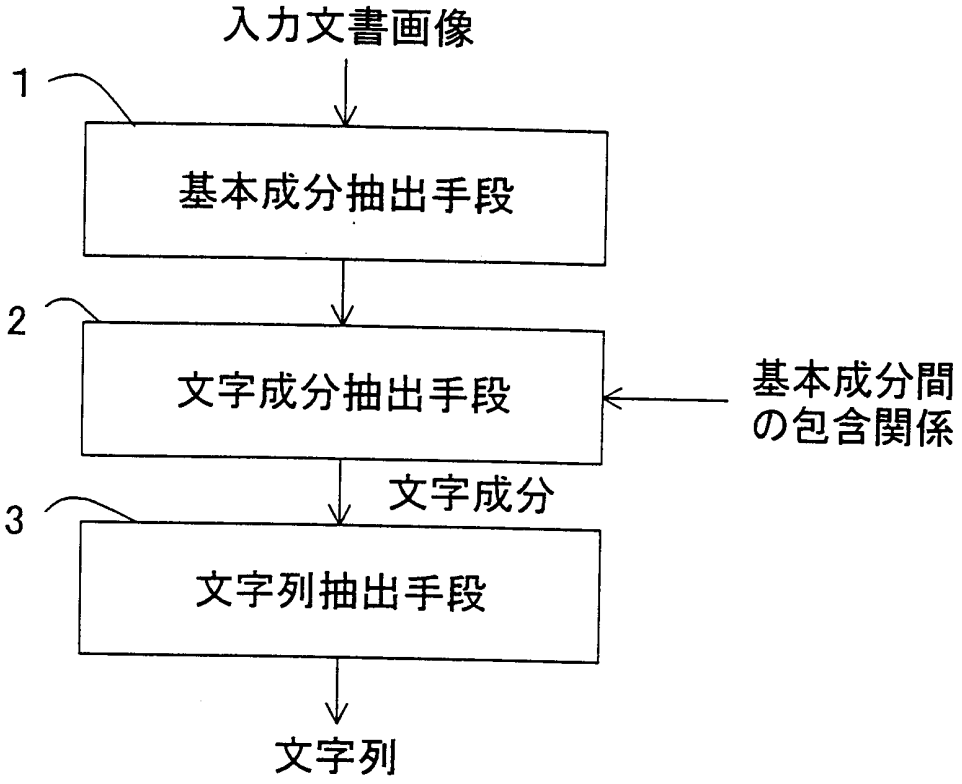


図1

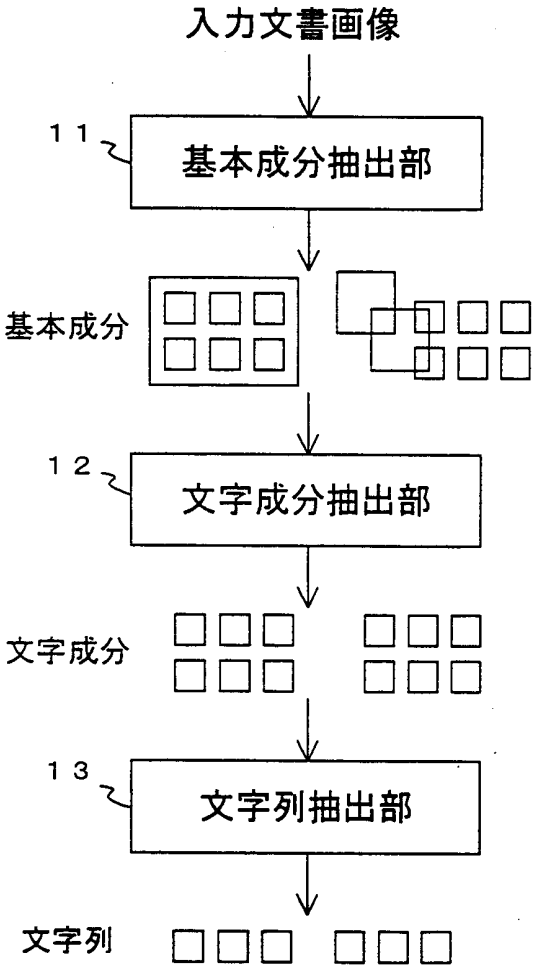


図2

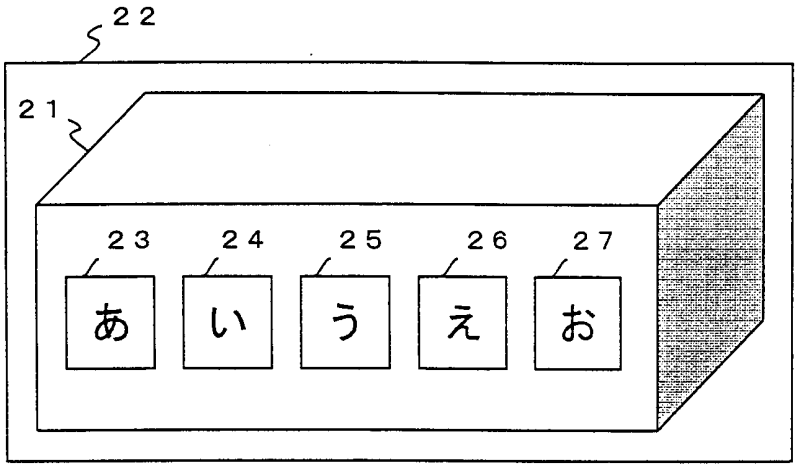


図３

