モンゴル文字認識システムの研究開発*

松尾崇史** 田口浩太郎*** 武富　敬****

Development of Mongolian Character Recognition System

Takafumi MATSUO, Kotaro TAGUCHI, Hiroshi TAKETOMI

1．はじめに

文字認識は，画像中の文字を、コンピュータで処理可能なテキストデータに変換する技術である。英語や日本語の文字認識研究は古くから行われ，現在では高精度での認識が可能である。しかし，世界には機械認識の方法が確立されていない文字が数多く存在する。それらの文字は，独自の表記規則などによって，文字認識における障壁が生じているといえる。その一例として，モンゴル文字が挙げられる。モンゴル文字は，現在でも使用されているにもかかわらず，文字認識の方法は未だ確立されていない。

そこで，本研究では，このモンゴル文字を対象とした文字認識システムの開発を目指している。本稿では，まずモンゴル文字の特徴や表記規則について述べ，それを踏まえて開発した認識システムの評価実験について，結果と考察を述べる。

2．モンゴル文字の特徴[1]

2．1　表記

モンゴル文字で書かれた文書の一部を図1に示す。モンゴル文字は縦書きで，左から右に改行する。また，この文字は続け字であるため，文字が単語ごとに線で繋がっている。そのため，文字同士の境界の判別が困難になる。

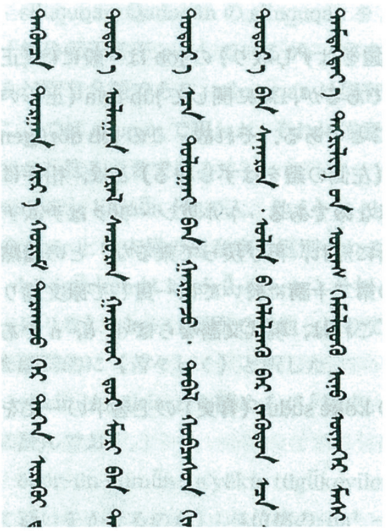

図1　モンゴル文字の文書

2．2　音と文字の対応

表1に，モンゴル文字の要素数を示す。モンゴル文字は，音と文字が1対1で対応していない。つまり，同じ音でも，単語中のどこに現れるかで，語頭形，語中形，語末形の3つの文字を使い分ける。母音の独立形とは，その母音1字で単語を構成するときに用いる形である。子音における中間語末形は，単語の途中で文字が途切れる時に使われる形である。また，違う音でも同じ形の文字を用いる部分もある（図2）。これらを形状のみから区別することは難しいため，表記規則を考慮する必要がある。

表1　モンゴル文字の要素数

	子音	母音	数字	記号
語頭形	27	7	10	6
語中形	30	10		
語末形	13	6		
中間語末形	6	-		
独立形	-	8		

* 原稿受付　平成22年9月24日

** 佐世保工業高等専門学校　専攻科
電気電子工学専攻
*** 熊本工業高等専門学校　電子情報システム工学専攻
**** 佐世保工業高等専門学校　電子制御工学科

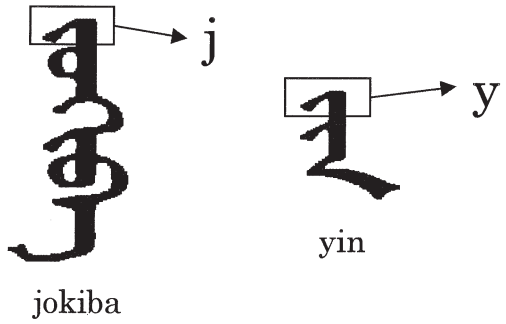

図２　同じ形状で音が異なる子音

２．３　母音の抱え込み

モンゴル文字は，文字を線でつなげて表記されるが，ある一定の音節のときは，子音の中に母音が入り込むような形で表記される。これを“抱え込み”と呼ぶことにする。その一例を図３ に示す。

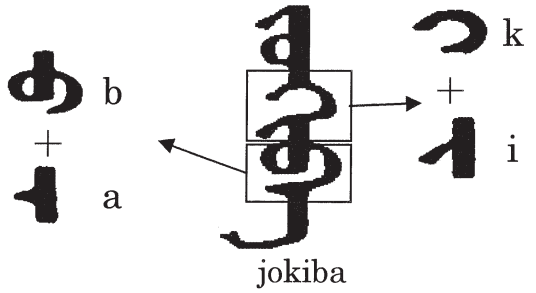

図３　母音の抱え込み

３．モンゴル文字認識システム

開発したモンゴル文字認識システムの概要を図 4 に示す。従来の文字認識では，周辺分布により文字領域を切り出して，その文字を認識するという方法が一般的である。しかし，モンゴル文字は前述の通り，文字が線で連続しているため，周辺分布で文字を切り出すことは難しい。そこで，本システムでは，文字を連結している中央の線を消去し，その後で切り出しを行う。

