(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号
特許第3101153号
(P3101153)
(45)発行日　平成12年10月23日(2000.10.23)　(24)登録日　平成12年８月18日(2000.8.18)

(51)Int.Cl.⁷	識別記号	ＦＩ	
Ｇ０６Ｋ　9/68		Ｇ０６Ｋ　9/68	Ｇ
Ｇ０６Ｆ　17/22	５２０	Ｇ０６Ｆ　17/22	５２０Ｂ

請求項の数１(全 11 頁)

(21)出願番号　特願平6－143583
(22)出願日　平成６年６月24日(1994.6.24)
(65)公開番号　特開平8－16711
(43)公開日　平成８年１月19日(1996.1.19)
審査請求日　平成10年１月30日(1998.1.30)

(73)特許権者　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号
(72)発明者　船山　竜士
大阪府大阪市阿倍野区長池町22番22号
シャープ株式会社内
(74)代理人　100079843
弁理士　高野　明近

審査官　脇岡　剛

(56)参考文献　特開　平４－151761（ＪＰ，Ａ）
特開　昭62－251986（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.⁷，ＤＢ名)
G06K　9/68
G06F　17/22

(54)【発明の名称】　日本語入力装置

(57)【特許請求の範囲】
【請求項１】　読みのわからない文字を入力する際に、形の似ている文字で読みのわかっている文字をキー文字入力する漢字入力手段と、オンライン手書き文字認識用辞書及び光学文字認識用辞書を利用して、似ている文字群と該文字群の各文字の類字度を集めて自動生成される類字辞書と、前記漢字入力手段により入力されたキー文字に基づいて前記類字辞書から形の似ている文字を検索し、類字度を評価する類字検索手段とを備えた日本語入力装置であって、前記漢字入力手段から見方の異なる複数のキー文字を入力し、該キー文字ごとに前記類字辞書から形の似ている文字を検索することにより目的とする文字の候補を絞り込み、キー文字と類字度の高い順に候補文字を表示させることを特徴とする日本語入力装置。
【発明の詳細な説明】

【０００１】
【産業上の利用分野】本発明は、日本語入力装置に関し、より詳細には、日本語入力が可能な機器において、読みのわからない文字を入力する際、形の似ている文字を集めた辞書を用いて、その字に形が似ている文字をまず入力してから前記辞書を用いて目的の字を検索するようにした日本語入力装置に関する。例えば、ワープロ、パソコン、オフコン、ワークステーション、電子手帳など、辞書を検索する方式を用いた日本語入力方式を持つすべての機器に適用されるものである。
【０００２】
【従来の技術】現在主流になっているかな漢字変換は、読みをかな入力し、入力されたかなを変換することによって目的の字を得るようになっているが、読みのわからない文字を入力する方法としては、コード入力や部首入

力、画数入力などがある。コード入力は、図２５に示すように、字や記号と１：１に対応したコード（ＪＩＳ、シフトＪＩＳ、区点コードなど）を、通常は紙に印刷された一覧表（通常、第１水準の漢字は音読みでのよみがなの５０音順、第２水準は部首の画数順の配列になっている）から拾い出し、そのコードを入力して変換することによって目的の字を得るものである。一覧表が紙に印刷されておらず、電子化されている場合もあり、その場合は、コードを入力するのではなく、ディスプレイ上で紙の上と同じように目的の字を探し出し、ポインティングデバイスなどを用いてその字を指定することもある。

【０００３】部首入力は、図２６に示すように、目的の字がどの部首に属するかを、通常は紙に印刷された一覧表から拾い出し、そこに書いてあるコードを入力して目的の字を得るものである。一覧表が紙に印刷されておらず、電子化されている場合もあり、その場合は、コードを入力するのではなく、ディスプレイ上で紙の上と同じように目的の字を探し出し、ポインティングデバイスなどを用いてその字を指定することもある。

【０００４】画数入力は、文字の総画数が同じものを集め、画数の順にその漢字群を配列した表から、目的の文字を目視によって検索するものである。紙に印刷された一覧表から、該当する文字のコードを探し出し、そのコードを入力して目的の文字を得る場合と、一覧表が電子化されており、ディスプレイ上で紙の上と同じように目的の文字を探し出し、ポインティングデバイスなどを用いてその文字を指定する場合とがある。

【０００５】従来の日本語入力装置について記載した公知文献としては、例えば、特開平２－１８６６１号公報がある。この公報のものは、音訓入力手段と部首入力手段と画数入力手段の少なくとも１つの入力手段と、字形構成要素辞書及び字形構成要素入力手段とによって、それぞれの辞書の共通をとって候補文字を表わすもので、候補対象漢字の数を減らすことができるというものである。すなわち、読みや部首、画数の情報を複合して、複数の条件を設定し、目的の文字を絞り込むというものである。

【０００６】また、特開平５－１２２４９号公報のものは、文書作成時に画面に表示された漢字パターンを、例えばカーソルで指定すると、部首解析処理手段は、部首解析用テーブルに基づいて、指定された漢字の部首名及びその部首の位置を解析し、検索出力手段は、部首解析処理手段による解析結果に基づいて、指定された漢字の部首名及びその部首の位置に対応する全ての漢字コードを部首辞書メモリ手段から検索する。これにより、部首名がわからない場合でも、同一の部首名を有する漢字を指定するだけで、目的の漢字を得ることができる。また、目的の漢字を構成する部首名の一部しかわからない場合でも、その部首名に基づいて目的の漢字を得ることができるものである。

【０００７】

【発明が解決しようとする課題】パソコンや電子手帳などの日本語入力が可能な情報機器などにおいて、通常は読みをかなで入力し、入力されたかなを漢字などに変換する方式などを用いているが、それだけでは、読みのわからない字や記号を入力することができない。そのような文字を入力する場合、従来の方法では、総画数順に漢字が配列された表を、紙上あるいは電子の形で持ち、画数を手ががりに目的の字をその表から目視によって検索していた。あるいは、部首ごとに漢字を配列した表を、紙上あるいは電子の形で持ち、部首を手がかりに目的の字を、目視によって検索していた。通常は、部首の画数順に同じ部首を持った漢字群が配列され、同じ部首を持った漢字群の中では、総画数の順に字が配列されている。

