ウイグル文字認識に関する研究*

4N−01

庫尓班江　肉孜†　矢口 博之　市野 学‡
東京電機大学§

1　はじめに

ここ20年のコンピュータ・テクノロジーの発展と普及等がだんだん中国の新疆ウイグル自治区にも影響を及ばしている．文字認識の研究は日本と米国では古くから研究され，実用化の進んだ分野である．これに対して，ウイグル文字認識の研究はまだ始まったばかりである[1]．ウイグル文字はアルタイ語系で，文字の構造はアラビア文字に非常に似ている．従ってウイグル文字認識の研究はウイグルだけでなくアラビア文字を使っている他の国々にとっても非常に重要な意味があると考える．本論文では，ウイグル文字認識システム作成への第一段階として，ウイグル文字の構造を分析し，ウイグル文字データベースの構築を行う．

2　ウイグル文字の特徴

ウイグル文字は中国の新疆ウイグル自治区等で使われている文字で，字種は32個である．但し，位置によって形が変化し，124 個のアルファベットをもっている(図1)．以下はウイグル文字の特徴について述べる．

（1）　ウイグル文字は右から，左にかけて一つの水平線によって続け書きするのが一般的である．ここではこの水平線を基本線と呼ぶことにする(図2(①, ②))．

（2）　一つの基本文字は位置によって，二つから四つまでの変化型がある．すなわち字首型(HEAD)，字中型(MIDDLE)，字末型(TAIL)と独立型(ISOLATED)がある．単語あるいは音節の始まる位置で，出現するときには，字首型を使う（ شادلىق ）．字首型は次のアルファベットと続けて書かれる．単語あるいは音節の中間で出現する時には，字中型を使う（ ئىشتىن ）．字中型は前後のアルファベットと続けて書かれる．単語あるいは音節の末で出現するときには，字末型を使う（ ئىش ）．字末型は前のアルファベットと続けて書かれる．一つのアルファベットが一つの単語あるいは一つの音節を構成するときは，独立型を使う（ يالاش ）．独立型は前後のアルファベットと独立して書かれる．

T	M	H	I	No	T	M	H	I	No
ق	ﻗ	ﻗ	ق	17	ﺋﺎ			ئا	1
ك	ﻛ	ﻛ	ك	18	ﺋﻪ			ئە	2
گ	ﮔ	ﮔ	گ	19	ﺑ	ﺒ	ﺑ	ب	3
ﯽ	ﯿ	ﯾ	ي	32	ف	ﻓ	ﻓ	ف	16

図1　ウイグル文字アルファベット表

（3）　一つの単語は一つあるいはいくつかの音節から構成される(図2(a, b, c))．

---

*Recognition of off-line Uighur characters
†Kurbanjan Rozi
‡Hiroyuki Yaguchi, Manabu Ichino
§Tokyo Denki University

（4） 124 個のアルファベットのうち，92 個のアルファベットはメインストロークに加えて，1～3個の点や短い曲線をセカンドストロークとして有する．セカンドストロークはメインストロークの上下あるいは内部に配置される．

（5） 字種によって文字幅にかなりの違いがある（図3）．

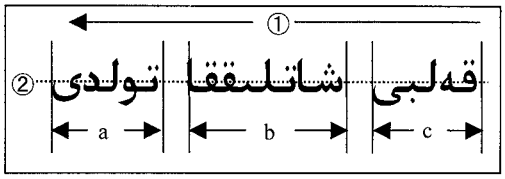

図2 ウイグル単語の構成
①：書き方向 ②：基本線 a：5 のアルファベット
b：8 のアルファベット c：5 のアルファベット

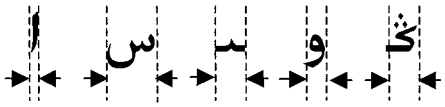

図3 文字幅の違い

3 データベースの構築

文字認識の研究では文字データベースが非常に重要である．日本語の場合は電子技術総合研究所が構築したETL8,9 文字データベースを使用して研究が進められている．しかし，ウイグル文字データベースはまだ構築されていない．そこで，本研究では手書き文字と印刷文字からウイグル文字データベースを構築した．

ウイグル文字を良く観察すると（図 4）で示すように右のアルファベットの ﺋ 部分と左のアルファベットの ﺌ 部分がほぼ同じであることがわかる．そこで，ﺋ 部分と ﺌ 部分を二つの補助アルファベットと決めることで，全部で 20 個のアルファベットを減らすことができる．従来の 124 個のアルファベットが 106 個のアルファベットになる．以下 106 個のアルファベットからデータベースを構築する．手書き文字認識の研究で使用するために，ウイグルの中学生 50 人に書いてもらった文字から各字種 50 サンプル，合計 5300 サンプルを収集した．その他ウイグル印刷文字で良く使っている 9 つのフォントから合計 954 サンプルを収集した．構築方法はまず，書いてもらった文字とウイグルで出版された本や雑誌に印刷された文書をスキャナーを使って画像データ（解像度 300dpi）として読み取る（図 5）．読み取った画像から 106 個のアルファベットを一つ一つ切り取り，各字種を 2 値化し，64×63 ドットで登録する．

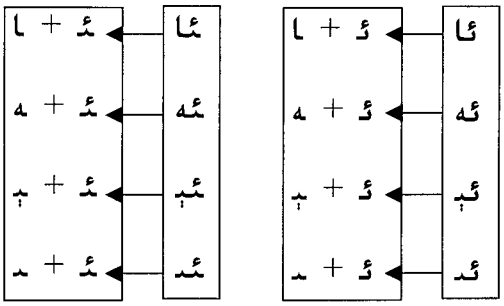

図4 ﺋ と ﺌ の補助アルファベット

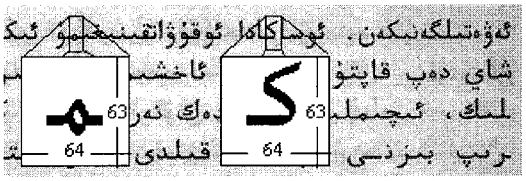

図5 ウイグル文字データベースの構築

4 まとめ

本研究では，ウイグル文字認識の研究の第一段階として，ウイグル手書き文字データベースと印刷文字データベースの構築を完成した．

今後は，文字の切り出しそして文字認識について更に検討を行う予定である．

文 献

[1] アニワル イミン，川島 稔夫，青木 由直："階層的手法を用いた手書きウイグル文字認識"，信学論(D-II)，j78-D-II，12，pp. 1787-1793(1995)．

[2] 庫尔班江 肉孜，市野 学："印刷ウイグル文字認識に関する研究"，中国新疆ウイグル自治区海外学者学術交流会(1999)．