ここで切り出されるものは文字とは限らず，ばらばらになった“文字のパーツ”が切り出される場合もある。そのため，切り出したものは一旦“文字構成パターン”と見なして認識し，その後，モンゴル文字の表記規則を用いた組み合わせ処理を行うことで，文字として出力する。

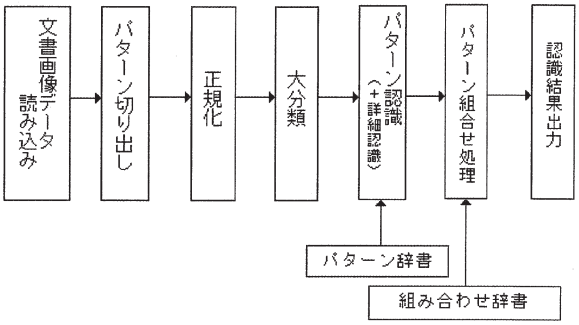

図４　モンゴル文字認識システム

３．１　文字構成パターン切り出し

文字を繋げている中央線を検出するには，単語ごとに垂直方向の黒ラン数の最大値をとり，その変化量を見る。変化量が急激に大きくなる部分を中央線の左端とし，変化量が急激に小さくなる部分を中央線の右端とする。その後，水平方向に周辺分布をとって文字構成パターンの境界線を定める。最後に，元の単語画像と，境界線を重ね合わせることで文字構成パターンを切り出す。この一連の流れを図 5 に示す。

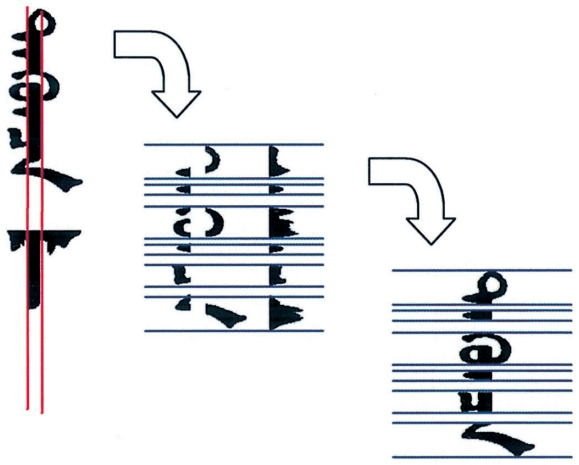

図５　中央線の消去

３．２　正規化

正規化処理では，文字構成パターンの縦と横の比に応じて３種類のサイズに分ける。これは，文字構成パターンには縦横比が様々なものが存在するためである。これにより，文字構成パターンの特徴を保持できるとともに，分類による比較回数の削減にもなる。正規化のサイズを表 2 に示す。

表 2　正規化サイズ

縦横比	正規化サイズ
~ 1.15	64 × 96
1.15 ~ 1.18	64 × 64
1.18 ~	64 × 128

※縦横比＝横サイズ / 縦サイズ

3．3　大分類

大分類処理では，簡単に抽出できる特徴を用いて，文字構成パターンを大まかに分類する。これにより，形状が似たパターンをカテゴリで分けて，誤認識の可能性を減らすことができる。また，サイズ分類と同様に，比較対象のパターンが減ることで高速化にもつながる。

大分類処理に用いる指標は，中央線の左側に出る線の数，中央線右側の画素の有無，閉ループの数の 3 つである（図 6）。この大分類処理と，サイズ分類を合わせて，文字構成パターンは 23 カテゴリに分類され，最大で 18 パターン，最小で 1 パターンのカテゴリが存在する。

図 6　大分類に用いる指標

3．4　文字構成パターン認識処理

文字構成パターンの認識処理では，パターン画像から特徴量を抽出し，特徴量を成分とした特徴ベクトルを，辞書ファイルの特徴と比較し，最も類似度の高いものを認識結果とする。特徴量には 4×4 のメッシュ特徴量，識別関数に単純類似度（コサイン類似度）を用いる。特徴ベクトルの次元数は，正規化のサイズに応じて異なる（表 3）。

表 3　特徴ベクトルの次元数

縦横比	正規化サイズ	ベクトル次元数
~ 1.15	64 × 96	384
1.15 ~ 1.18	64 × 64	256
1.18 ~	64 × 128	512

酷似した文字構成パターンについては，メッシュ特徴だけでは誤認識することがあるので，別途に詳細認識を行う。詳細認識では，文字構成パターンの局所的な特徴をとって，類似したもののうち，どのパターンに該当するかを定める。