【０００８】しかし、総画数より検索する方式だと、同じ画数を持つ漢字は膨大な数に昇り、そこからひとつの文字を探し出すのは、非常に困難であり、また、画数がはっきりと特定できない文字も存在するため、検索はさらに困難を極める。また、部首より検索する方式だと、目的の漢字の部首がきちんとわかっていなければならないことが条件となる。漢字によっては、部首が何になるのかわからないものもあり（例えば、彗星の「彗」の部首は「彑」である）、その場合、目的の字を探し出すのは非常に困難である。また、多くの部首では、同じ部首を持つ文字の数が、非常に多くなり、同様に検索には困難が伴う。

【０００９】また、コード入力の場合、一覧表は通常、第一水準ならば音読みでのよみがなの５０音順に配列されており、もとより読みのわからない字を入力しようとしているわけであるから、これでは役に立たないことがわかる。第２水準ならば部首の画数の小さい順に配列されており、目的の字の画数を数えるという手間がかかる上に、部首のわからない文字を探すには相当手間がかかる。部首入力の場合は、目的の字の部首がわからない場合は探すのに大変な手間がかかるし、同じ部首を持つ字が多く存在する場合の検索も時間がかかる。総画数入力の場合は、目的の文字の画数を数えるのがまず面倒であり、多画の文字であれば、画数の数え間違いなどが発生して検索の効率が悪い。また、同じ画数の文字も多く存在するので、目的の文字を探すには手間がかかる。

【００１０】前述した特開平２－１８６６１号公報に示されている方法を用いた場合、「読み」はわからないわけであるから、部首と画数から目的の文字を検索するわけであるが、部首がわからないものに関しては、総画数入力と同じ条件になるし、部首がわかっても、文字の画数を数えるという作業は、非常に根気のいる作業であり、あまり、効率のいい検索方法であるとは言えない。一覧表が電子化されている場合も同様で、ディスプレイ上で表示できる字の数は、紙の上で表示できる数より少

ない場合が多く、その場合、目的の字を探すにはさらに時間がかかることになる。

【００１１】また、類字変換を実現するための類字辞書を人力で作成することにおいて、形の似ている字を人手を使って集め、それらの類字度を主観によって評価するとなると、似ている字がまだあるにもかかわらず辞書から洩れてしまったり、類字度の評価が類字グループで統一されていなかったり、また、使用者と辞書作成者で類字の評価に違いがある場合など、使用感が著しく損なわれる可能性がある。もとより、人手を使ってこの作業を行うとなると、それにかかる労力・時間（コスト）は大変なものとなる。

【００１２】また、前述したように、特開平５－１２２４９号公報のものは、同じ部首を持つ漢字から目的の字を検索するだけのものである。本発明はあくまでも、読みのわからない字を入力するために、「形の似ている字（＝類字）」をキーにして、類字辞書を引き、目的の字を検索するものである。本発明においても、共通に持つ部首をキーとして目的の文字を検索する手段を有しているが、それはあくまでも一つの手段であり、本発明では、部首が違っていても、形が似ている字（「瓦」と「互」など）も、検索の対象としている。

【００１３】また、特開平５－１２２４９号公報においては、ただ単に、部首の同じ字を辞書から拾ってくるだけであるが、本発明では、「類字度」という概念を導入し、より似ている字を候補の先に配列することができるようになっている。例えば、特開平５－１２２４９号公報では、「苑」を入力しようとして、「宛」をキーにしても、通常では、「うかんむり」の文字がたくさん出てきてしまい、効率的に「苑」を見つけることはできない。また、部首の位置を指定すれば、もう少し候補を絞ることができるかもしれないが、この作業自体が非効率的である。そこで、本発明における辞書の自動生成方法に従って作成された辞書を用いて類字検索を行えば、特開平５－１２２４９号公報より遥かに効率的に、読みのわからない文字を検索することができる。

【００１４】本発明は、このような実情に鑑みてなされたもので、日本語入力が可能な機器において、読みのわからない文字を入力する際、形の似ている文字を集めた辞書を用いて、その字に形が似ている文字をまず入力し、前記辞書を用いて目的の字を検索するようにした日本語入力装置を提供することを目的としている。

【００１５】

【課題を解決するための手段】本発明は、上記目的を達成するために、読みのわからない文字を入力する際に、形の似ている文字で読みのわかっている文字をキー文字入力する漢字入力手段と、オンライン手書き文字認識用辞書及び光学文字認識用辞書を利用して、似ている文字群と該文字群の各文字の類字度を集めて自動生成される類字辞書と、前記漢字入力手段により入力されたキー文字に基づいて前記類字辞書から形の似ている文字を検索し、類字度を評価する類字検索手段とを備えた日本語入力装置であって、前記漢字入力手段から見方の異なる複数のキー文字を入力し、該キー文字ごとに前記類字辞書から形の似ている文字を検索することにより目的とする文字の候補を絞り込み、キー文字と類字度の高い順に候補文字を表示させる日本語入力装置であることを特徴とする。

【００１６】

【作用】前記構成を有する本発明の日本語入力装置は、読みのわからない文字の検索を簡単に行うことができる。すなわち、ほとんどの漢字は、それほど多くない、いくつかの部品から成ると考えられる。例えば、「腕」という字は、「月」「宀」「夕」「巳」という４つの部品から成る。また、「苑」という字は、

【００１７】

【表１】

「艹」

【００１８】「夕」「巳」の３つの部品から成る。「腕」と「苑」では「夕」「巳」の部品が共通であり、従って、この二つの字は、似ているということが言える。共通の部品の割合が多いほど、二つの字の形の「類字度」は高いということが言える。こういったタイプの形の似ている字の集合は、オンライン手書き文字認識のための、認識用の辞書を利用して作成することができる。また、部品が共通でなくても、例えば「互」と「瓦」のように、形が似ている文字もまたある。これは、光学文字認識用の辞書を利用して作成することができる。

【００１９】このように、形の似ている字と、それがどれだけ似ているかの指標（類字度）を要素としてもつ「類字辞書」を用意し、この辞書を利用して読みのわからない文字を検索する。そのためには、まず、入力したいが読みのわからない文字に似た字を、通常のかな漢字変換方式などを用いて入力する。そして、その字をキーにして、類字辞書を引く。同様に、別の似ている字をキーにして類字辞書を引き、目的の字を絞り込んでゆく。このようにして、（複数の）キー文字から候補を得、その中から目視で目的の字を見つけ出す。従来の総画数表や部首別表を目視して探し出すことに比べ、類字辞書を引いて検索すると、少ない候補文字の中から検索すれば良いため、検索時間は圧倒的に有利になる。

