2P－5 文字枠コード法およびペリフェラルパターン法を用いたタイ語文字認識

パニダー アナンパッタラチャイ†, 林 俊成†, 成田 誠之助†, Nucharee Premchaiswadi‡

早稲田大学理工学部電気電子情報工学科

1 はじめに

タイ語文字はパターン数こそ76文字と少ないものの，酷似している文字が多く，文字認識の分野においては大変困難な文字セットとして知られている．これら類似しているパターンを分類する際，文字のわずかな変形や文字線幅の変動などは，誤認識の主な原因となる．そのような文字の文字認識に適用できる手法としては，漢字の文字認識の粗分類手法としても用いられる文字枠コード法およびペリフェラルパターン法が挙げられる．これらの手法は，線幅の変動や多少の変形にもかかわらず，安定した分類をすることができる．

2 タイ語文字の特徴

文字の特徴について，日本語，タイ語と英語を表1で比較した．比較的認識率の高い英語，日本語と認識率の低いタイ語との大きな違いは，表1中，“文字の位置関係”と“基線から離れた文字”に見られる．

表1: タイ語文字の特徴

	日本語	タイ語	英語
文字の高さ	同一	可変	可変
文字の幅	同一	可変	可変
文字の位置関係	横	縦横	横
基線からの相対的位置	同一	異	異
基線から離れた文字	なし	存在	なし
文字の種類	多数	76	52
文字間の空白	あり	あり	あり

3 タイ文字データ

タイ文字は子音，母音，音調と記号及び数字から構成される．本稿では旧字を除いた76文字を扱うことにする．それらを図1に示す．

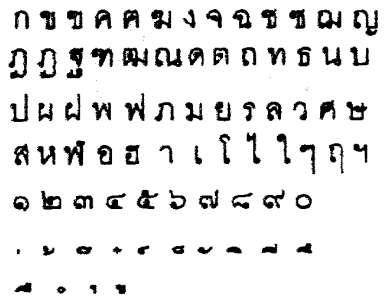

図1: タイ文字データ

4 文字分類処理

画像入力をして以下の処理手順を行う．

4.1 傾き補正

類似している文字が多く，少し傾いてもパターンマッチングを行う際に分類する事が困難になる．傾き補正には[1]の手法を用いた．

4.2 文字切出し

まず，ヒストグラムを用いて入力画像の黒画素濃度を調べ，一行ずつ切り出す．各文字の切り出しはラベリングで行う．

4.3 文字の大まかな分類

タイ文字はほとんどが一筆書きのようになるので，線が途切れることがない．この線の連続性を利用して文字枠コードを作成する．文字枠コードでは，まず文字枠の内側15%の矩形領域を四辺ごとに設定し，それぞれ矩形領域内の線にラベルをつけて，各矩形領域に含まれるラベルの長さを調べる．それらを長いラベルと短いラベルに分類して合計し，コードを作成する．例えば，一矩形領域に長いラベルが一つと短いラベルが一つあった場合，文字枠コードは11となる．長いラベルがな

†Panida ANANPATTARACHAI, Chunchen LIN, Seinosuke NARITA
†Waseda University. 3-4-1 Ohkubo Shinjuku-Ku, Tokyo 169
‡Department of Computer Engineering, Khonkaen University, Thailand

く，二つの短いラベルがある場合は02となる．これを四辺に利用するので，文字枠コードは8桁の数字になる．文字枠コードの例を図2に示す．

実験ではスキャナーからの入力を用いるため，位置のずれによって同じデータからでも出力ファイルに誤差が生じる．この誤差によって，文字の形が変わり，コードが変化してしまうことがある．この問題を解決するため，文字枠のサイズを深さ15% だけではなく10% と 20% も測定し，それらを組み合わせて分類した．その結果，全部で66種類のコードが発生した．ただし，同じ文字がいくつかコードを持つことになる場合もある．分類の例を図3に示した．ここでは基線の上下にある母音は別に処理するものとする．

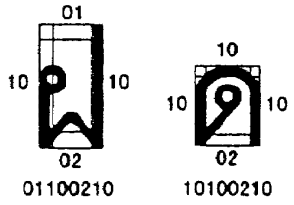

図 2: 文字枠コードの例

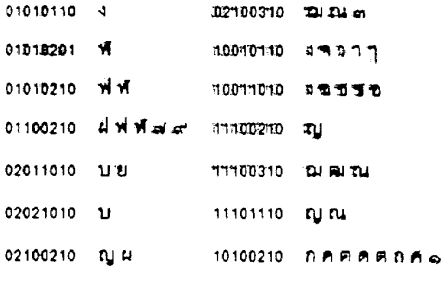

図 3: 分類の例

4.4 グループ分類

以上から，文字を大まかにいくつかのグループに分類することができた．しかし，グループの中には同一のコードを持つ異なる文字が存在するという問題点がある．この問題点を解決するために，ペリフェラルパターン法[2] を用いた．ここで各文字を矩形として上と下各面を四分割し，それぞれの面積を計算して[2] 中の類似度式を使うことにより分類を行った．

5 実験結果

基準となるデータは読み取り解像度400dpi，サイズ20pointから作成した．これらのデータ (図 4) は Association of computer in Thailand under patronage がタイ語文字認識を実験するために定義したデータセットに基づいている．基準と異なるサイズ，解像度に上記の分類手法を用いて分類を試みた場合，分類率は図 5と図 6のようになった．

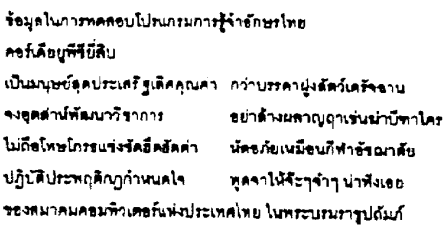

図 4: 基準データ

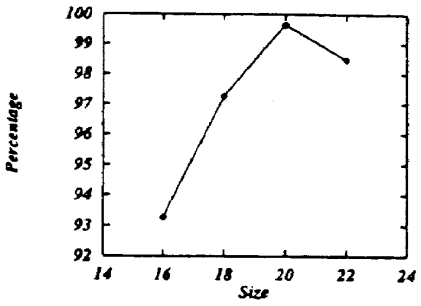

図 5: 解像度固定の実験結果

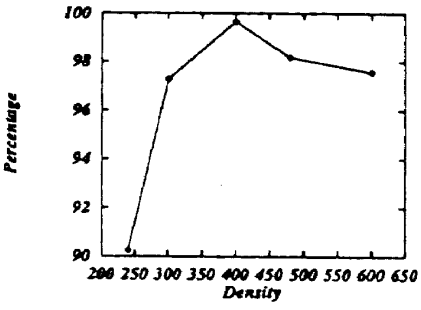

図 6: サイズ固定の実験結果

6 まとめ

基準データと同じサイズ，解像度の入力データを提案手法で分類した場合99.69% の分類率が得られた．誤差の原因としては，スキャナーのノイズや，二文字がデータの上では結合してしまっている場合があるためである．

参考文献

[1] 秋山 照雄，増田 功：書式指定情報によらない紙面構成要素抽出法，電子通信学会論文誌' 83/1 Vol.J66-D No.1

[2] 梅田 三千雄，有野 幸夫：粗いペリフェラルパターンよるマルチフォント印刷漢字の分類，PRL78-4 1978

[3] 森 健一，坂井 邦夫：2,000 字種を 100 字 / 秒で読む印刷漢字 OCR の開発，NIKKEI ELECTRONICS，1977 P.102-127