3．5　文字構成パターン組み合わせ処理

認識した文字構成パターンは，そのまま文字になるものがほとんどだが，隣接する文字構成パターンと組み合わさって文字を成すものもある。したがって，それらを組み合わせて文字とする処理を行う。

まず，抱え込みなど，1 つのパターンに複数の文字構成パターンが含まれるものに対して，分解処理を行う。この対応付けは外部ファイルから読み込む。

そして，組み合わせる際は，基本的に 1 音節が子音と母音の 1 セットになることを前提に組み合わせる。ただし，例外として，以下のような 12 通りの音節が存在する。これらを十二字頭[2]と呼び，これらに限っては，子音の後に子音が続くことが許される。

1. 母音のみ
2. 母音＋母音 i
3. 母音＋子音 r
4. 母音＋子音 n
5. 母音＋子音 ng
6. 母音＋子音 G
7. 母音＋子音 s
8. 母音＋子音 d
9. 母音＋子音 b
10. 母音＋母音 u
11. 母音＋子音 l
12. 母音＋子音 m

4．認識実験

4．1　文字構成パターン認識実験

これまでに述べた方法をもとにして，文字構成パターンの認識実験を行った。入力はスキャナから読み込んだモンゴル文字の文書画像 1 枚を用いる。結果を表 4 に示す。

表 4　文字構成パターン認識実験結果

入力パターン数	728
誤認識パターン数	26
文字パターン認識率	96.4%

４．２　文字認識実験

文字構成パターンの組み合わせも含めた，文字認識実験の結果を表５に示す。

表５　文字認識実験結果

入力文字数	745
誤認識文字数	70
文字認識率	90.6%

５．考察

５．１　文字構成パターン認識実験の考察

文字構成パターンの認識では，中央線の誤検出による誤認識が多かった。中央線の誤検出の原因は，スキャナからの画像の傾きであった。これにより，文字構成パターンがうまく切り出されないだけではなく，大分類処理において誤分類が起こることもあった。

これに対する改善策として，ハフ変換[3]を用いることが有効であると考えられる。ハフ変換は，画像中の直線検出に用いられるものであり，全ての黒画素について直線式を計算するため，計算量が多い。しかし，モンゴル文字の中央線はほとんど鉛直方向になっているため，走査する角度の範囲を鉛直方向から＋5°～－5°ほどにすれば，計算量を減らすことができる。

５．２　文字認識実験の考察

文字認識実験で生じた誤認識の文字を，原因別に分類したものを表６に示す。

表６　誤認識文字の原因別分布

文字構成パターンの誤認識	24
a またはｅと G, n の混同	19
ng と(a+k, e+g)等の混同	12
(u+n)と d の混同	6
b と u の混同	4
y と i の混同	3
その他	2

表６から，最も誤認識の多かった原因は，文字構成パターンの誤認識による 24 字であることが分かる。このことから，文字構成パターンの認識率を100%にすれば，およそ 94%まで，文字認識精度の向上が見込まれる。

また，組み合わせ処理でａやｅの母音と，ｑやｎの子音などのように，同形文字のなかでも，組み合わせ次第で母音にも子音にもなりうる文字構成パターンを混同して誤認識することが多い。これらについては，新たに別処理が必要となる。

ｂとｕの混同では，中間語末形の後に１字だけ現れるｕを，同形のｂと誤認識していた。これに対しては，単語中の文字構成パターン数で対処することができると考えられる。

６．今後の展望

システム全体を通しての実験の結果，中央線を消去する方法で 90%を超える認識率を挙げられることが分かった。中央線検出は，文字構成パターン認識処理だけではなく大分類処理にも大きく関わっている。そのため，本システムを実用的なレベルにするには，中央線の検出処理において高い精度が求められる。具体的には，文字構成パターンの誤認識なども考慮に入れると，99%以上の検出率が必要であると考えられる。

今後の展望としては，マルチフォント認識への発展が期待される。マルチフォント認識では，フォントにより形状の違いが生じるため，メッシュ特徴による認識は難しくなる。そのため，線の方向や交点などから特徴を解析する，構造解析手法をとることが望ましい。例として，細線化処理を施した後に，方向線素特徴を抽出するという方法が考えられる。

参考文献

[1] 小沢重男：“蒙古語文語文法講義”，大学書林（1997）

[2] 高橋まり代：“検証（続）”，言の葉，2003-5-1, http://mariyot.ld.infoseek.co.jp/confirm2.htm（参照　2010-09-15）

[3] 高木幹雄，下田陽久：“新編　画像解析ハンドブック”, pp.1255-1256,　東京大学出版会,(2004)