【００２０】

【実施例】実施例について、図面を参照して以下に説明する。図１は、本発明による日本語入力装置の一実施例を説明するための構成図で、図中、１は漢字入力手段、２は類字検索部、３は類字辞書、４はバッファ管理部、５はバッファ部、６は候補表示・選択部、７はかな漢字変換辞書等である。漢字入力手段１は、類字入力以外の

漢字入力を行うもので、かな漢字変換手段などで実現する。該漢字入力手段１によりキー文字を与えると類字検索部２により類字辞書３を検索する。類字辞書３は、日本語入力が可能な機器において、形の似ている文字群とその類字度を集めたものである。前記類字検索部２により候補文字を与えると、バッファ管理部４ですでにバッファ部５に記憶されている文字と新しい候補文字とから新しいバッファを作成し、候補表示・選択部６で候補文字を表示・選択する。なお、かな漢字変換辞書７は、前記漢字入力手段１によるキー文字を与える場合に、必要に応じて用いられるもので、かならずしも必要とするものではない。

【００２１】図２は、類字辞書の構造を示す図である。類字変換を行うためには、まず類字辞書を用意する必要がある。ある字に注目し、その字に似ている字を集める。例えば、「腕」という字をキーにそれに似ている字を列挙してみると、「椀婉宛苑蛇怨」などの字が見つかる。これらはお互いに形が似ており、その一群の字の集合を「類字グループ」と呼ぶ。類字グループの集合が類字辞書である。そして、これら類字グループの要素文字に点数を付ける。この点数は、キーとなる字に対する類字度であり、点数が大きいほどより形が似ているということを意味する。類字辞書は、すべての文字をキーにしてそれに似た字を羅列し、各自に類字度を付加したものの集合である。

【００２２】図２において、まず、ＪＩＳコードの順に、キー文字に対する類字グループへのポインタが置かれる。辞書の先頭は、ＪＩＳコードの１番目である、２１２１（１６進数）の文字（全角スペース）に対する、類字グループへのポインタが置かれることになる。そのポインタの先には、全角スペースの類字文字とその類字度が配列される。以下、ＪＩＳコード順に、類字グループへのポインタが羅列される。

【００２３】例えば、「腕」の類字は、「腕」のＪＩＳコードである４Ｆ５３のポインタ格納場所に格納されているポインタが示す場所から、その類字度と共に配列されている。「腕」の類字として「椀」「婉」「宛」「苑」「蛇」「怨」の６字があり、類字度がそれぞれ、９０、８０、７０、６０、５０、４０と仮定する。そうすると、「腕」の類字グループへのポインタが示す場所から、「椀，９０，婉，８０，宛，７０，苑，６０，蛇，５０，怨，４０，０」と配列されることになる（最後の０は、エンドコード）。このように、すべてのＪＩＳ文字に関して、その類字を配列した辞書を類字辞書として使用する。

【００２４】次に、類字辞書を検索して入力した文字に似た字を出力する動作について説明する。図３は、類字変換の手順を示すフローチャートである。以下、各ステップ（Ｓ）に従って順に説明する。読みのわからない文字を入力するために、この方式では、形の似ている字で読みのわかっている文字をまず入力する（Ｓ１）。これは、通常のかな漢字変換を用いるなどして、入力する。そして、その文字をキーにして、類字辞書を検索する。

【００２５】次に、キー文字のＪＩＳコードをもとに、その文字の類字グループを指すポインタを格納しているアドレスを算出する。そのアドレスからポインタを読み出し、類字グループの先頭アドレスを得る。そのアドレスからキー文字の１つ目の類字及びその類字度を読み出し、バッファに格納する。ポインタを進め次の類字及びその類字度を読み出す。エンドコードを読み出すまでこの動作を続ける。これによって、バッファには、キー文字の全ての類字と類字度が格納されることになる（Ｓ２）。

【００２６】同様に、別のキー文字がある場合（Ｓ４）、同じ検索を行い、類字及び類字度を読み出す。そして、そのキー文字から検索された類字の中に、既にバッファにある文字と同じものがあれば、その類字度を、バッファにある文字の類字度に加え、それをその文字の新たな類字度とする。つまり、別々のキー文字から同じ文字が抽出された場合、類字度を足して、候補表示の優先順位をあげるのである。新たに読み出された類字が、バッファの中にない場合は、バッファにその文字を加える。バッファが溢れる場合は、類字度の小さいものから削除する（Ｓ５）。

【００２７】このようにして、キー文字を複数指定し、探している可能性の高い文字を優先して前の方に表示することにより、目的の文字を探し易くすることができる。類字度順に配列された候補文字の中から目的の文字を選択し、これをもって類字変換が完了する（Ｓ６）。

【００２８】図４は、図３に示す類字変換の手順に従って目的の文字を選択する動作例を示すフローチャートである。以下、各ステップ（Ｓ）に従って順に説明する。ここでは、「怨」という字を入力したいのだが、その読みがわからない場合を想定して考える。まず、この字に似ている字を、キー文字として入力する必要がある。ここでは、「宛」という字をまず、キー文字として入力している。キー文字の入力は、通常のかな漢字変換などを用いて行う（Ｓ１１）。そして、このキー文字を類字変換する。候補文字として、「椀婉宛苑蛇怨…」が見つかり、類字度順に候補文字として表示される（Ｓ１２）。

【００２９】次に、キー文字「怒」を入力する（Ｓ１３）。このキー文字を類字変換すると、「怨努恋忘…」などが候補として上がるが、すでに「宛」を類字変換した候補が上がっているため、それと統合した新しい候補文字が表示されることになる（Ｓ１４）。ここでは、「宛」と「怒」を類字変換し、目的の「怨」が、候補の先頭に来ていることがわかる（Ｓ１５）。