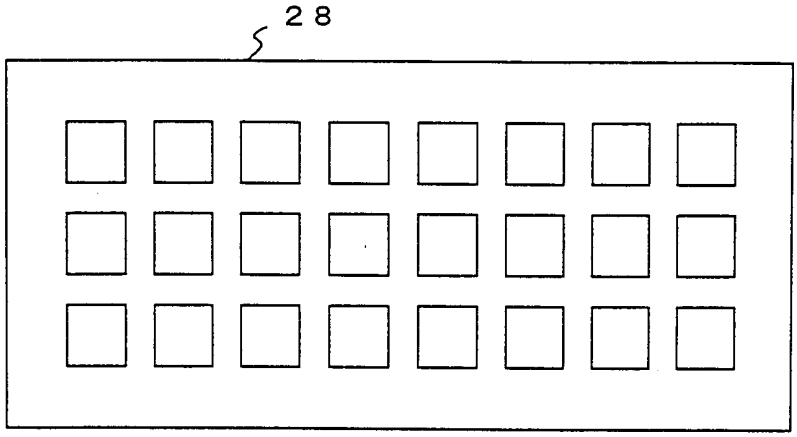


図4

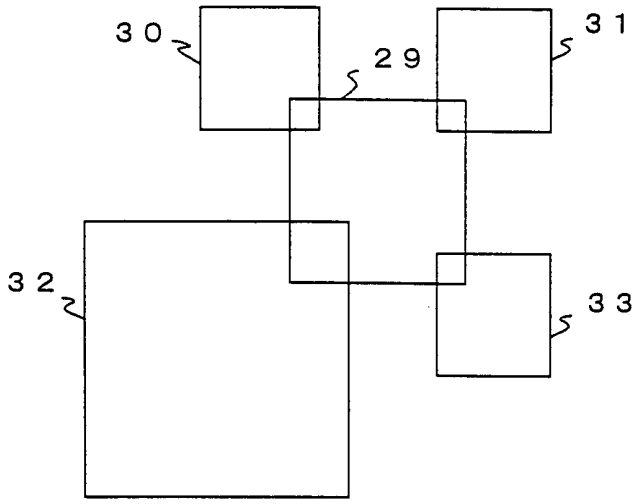


図5

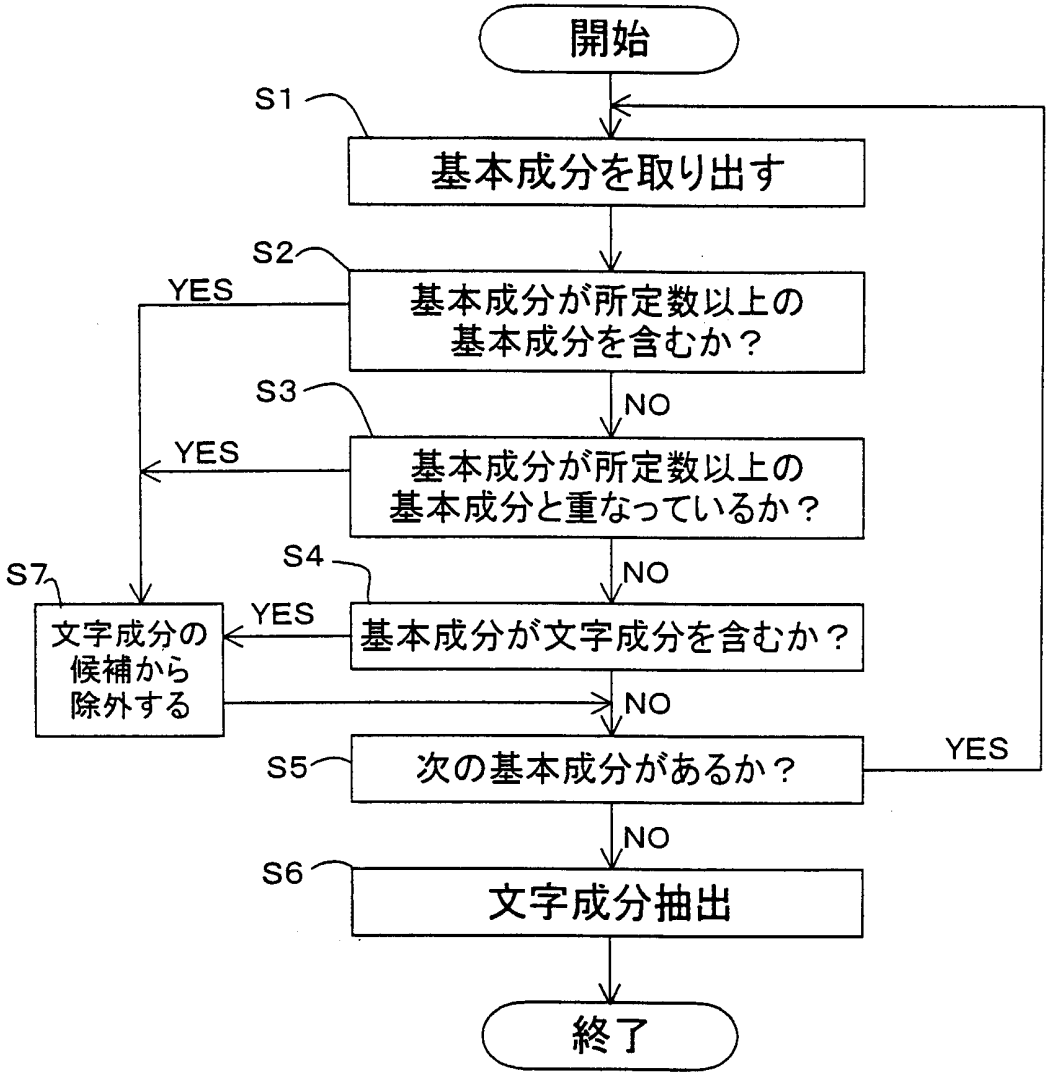

開始
S1
基本成分を取り出す
S2
基本成分が所定数以上の
基本成分を含むか？
YES
NO
S3
基本成分が所定数以上の
基本成分と重なっているか？
YES
NO
S4
基本成分が文字成分を含むか？
YES
NO
S7
文字成分の
候補から
除外する
S5
次の基本成分があるか？
YES
NO
S6
文字成分抽出
終了

