情報処理学会第42回(平成３年前期)全国大会

5D−10

ＯＣＲの認識率アップ法とそのシステムの簡素化

熊谷勝彦 , 鈴木真一 , 上野浩司
木更津工業高等専門学校 電気工学科

はじめに

近年、日本語ＯＣＲシステムの実用化が進んでいる。しかし、認識率向上のためにＯＣＲシステム自体が大規模化し、実使用に耐えるＯＣＲシステムは、専用のハードウェアを用いているのが現状である。そのため価格も高額になり、エンドユーザーまで普及するのは難しい。そこで、大規模化したＯＣＲシステムを認識率の向上をはかりながら簡素化し、パーソナルコンピューター上で利用できる方式を考案したので報告する。

概　要

［１］文字パターンの取得

文字認識をしようとする場合、どのような方法で、【文字】を文字として、取り出すか、ということになるが、今回の研究では、文字を１文字１文字の文字パターンとして、切り出す方法を使った。

文字の切りだしは、パソコンなどのプログラムのリストなどを想定している。

したがって、文字は横書きで、順序よく一列一列続いているものを、仮定する。

まず、文字の列を切り出すのだが、この場合、文字のデータは、デジタルの１と０のドットとして処理されている。文字の存在するドットの場所は、１のデータ、空白は０のデータで、構成されている。

文字の列として切り出すには、この１のドットのかたまりを、識別すればよい。

文字と空白とを、区切れば、文字列として切り出せることになる。

こうして、列として切りだした文字列を、さらに一文字ずつに、切り出す。

この場合も、先ほどと同じく、文字と空白とを、ドットを調べて識別し、１文字づつ切りだしていく。

［２］データの加工

そして、切り出された文字は、たて数ドット×よこ数ドットで、構成される。

このドットを、２進数のビットとして扱い、よこ方向にビットの重みをかける。

こうして、文字のドットとしての、情報は、ビットとしての、重みを持った情報として、扱われる。今回の、著者らの場合は、一文字を３０×３０ドットの情報として扱っている。（図１参照）

［３］データからの特徴抽出

縦３０×横３０ドットの範囲で構成された文字の情報は、［２］の方法により

Recognition Rate Up Method of OCR and Simplification of the System
Katuhiko KUMAGAI , Sinichi SUZUKI , Koji UENO

３０個のデータとなる。このデータから特徴を抽出するのに著者らは、標準偏差・変曲点数・基準パターンからの誤差を用いた。ここで変曲点数は、どの場合に有効であるかというと、活字とはいえ、半角，全角など文字のサイズが何種類か存在するため、文字が横方向に伸縮を起こしたデータに対しては、標準偏差及び、基準パターンからの誤差だけでは、ほとんど対応することができない。しかし、変曲点数だけは、文字が横方向に伸縮を起こした場合に於いても、あまり変化しないからである。

［４］データベース（基準文字パターン管理及び、検索用）

ＯＣＲシステムに於いて、データベースはその核をなすものであり、文字の特徴抽出データを効率的に検索するために、次のような仕様とした。

① 登録されているデータは、標準偏差・変曲点数・基準パターンの内、どの項目からも、自由にアクセスでき、認識の対象である文字データとの誤差が最も小さいものが、候補として出力される。

② 認識結果が正しかった場合、データベース内の許容範囲記録エリア（標準パターンの特徴から、どの程度外れている状態で、認識したかを記録する領域。）に誤差が記録され、次回からは、その許容範囲で検索が行われる。

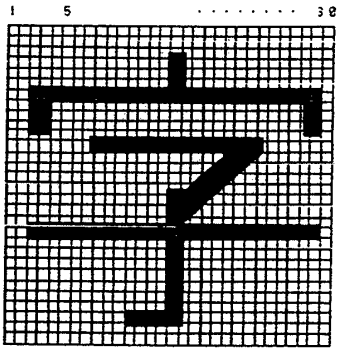

図１

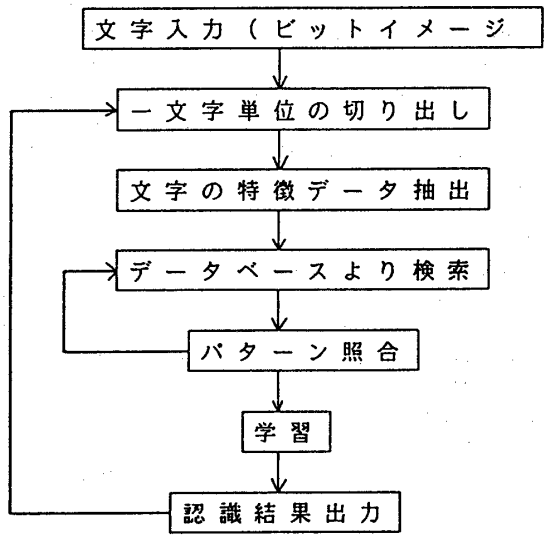

図２　処理の流れ

今後の課題

文字の切り出しについては、文字を、空白を持って切りだし、一文字として識別しているのだが、文字と文字が十分離れていない時、空白がないときには、一文字として切りとることができない。このような、多種多様な書式にも、対応できる様にしていくことが必要である。

検索方法については、基準の文字パターンから大きく外れる（字体そのものが極端に異なる場合。）と、誤差の範囲だけでは処理しきれず、認識率が著しく低下してしまうので、多種字体へ対応できるシステムにする必要がある。