【００３０】次に、類字辞書の自動作成方法について説明する。この類字辞書の自動作成方法には、（１）光学文字認識用辞書を用いる場合と、（２）オンライン手書

き文字認識用辞書を用いる場合とがある。ある文字と、別の文字がどれだけ似ているかを定量的に表現するのは難しい。また、個々の文字について、それに形の似ている文字を見つけ、類字度を評価して、類字辞書を作成するのは、人手を使って行うため、大変な労力・時間を必要とする。ここでは、光学文字認識やオンライン手書き認識の手法を用い、類字辞書の自動生成を行う手法を説明する。

【００３１】まず、光学文字認識用辞書を用いて類字辞書を自動生成する方法について説明する。これは通常、認識すべき各文字のドット情報をベクトルの形で持ち、これと、ＯＣＲ（Optical Character Reader：スキャナなど）で入力されたドットパターンとのマッチングを行うものである。図５（ａ），（ｂ）にその概念を示す。ここでは、８ドット×８ドットの文字を認識することを考える。

【００３２】まず、８ドット×８ドットの格子に、文字パターンが記録されている。それを、ドットがＯＮになっている部分を１、ＯＦＦになっている部分を０とし、左上から順にその値を並べ、６４次元のベクトルとして表現する。このベクトルが、認識対象としている各文字について存在している。この各文字のベクトルの集合が、文字認識用辞書である。

【００３３】ここで、ＯＣＲから文字の入力がある。１文字８ドット×８ドットになるよう正規化されたデータを用い、辞書データとのマッチングを行う。辞書作成時と同様に、入力された８ドット×８ドットのデータを６４次元のベクトルにし、それと辞書に記録されている各文字のベクトルとの距離を計算する。この２つのベクトルの距離が一致度である。辞書にある全ての文字との一致度を計算し、一番、一致度が高いものを、認識結果として出力するのが、光学文字認識の基本手法である。

【００３４】

【数１】

一致度の計算は、

辞書データの特徴ベクトルを$\vec{a}$＝(０,０,０,０,１,０,０,０,…,０)

入力データの特徴ベクトルを$\vec{b}$＝(０,０,０,０,１,０,０,０,…,０)

とすると、

$$\frac{(\vec{a}\cdot\vec{b})}{|\vec{a}||\vec{b}|} = \text{一致度}$$

で求められる。

【００３５】従って、この光学文字認識の手法を用い、辞書データとＯＣＲからの入力データのマッチングを行うのではなく、辞書にある各文字間でマッチングを行う。マッチングの手法は全く同じで、各文字の６４次元ベクトル間の距離を求め、それを類字度とする。各文字について、一定の類字度以上を持つ文字とその類字度を、目的文字の類字グループとして辞書に登録する。

【００３６】以下、光学文字認識用辞書を用いて類字辞書を自動生成する方法について、さらに詳細に説明する。

類字度の評価法

光学文字認識の原理について説明すると、システムが内部に持っている文字のマトリックスパターンと入力されたマトリックスパターンとが、どれだけ似ているかということを、そのマトリックスから得られるベクトルデータの距離でもって評価するというものである。話しを簡単にするために、３ドット×３ドットのデータで説明する。

【００３７】図６（ａ）を辞書にあるマトリックスパターン、図６（ｂ）を入力パターンとする。図６（ａ）、図６（ｂ）の３×３ドットのマトリックスパターンは、ドットがある部分を１、ない部分を０とすると、３×３＝９個の０、１の羅列、すなわち、９次元のベクトルデータで表わすことができる。これら、２つの９次元ベクトルデータの「距離」は、以下の表２で表わされる。

【００３８】

【数２】

$$\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|} = \text{類字度}$$

【００３９】従って、辞書にあるパターンと、入力パターンが全く同じであれば、類字度は１となり、２つのパターンがずれていればいる程、０に近くなる。この値を類字度とする。

【００４０】マトリックスデータより類字辞書を作成する方法

図７は、マトリックスデータより類字辞書を作成するためのブロック図で、図中、１１は光学文字認識用辞書、１２は類字度評価部、１３は光学文字認識版類字辞書である。光学文字認識用辞書１１には、入力されたマトリックスパターンと比較するための、各文字のマトリックスパターンデータがストアされており、通常は、図８に示すように、文字コード順に並んでいる。図８に示すマトリックスパターンデータから、類字辞書を自動作成するために、類字度評価部において、前述した類字度の計算法により、まず、図８の辞書の最初の文字（ＪＩＳコード２１２１の文字）と、辞書の残りの文字全部との間の類字度をそれぞれ計算する。そして、ある一定の類字度（例えば０.７）以上の文字について、その文字をＪＩＳコード２１２１の文字の類字として類字辞書に登録する。

【００４１】次に、ＪＩＳコード２１２２の文字と、それ以外の、辞書に含まれる全部の文字との類字度を計算し、同様にして、一定の類字度以上の文字を、ＪＩＳコード２１２２の文字の類字として、類字辞書に登録する。同様にして、すべての文字について、残りの文字との類字度の値が一定以上の文字を、その文字の類字として類字辞書を作成するのである。このようにして、光学文字認識版類字辞書１３ができる。

【００４２】以上が、光学文字認識用辞書を用いて類字辞書を自動作成する方法についての説明であるが、次に、オンライン手書き文字認識用辞書を用いて類字辞書を自動作成する方法について説明する。

【００４３】図9は、オンライン手書き文字認識の手法を説明するための図で、図１０は、オンライン手書き文字認識用の辞書構造の概念図である。オンライン手書き文字認識であるが、これは、文字を単純なストローク（基本ストローク）の組合せとして捉え、オンラインによる手書きの入力を基本ストロークごとに認識し、基本ストロークの集合がサブパターンで、そのサブパターンの集合が文字という風に、ボトムアップ的に文字を認識するものである。

【００４４】従って、オンライン手書き文字認識用辞書には、文字がどのようなサブパターンから構成されているのかを示す情報があり、これを類字辞書の生成に役立てる。すなわち、サブパターンとは、認識のアルゴリズムにもよるが、漢字における、へん、つくり、あるいは、それ自身が別の漢字になるような漢字の構成部品であり、ほとんどの漢字はいくつかのサブパターンの組合せで構成されている。従って、これらサブパターンを共有する文字どうしは形が似ているということになり、共有する割合が大きいほど、類字度が高いといえる。