図6

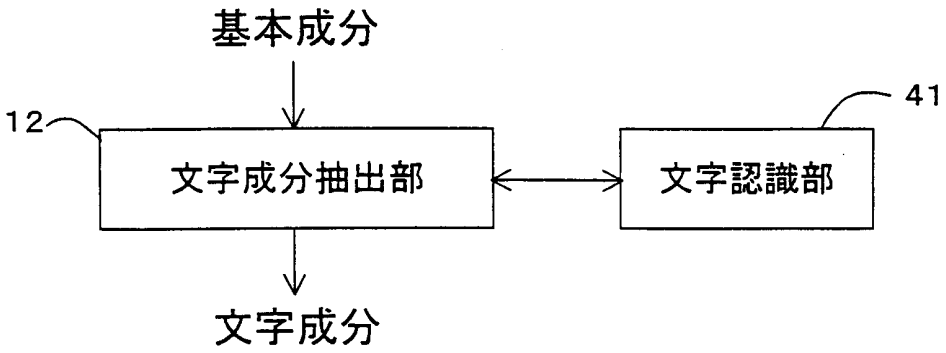


図7

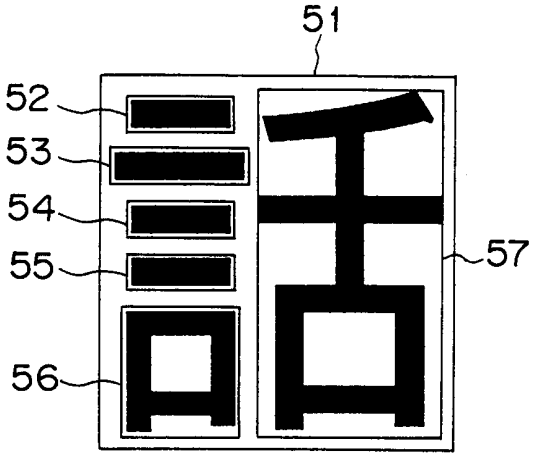


図 8

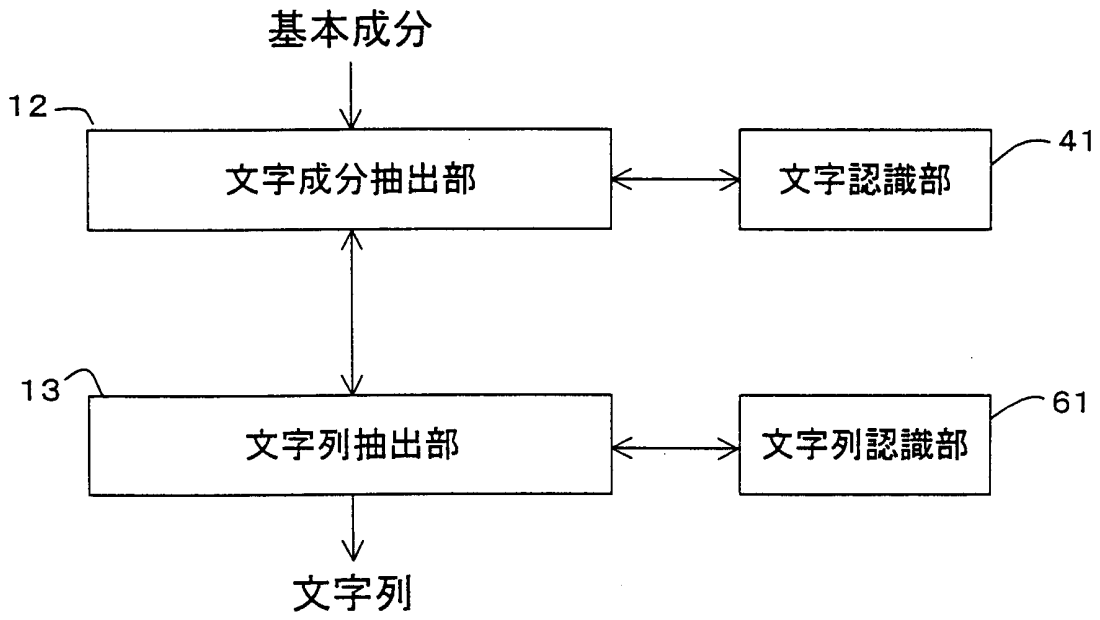


図9

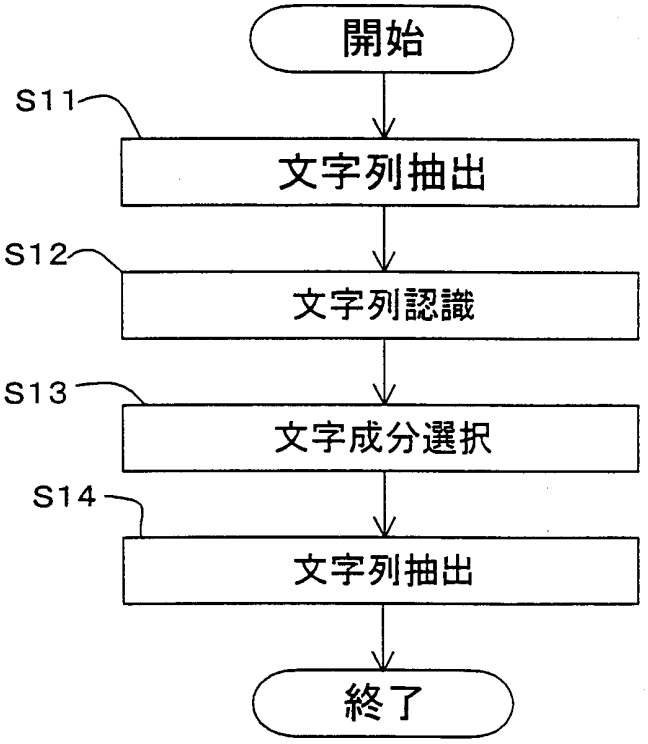


図10

