つぶれ文字を認識するための一方法

電子制御工学科　　志久　修

ご存知のようにＯＣＲ（光学式文字読取り）とは，紙にかかれている文字を認識し，その文字に対応する電子テキストを出力する装置です．読み取る対象に応じ，活字ＯＣＲ，手書き文字ＯＣＲ，帳票ＯＣＲなどがあります（ちなみに本校の授業評価アンケートの読み取りに使っているのは帳票ＯＣＲです）．これらのなかで活字ＯＣＲは速度･認識性能とも非常にすばらしく，さらに低価格で十分実用レベルにあると思います．私が使っている活字ＯＣＲソフトはスキャナのおまけとしてついてきたものですが，それでも文字サイズが１０ポイント以上で，文書のレイアウトが単純で，プリンタから直接出力された高品質な文書なら１００％近い精度で電子テキストに変換してくれます．

しかし，活字ＯＣＲにもまだまだ弱点があります．例えば，複雑なレイアウトの文書や図表が混在した文書がうまく読めないし，コピーやファックスで品質が落ちた文字も正しく読むことができません．特に品質が落ちた文字が読めないのは致命的な弱点となっています．というのも，活字ＯＣＲには電子テキストが残っていない古い文書の読み取りに大きな期待が寄せられていますが，このような文書にはコピーやファックスされたものも多く含まれており，さらには紙自体の劣化により，文字の品質が悪くなっているからです．これらの品質が悪い文字を高精度に読み取ることができるようにすることが，活字ＯＣＲの重要な研究目標となっています．

今回は品質が悪い文字の例として，つぶれた文字を対象とします．つぶれ文字対策として，①つぶれた文字を認識のための標準文字パターンとして登録する方法，②文字のつぶれた部分は認識に使わず, きれいな部分だけを使う方法, ③つぶれの影響を受けにくい特徴を用いる方法, などがあります．ここで述べるのは，③のアプローチにもとづくものです．

文字認識方法としてこれまでたくさんの方法が開発されていますが，今回は文字を線図形として表現する方法を採用します. この理由は, 郵政研究所が行った文字認識方法の大規模な比較実験により, 文字を線図形として表現する方法の有効性が明らかにされているためです．

文字を表す線図形の代表例を図１に示します. 図(a)はもとの文字画像で, 図(b)は文字線の中心線（方法Ａ），図(c)は輪郭線（方法Ｂ）です．中心線は文字線の太さに依存せず，また輪郭線は文字の外形が表現できる利点がそれぞれあり，文字認識によく用いられています．

しかし，図２(a)のようなつぶれ文字に対しては，中心線が正しく求められない(図２(b))，内部の輪郭線が求めらない(図２(c))，などの問題があります．その対策として，図２(d)のようにつぶれていない部分は中心線を使い，つぶれた部分はその輪郭線を用いる方法（方法Ｃ），図２(e)のように文字の輪郭線とつぶれ部分の輪郭線を用いる方法（方法Ｄ）が考えられます．方法Ｃは東北大から提案された方法で，方法Ｄは今回比較する方法です．

図１　中心線（方法Ａ）と輪郭線（方法Ｂ）

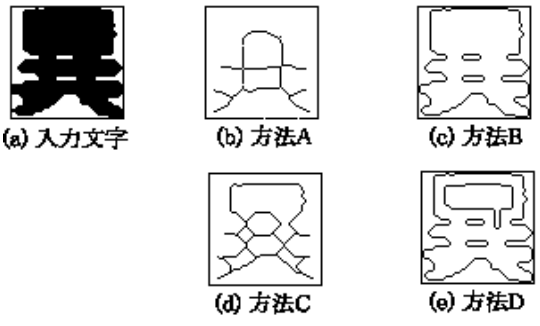

図２　つぶれ文字に対する４つの方法

以上の４つ方法を用いて実験を行いました.実験では図３のような品質を変えた文字画像を用意し，品質毎に文字認識率を調べていきました．詳細な実験条件は表１のとおりです.

図４に実験結果を示します．図４では品質を表す尺度として，文字画像を 64×64 画素で表したときの文字の平均線幅を使っています．線幅が太くなるほど文字のつぶれが生じ，逆に線幅が細くなるとかすれが生じていることを表しています．図４より線幅 8 程度で認識率がピークとなり，つぶれやかすれがひどくなるに従い認識性能は低下しています. つぶれ文字（線幅が太い場合）に注目すると，４つの方法とも認識率は低下していきますが，方法Dが他の方法に比べ若干低下の割合が小さくなっています．さらにかすれ文字（線幅が細い場合）に対しては方法ＢとＤは同じ認識率であることがわかります．以上の結果より，方法Dはつぶれ文字の認識に若干効果があり，さらにかすれ文字に対しては悪影響を与えないことがわかります.

ここで述べた簡単な改良で若干の性能向上が得られますが，さらに文字品質に対して頑健な認識方法の開発が必要だと思います.

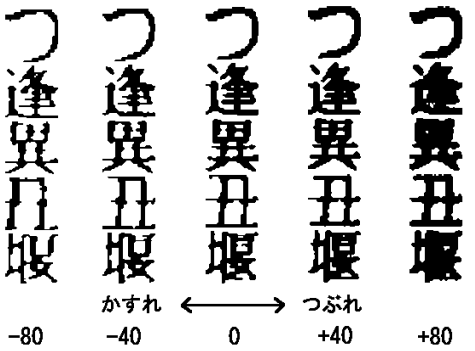

図３　文字データの例

表１　実験条件

文字データ	文字種 3,169 字種（ひらがな･カタカナ･数字･アルファベット･漢字） 学習パターン数 30 パターン／1 文字　（プリントアウトした文字） 評価パターン数：6 パターン／１文字　（コピーした文字）
特徴量	方向線素特徴量（196次元）
識別方法	部分空間法

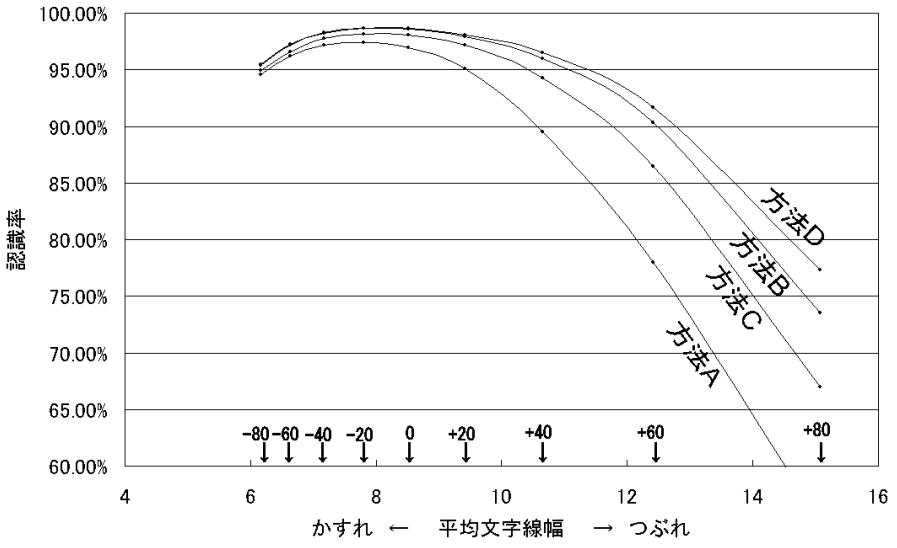

図４　実験結果