【００４５】例えば、「腕」という字は、「月」「宀」「夕」「巳」の４つのサブパターンで構成されている。サブパターンの一致率が高いほど、各文字間の類字度が高くなるように設定する。また、出現頻度の高いサブパターンほど、類字度へのウェイトを高くすることにより、より正確な辞書を生成することができる。そのような方法で、全ての文字に関して、他の文字との類字度を計算し、一定の類字度以上を持つ文字とその類字度を、目的文字の類字グループとして辞書に登録する。

【００４６】以下、オンライン手書き文字認識用辞書を用いて類字辞書を自動生成する方法についてより詳細に説明する。

類字度の評価法

オンライン手書き文字認識用辞書は、図１０に示すような構成になっている。この例では、「結」は「糸」「士」「口」の３つのサブパターンから構成され、「詰」は、「言」「士」「口」の３つのサブパターンから構成されていることを示している。具体的には、辞書は各々図１１～図１３に示すような構成になっている。

【００４７】図１１に示す基本辞書には、文字（「詰」「結」など）が、どういったサブパターンから構成されているかを示す情報が含まれている。図１２に示すサブパターン辞書には、各サブパターンが、どういった基本ストロークから構成されているかを示す情報が含まれている。図１３に示す基本ストローク辞書には、各基本ストロークがどういうものかを示す情報が含まれている。

【００４８】さて、類字度の評価法であるが、２つの文字の間で共通のサブパターンが多いほど、２つの文字はより似ていると言うことができる。例えば、「苑」「宛」「草」の３つの文字について考えてみる。それぞれの文字は、図１４に示すようなサブパターンから構成されるとする（サブパターンの選び方は、認識システムにより違っている）。

【００４９】「苑」と「宛」では、「夕」「巳」が共通であり、３つのサブパターンのうち、２つが共通と言うことになる。一方、「苑」と「草」では、

【００５０】

【表２】

「艹」

【００５１】が共通であり、同じく３つのサブパターンのうち、１つが共通と言うことになる。従って、ごく単純に考えれば、「苑」と「宛」の類字度は２／３＝０．６７、「苑」と「草」の類字度は、１／３＝０．３３ということになる。

【００５２】しかし、共通なサブパターンの比率のみを２つの文字の類字度とすると、類字度が同じ値のものが数多く出現してしまい、これで構成した辞書を用いると、候補が極端に多かったり、少なかったり、ということになり、本発明の本来の目的から外れてしまう。そこで、サブパターンごとに重みを設定することにより、この問題を解決することができる。いま仮に、図１５に示すように重みを設定したとする。

【００５３】重み付けの原則は、「多くの文字に共通に現れるものは小さく、滅多に現れないものほど大きくする」ということである。例えば、「くさかんむり」や「うかんむり」を持った文字は非常に多く存在し、従って、これらを共通に持つ文字どうしは、それほど似ているとは思わないであろう。しかし、「巳」のように、それほど多くの文字に現れないサブパターンを共通に持つ文字どうしは、より似ていると感じるはずである。

【００５４】従って、サブパターンの出現頻度に合わせて重み付けを行うことにより、より似ていると感じるであろう文字どうしの類字度を高くすることができる。この重みに関しては、全ての文字のサブパターンの出現頻度を調べ、その値に比して決定することが原則となるが、利用者の主観により、似ていると感じるものは、必ずしも出現頻度に依存するとは限らない。従って、より使い易い辞書を作成するためには、重みの調整が必要と

なる。

【００５５】さて、図１５に示す重みにしたがい、「苑」と「宛」、「苑」と「草」の類字度を計算すると、図１６（ａ）～（ｄ）に示すようになる。図１６（ａ）に示す重みにしたがうと、共通なサブパターンは「夕」と「巳」だから、「苑」から見た「宛」の類字度は、図１６（ｂ）に示すように０.８６となる。また、図１６（ｃ）に示す重みにしたがうと、共通なサブパターンは

【００５６】

【表3】

「艹」

【００５７】だから、「苑」から見た「草」の類字度は、図１６（ｄ）に示すように０.１４となる。この式を一般化すると、文字Ａから見た文字Ｂの類字度は、図１７に示すようになる。この式を見ればわかる通り、類字度は０から１の値をとり、１に近いほど類字度が高いということになる。

【００５８】オンライン手書き認識用辞書から類字辞書を作成する方法

図１８は、オンライン手書き認識用辞書から類字辞書を作成するためのブロック図で、図中、２１はオンライン手書き文字認識用辞書、２２は類字度評価部、２３はオンライン手書き文字認識版類字辞書、２４は重み評価部、２５は重みテーブルである。前述した図１１～図１３に示す辞書により、図１７の類字度評価法を用いて類字辞書を自動生成する。

【００５９】まず、オンライン手書き文字認識用辞書２１を用いて、重み評価部２４によりサブパターンの重みテーブル２５を作成する。図１１に示す基本辞書の全ての文字を調べ、サブパターンごとにその数をカウントする。最も多くカウントされたサブパターンの重みを１として、サブパターンの重みを正規化し、重みテーブル２５を作成する。前述したように、主観による類字度の影響を考慮した重みの調整を行うことが望ましい。次に、その重みに従って、類字度評価部２２により、ある文字とその残りの文字との類字度を計算する。その類字度が一定以上（例えば０.７以上）のものを、その文字の類字として、一定以上の類字度を持つものを類字辞書に登録して、類字辞書２３を作成する。

【００６０】以上が、オンライン手書き文字認識用辞書を用いて類字辞書を自動生成する方法についての説明である。次に、光学文字認識版類字辞書とオンライン手書き文字認識版類字辞書の統合について説明する。前述のようにして生成した光学文字認識用辞書から作成した類字辞書と、オンライン手書き文字認識用辞書から作成した類字辞書とをひとつにまとめ、図２に示すような類字変換のための類字辞書とする。

【００６１】具体的には、一つのキー文字に対し、光学文字認識版類字辞書とオンライン版類字辞書からそれぞれ類字を取り出し、それぞれに適当な重みを付けて類字度を足し合わせる。「音」の類字が図１９（類字度は満点が１００になるよう正現化している）に示すようになっており、光学文字認識版の重みを０.４とし、オンライン版の重みを０.６として「音」の総合類字度を計算すると、図２０に示すようになる。この値を「音」の「類字度」として類字辞書に登録する。ここでも、類字度が一定以上（例えば３０以上）の数値のみを辞書に登録することにより、登録数が多くなり過ぎないようにする。また、重みは仮に、０.４と０.６としたが、この値は、実際に複数の利用者が類字変換を行ってみて、最も適当と思われる数値に調整すべきである。