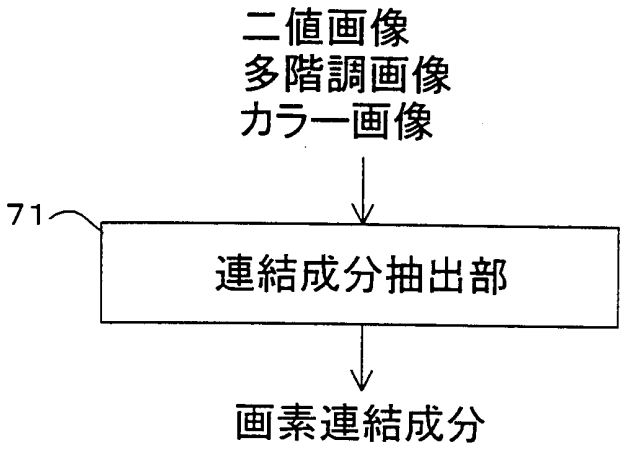


図11

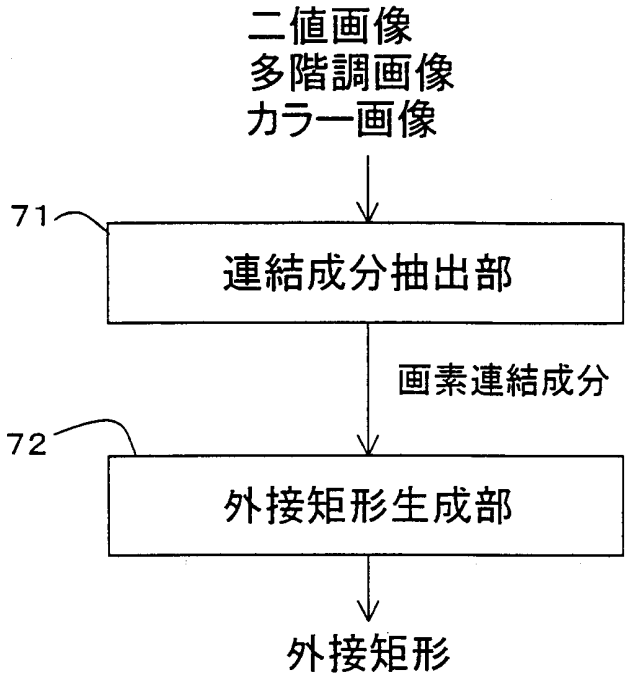


図12

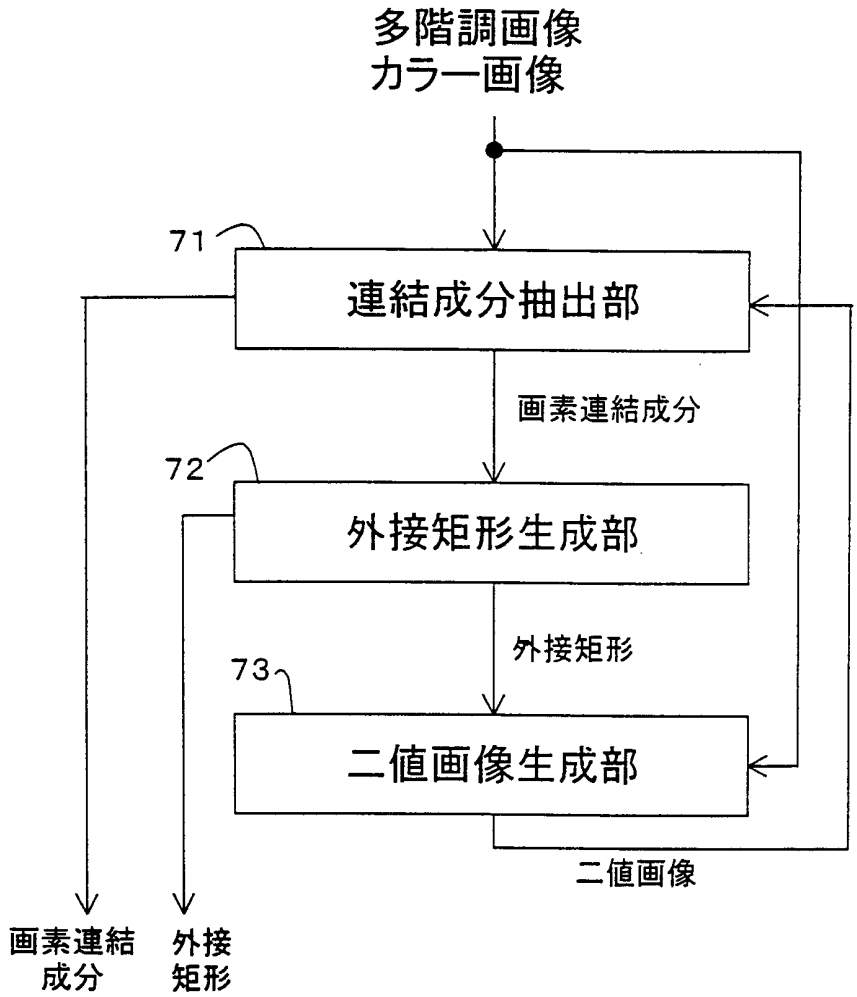

多階調画像
カラー画像
71
連結成分抽出部
画素連結成分
72
外接矩形生成部
外接矩形
73
二値画像生成部
二値画像
画素連結成分
外接矩形

図13