【００６２】次に、類字辞書の検索／表示方法について説明する。類字変換の手順は、図３に示している通りである。ここでは、辞書の検索などについて追加説明する。類字辞書は、図２に示したような構造になっている。ここでは、類字度を０から１００に正規化した数値を用いている。キー文字が与えられた場合、類字検索部２は類字辞書３を検索し、キー文字の類字とその類字度をバッファ部５に格納する。例えば、キー文字を「宛」とし、その類字と類字度が図２１に示すようになっている場合、バッファ部５は、図２２に示すようになる。

【００６３】次に、別のキー文字「怒」を与えたとする。「怒」の類字と類字度は、図２３に示すとおりとする。この類字と類字度をすでにバッファにあるものに追加する。もし、新しいキー文字に対する類字の中で、すでにバッファにある文字と共通なものがあれば、その類字度をバッファの中の類字度に加えると、図２４に示すようになる。バッファの大きさに限りがある場合は、類字度の小さいものから削除していく。また、類字度の閾値を適応的に変化させ、一定の類字度以上の文字だけバッファに残すようにしてもよい。閾値を適応的に変化させるとは、キー文字１つ指定の場合は、類字度５０以上、２つ指定の場合は７０以上、３つ指定の場合は１００以上ということである。このようにして、複数のキー文字を指定することにより、目的の文字を候補の前方に持ってくることが可能となる。以上の手法により、光学文字認識用辞書やオンライン手書き文字認識辞書から、類字辞書を自動生成することが可能になる。

【００６４】

【発明の効果】以上の説明から明らかなように、本発明によると、以下のような効果がある。

（１）現在使われている、読みのわからない文字を入力する方法では、効率的に目的の文字を探し出すことは困難であるので、「類字変換」機能を用いれば、これら従来の方式を用いた検索方法より、効率的に目的の文字を探し出すことができる。検索にかかる時間の短縮と操作の簡便化が達成される。また、読みがわからなくても、部首がわからなくても、目的の文字を検索することがで

き、さらに、画数を数える必要もない。

（２）類字辞書を、光学文字認識やオンライン手書き文字認識の手法を用いて自動生成することにより、類字辞書編成における字洩れの防止や類字度の客観的な評価、類字グループ間での類字度評価の統一が実現でき、快適な使用感を持つ類字変換のための辞書を作成することができる。また、正確な辞書を短時間かつ低労力（低コスト）で作成することができる。

（３）漢字入力手段から見方の異なる複数のキー文字を入力し、該キー文字ごとに類字辞書から形の似ている文字を検索し、目的とする文字の候補を絞り込み、キー文字と類字度の高い順に候補文字を表示させるので、目的とする文字の発見を容易にすることができる。

【図面の簡単な説明】

【図１】本発明による日本語入力装置の類字検索の一実施例を説明するための構成図である。

【図２】本発明における類字辞書の構造を示す図である。

【図３】本発明における類字変換の手順を示すフローチャートである。

【図４】本発明における類字変換の動作例を示す図である。

【図５】本発明における光学文字認識の手法を説明するための図である。

【図６】本発明における辞書パターン及び入力パターンを示す図である。

【図７】本発明におけるマトリックスデータより類字辞書を作成するためのブロック図である。

【図８】本発明における類字度を示す図である。

【図９】本発明におけるオンライン手書き文字認識の手法を説明するための図である。

【図１０】本発明におけるオンライン手書き文字認識用の辞書構造の概念図である。

【図１１】本発明における基本辞書を示す図である。

【図１２】本発明におけるサブパターン辞書を示す図である。

【図１３】本発明における基本ストローク辞書を示す図である。

【図１４】本発明における類字度の評価法を説明するための図（その１）である。

【図１５】本発明における類字度の評価法を説明するための図（その２）である。

【図１６】本発明における類字度の評価法を説明するための図（その３）である。

【図１７】本発明における類字度の計算を示す図である。

【図１８】本発明におけるオンライン手書き認識用辞書から類字辞書を作成するためのブロック図である。

【図１９】本発明における光学文字認識版類字辞書とオンライン手書き文字認識版類字辞書を示す図である。

【図２０】本発明におけるキー文字に対する総合類字度を計算した例を示す図である。

【図２１】本発明における類字と類字度を示す図である。

【図２２】本発明における類字と類字度のバッファの状態を示す図である。

【図２３】本発明における類字と類字度を示す図である。

【図２４】本発明における類字と類字度のバッファの状態を示す図である。

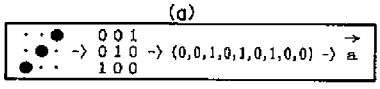
【図２５】従来のコード入力用の一覧表を示す図である。

【図２６】従来の部首入力用の一覧表を示す図である。

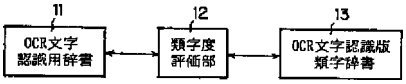
【符号の説明】

１…漢字入力手段、２…類字検索部、３…類字辞書、４…バッファ管理部、５…バッファ部、６…候補表示・選択部、７…かな漢字変換辞書等、１１…光学文字認識用辞書、１２…類字度評価部、１３…光学文字認識版類字辞書、２１…オンライン手書き文字認識用辞書、２２…類字度評価部、２３…オンライン手書き文字認識版類字辞書、２４…重み評価部、２５…重みテーブル。