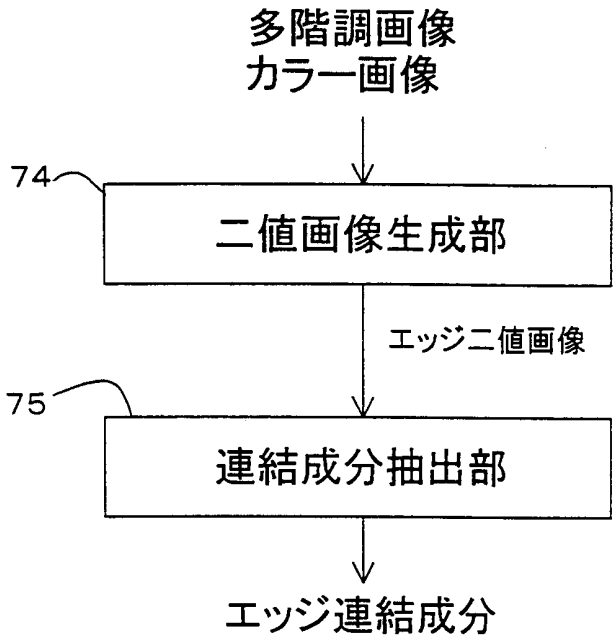


図14

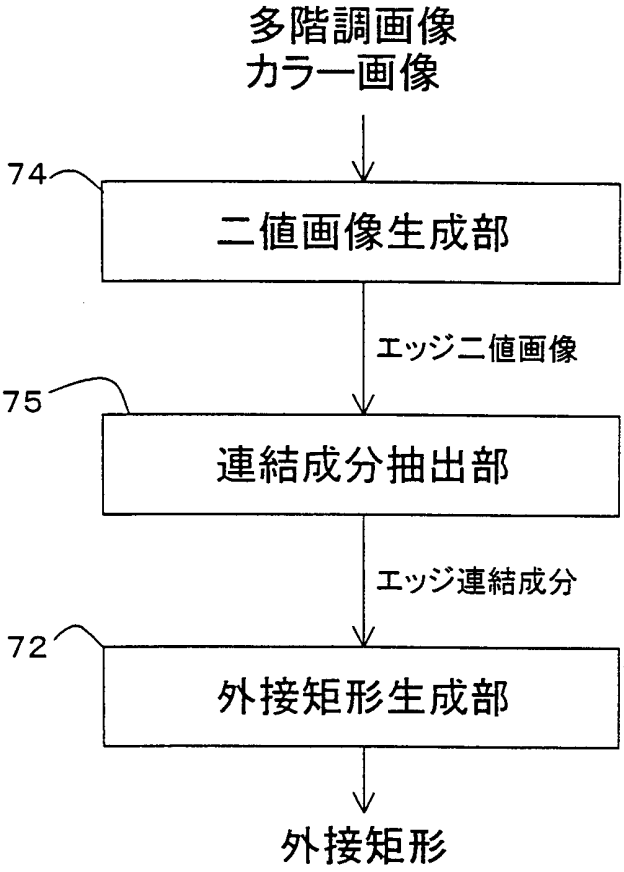

多階調画像
カラー画像
74
二値画像生成部
エッジ二値画像
75
連結成分抽出部
エッジ連結成分
72
外接矩形生成部
外接矩形

図15

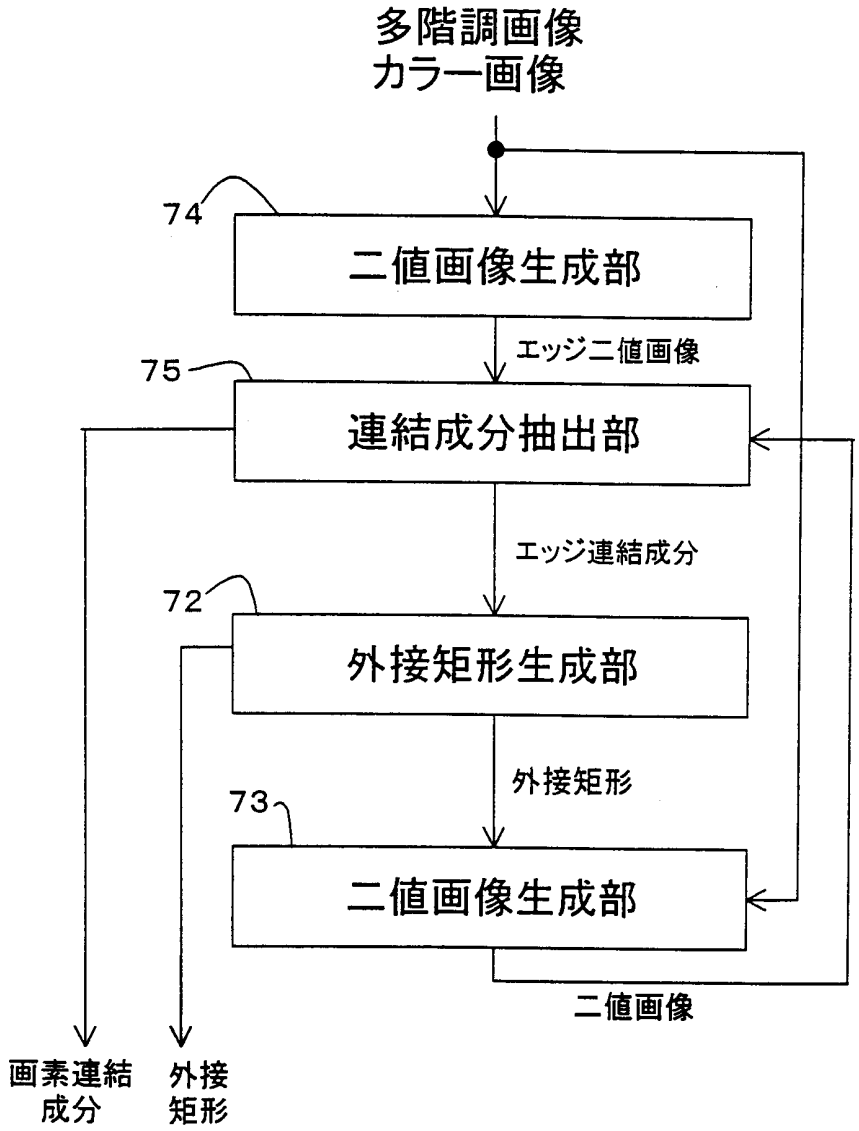

多階調画像
カラー画像
74
二値画像生成部
エッジ二値画像
75
連結成分抽出部
エッジ連結成分
72
外接矩形生成部
外接矩形
73
二値画像生成部
二値画像
画素連結成分
外接矩形

図16

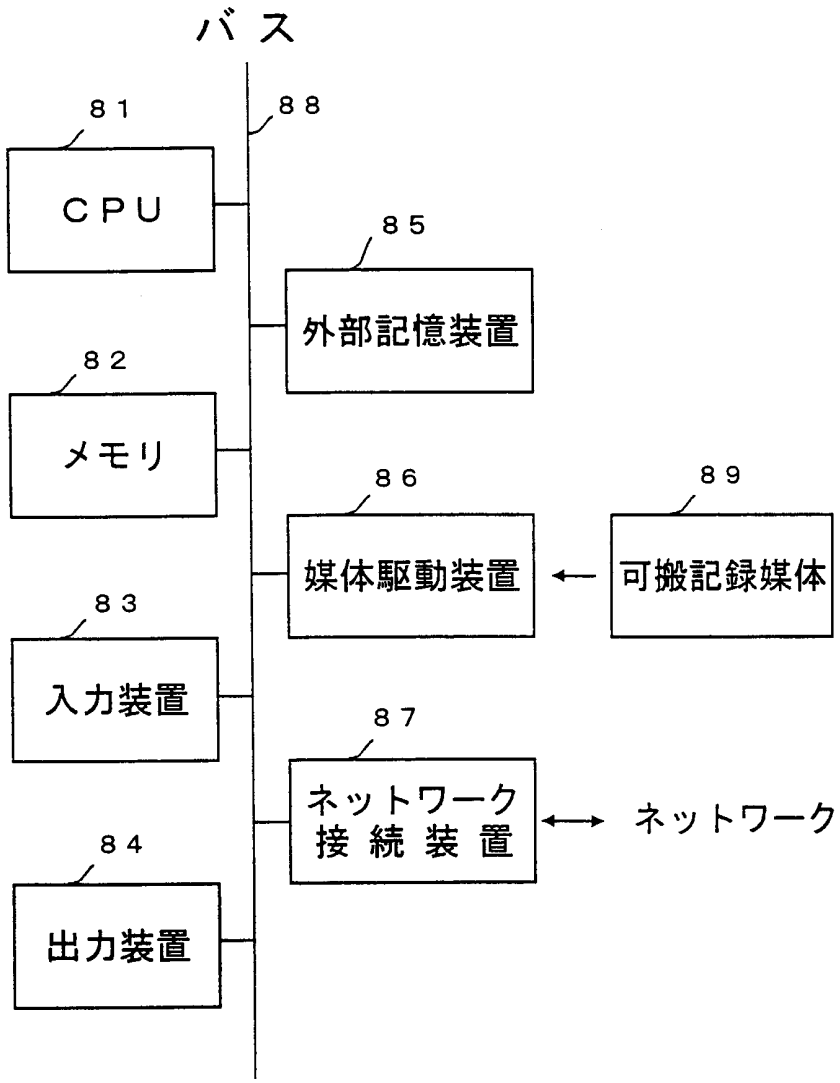

バス
88
81
CPU
85
外部記憶装置
82
メモリ
86
89
媒体駆動装置
可搬記録媒体
83
入力装置
87
ネットワーク
接続装置
ネットワーク
84
出力装置