【図６】

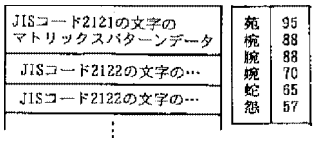


【図７】



【図８】

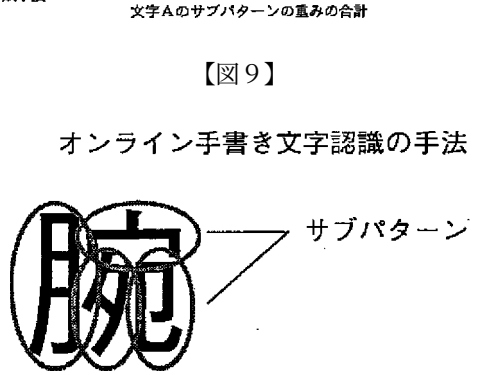
【図２２】



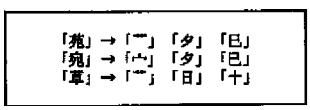
【図１７】

$$\text{類字度} = \frac{\text{文字Ａと文字Ｂで共通なサブパターンの重みの合計}}{\text{文字Ａのサブパターンの重みの合計}}$$

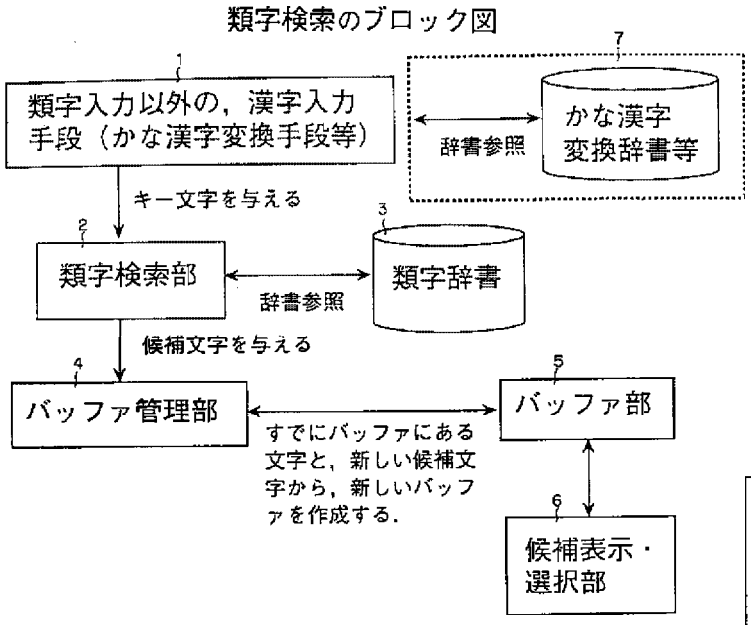
【図９】



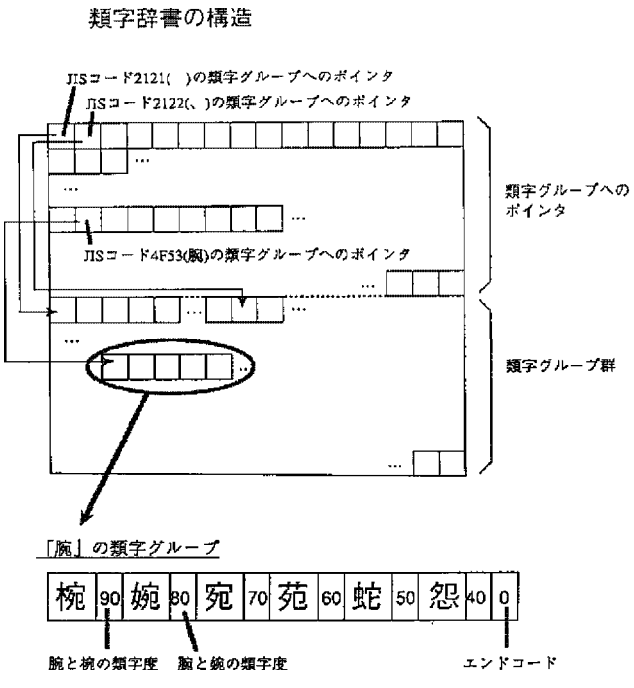
【図１４】



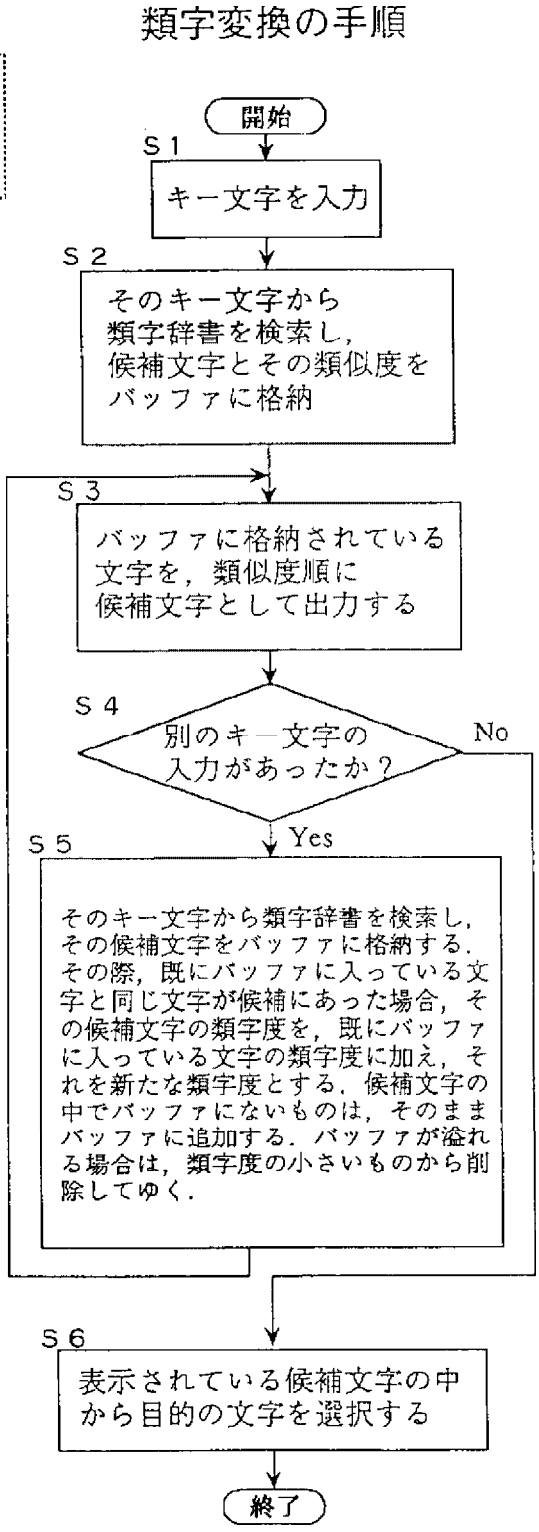
【図1】



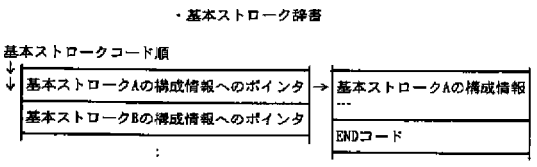
【図2】



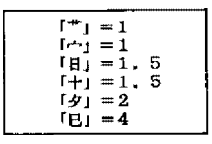
【図3】



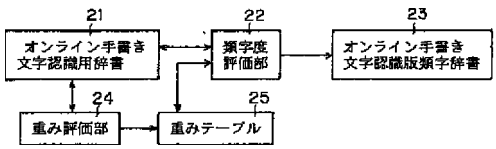
【図13】



【図15】



【図18】



【図４】

類字変換の動作例

（「怨」を入力したいのだが，読みが分からない場合の動作例）

S 11
キー文字：あて　通常変換　→　宛
（キー文字を入力するため，通常のかな漢字変換で「宛」の字を得る）

S 12
キー文字：宛　類字変換　→
候補文字：　苑椀腕婉蛇怨…
（一度目の類字変換では上のような候補文字が出現した）

S 13
キー文字：いかり　通常変換　→　怒
候補文字：　苑椀腕婉蛇怨…
（キー文字をもうひとつ指定する）

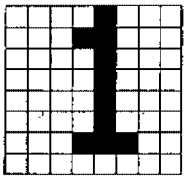
S 14
キー文字：怒　類字変換　→
候補文字：　怨苑努椀恐腕忘婉蛇…
（「怒」の類字には「怨努恋忘…」などがあり，前のキー文字と総合すると，上のような候補文字が得られる）

S 15
候補文字の中から，目的の字を選択する．ここでは，候補「怨」が先頭に来ている．

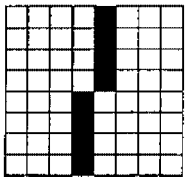
【図５】

OCR文字認識の手法

(a)
辞書データ



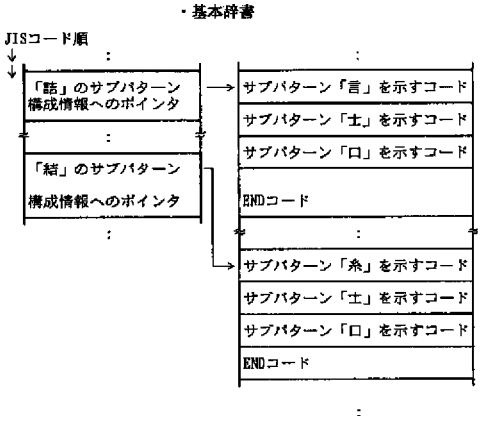
(b)
入力データ



【図２１】

キー文字「宛」：類字	苑	類字度	95
	椀		88
	腕		87
	婉		70
	蛇		65
	怨		57

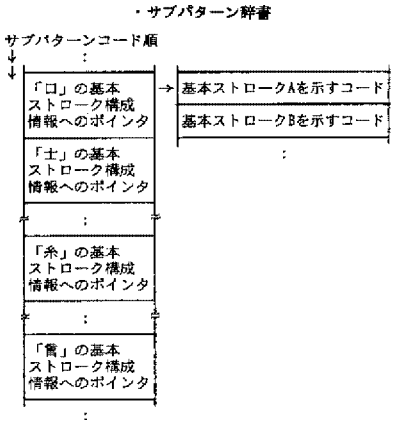
【図１１】



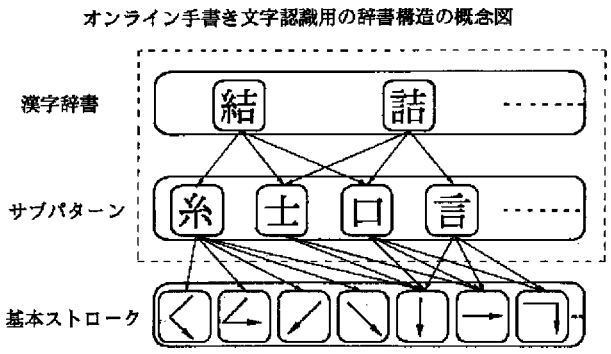
【図２４】

怨	128
苑	95
椀	88
腕	87
恋	85
努	72
婉	70
忘	67
蛇	65

【図１２】



【図１０】



【図１９】

OCR版類字辞書　キー文字「音」：類字	首	/ 類字度	72
	言		69
	旨		61
	音		59
オンライン版類字辞書　キー文字「音」：	早		66
	旨		66
	音		66
	日		50

【図２０】

	OCR版		オンライン版	
「旨」類字度 ＝	61×0.4	+	66×0.6	= 64
「音」	59×0.4	+	66×0.6	= 63.2
「早」	0×0.4	+	66×0.6	= 39.6
「日」	0×0.4	+	50×0.6	= 30
「首」	72×0.4	+	0×0.6	= 28.8
「言」	69×0.4	+	0×0.6	= 27.6

【図２３】

キー文字「怨」：類字	恋	/ 類字度	85
	努		72
	怨		71
	忘		67

(11)

【図１６】

(a)

「苑」と「宛」：重み「艹」＝1
「宀」＝1
「夕」＝2
「巳」＝4

(b)

$$\frac{\text{苑と宛の共通部分＝2（夕）＋4（巳）}}{\text{苑＝1（艹）＋2（夕）＋4（巳）}}=\frac{6}{7}=0.86$$

(c)

「苑」と「草」：重み「艹」＝1
「夕」＝2
「巳」＝4

(d)

$$\frac{\text{苑と草の共通部分＝1（艹）}}{\text{苑＝1（艹）＋2（夕）＋4（巳）}}=\frac{1}{7}=0.14$$

【図２５】

コード入力用の一覧表

10進	区点			0	1	2	3	4	5	6	7	8
16進		77JIS	JIS	0	1	2	3	4	5	6	7	8
イ	1700	893F	3120	院	陰	隠		韻	吋			
ウ	1700	893F	3120								右	宇
	1716	894F	3130	雨	臼	渦	嵐			唄	鬱	畝
	1732	895F	3140	雲								
	1732	895F	3140		在	薗	御			恢	悔	海
	1748	896F	3150	顔	芋							
	764	8950	3160									

【図２６】

部首入力用の一覧表