図１７

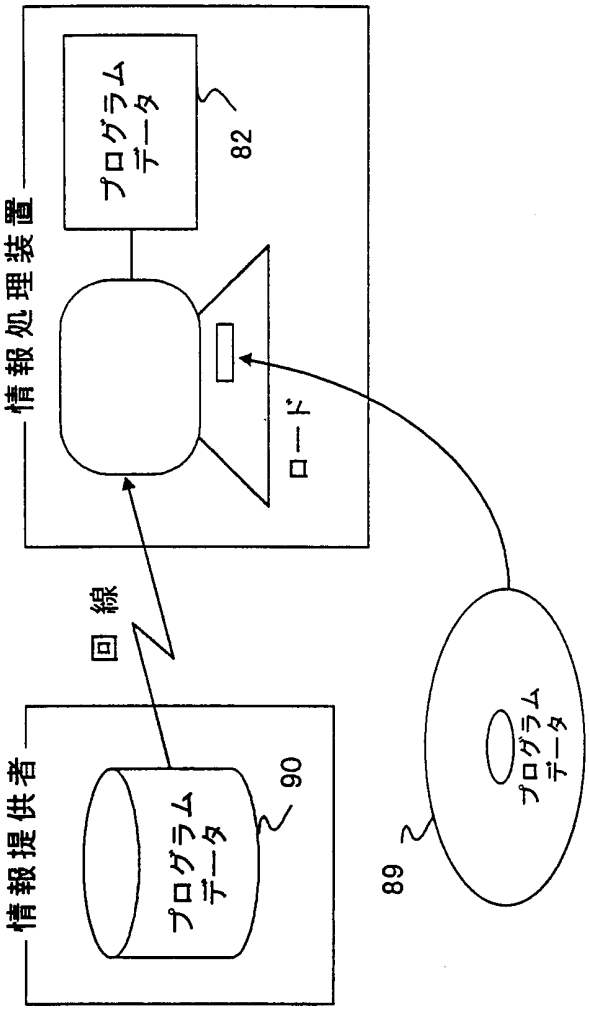


図18

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP99/01986

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[6] G06K9/20

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[6] G06K9/20

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	JP, 9-16713, A (Sharp Corp.), 17 January, 1997 (17. 01. 97) (Family: none)	1, 4, 15, 16, 5, 8-14
X Y	JP, 5-166002, A (Scitex Corp. Ltd.), 2 July, 1993 (02. 07. 93) (Family: none)	1, 15, 16 8-14
Y	JP, 7-168911, A (Matsushita Electric Industrial Co., Ltd.), 4 July, 1995 (04. 07. 95) (Family: none)	5
Y	JP, 5-28305, A (Fujitsu Ltd.), 5 February, 1993 (05. 02. 93) (Family: none)	14
Y	JP, 8-55188, A (Toshiba Corp.), 27 February, 1996 (27. 02. 96) & US, 5790701, A	8-13
Y	JP, 9-81743, A (Toshiba Corp.), 28 March, 1997 (28. 03. 97) (Family: none)	8-13

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 June, 1999 (15. 06. 99)	29 June, 1999 (29. 06. 99)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (July 1992)

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ９９／０１９８６

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl[6]　　G06K9/20

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl[6]　　G06K9/20

最小限資料以外の資料で調査を行った分野に含まれるもの

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
X Y	ＪＰ，９－１６７１３，Ａ （シャープ株式会社） １７．０１月．１９９７（１７．０１．９７）（ファミリーなし）	1, 4, 15, 16, 5, 8-14
X Y	ＪＰ，５－１６６００２，Ａ （サイテックス・コーポレイション・リミテッド） ０２．０７月．１９９３（０２．０７．９３）（ファミリーなし）	1, 15, 16 8-14

☒ Ｃ欄の続きにも文献が列挙されている。　　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
１５．０６．９９	29.06.99

国際調査機関の名称及びあて先	特許庁審査官（権限のある職員）	
日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	月野　洋一郎	５Ｈ　９４７２
	電話番号　０３－３５８１－１１０１　内線　３５３１	

様式ＰＣＴ／ＩＳＡ／２１０（第２ページ）（１９９８年７月）

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Ｙ	ＪＰ，７－１６８９１１，Ａ （松下電器産業株式会社） ０４．０７月．１９９５（０４．０７．９５）（ファミリーなし）	5
Ｙ	ＪＰ，５－２８３０５，Ａ （富士通株式会社） ０５．０２月．１９９３（０５．０２．９３）（ファミリーなし）	14
Ｙ	ＪＰ，８－５５１８８，Ａ （株式会社東芝） ２７．０２月．１９９６（２７．０２．９６） ＆ＵＳ，５７９０７０１，Ａ	8-13
Ｙ	ＪＰ，９－８１７４３，Ａ （株式会社東芝） ２８．０３月．１９９７（２８．０３．９７）（ファミリーなし）	8-13