中国語オンライン手書き文字認識：日本語のオンライン手書き漢字認識との比較と認識性能・文字入力効率の改善

南部 元[†] 川又武典[†] 丸山冬樹[†]
依田文夫[†] 池田克夫[††]

中国における使用を想定して，中国版モバイルコンピュータ用オンライン手書き中国語文字認識ソフトウェアを開発した．この開発は，日本語版のオンライン手書き文字認識方式をベースにしており，基本的には，その漢字認識ソフトウェアを中国語向けに変更・追加したものである．一般に，中国文字と漢字は印刷字形の類似したものが多いので，「中国文字は字種が多い」程度の違いしかないと考えられがちである．しかし，印刷字形が同じでも，オンライン手書き文字パターンは中国人と日本人とでは異なることが多い．しかも，個人差が日本人より大きい．また，字種が多いだけに類似文字も多い．本稿では，中国固有の簡体字と伝統的な繁体字を合わせて8063字種を認識対象として検討した．文字パターン収集は上海で行った．簡体字のうち，使用頻度の高い3942字種を標準セット，残りの簡体字と繁体字を合わせて4121字種を拡張セットとし，前者を400人分，後者を50人分収集した．標準セットの文字データを分析した結果，正しい画数で書かれたものは約30%（漢字では60%以上），そのうちで正しい筆順で書かれたものは約80%であった．そこで，続け書き文字の認識性能向上に焦点を絞り，認識方式を改良した結果，標準セット文字の平均認識率89%，第10位分類率98%を得た．文字入力効率改善には，8万単語の情報を用いた単語知識処理による候補の絞り込みが有効であることが分かった．

On-line Chinese Handwriting Character Recognition: Comparison with Japanese Kanji Recognition and Improvement of Input Efficiency

HAJIME NAMBU,[†] TAKENORI KAWAMATA,[†] FUYUKI MARUYAMA,[†]
FUMIO YODA[†] and KATSUO IKEDA[††]

Most Chinese characters are thought to have the same shape as their Japanese Kanji counterparts. Although this is true for most printed characters, it is not true for handwriting characters. This paper first describes the differences between handwriting character patterns in Chinese and Japanese based on an investigation of 8,063 categories of Chinese character patterns collected in Shanghai. 3,942 categories of the basic set were written by 400 people and 4,121 categories of the optional set by 50 people. Second, it is shown that Chinese characters are difficult to be recognized by adapting our original on-line Kanji recognition algorithm, because of the variety in number of strokes, stroke order and shape. Only 30% of the characters were written in the correct number of strokes (compared with 60% for Kanji), and only 80% of the correct samples were written in the correct stroke order. The original recognition algorithm was modified and improved to handle cursive style handwriting. The experimental results show a recognition rate of 89% and a 98% recognition rate is achieved for the 10th candidate. For practical Chinese character input, much higher recognition rates seem necessary and it is shown that a Chinese word dictionary can be used effectively together with the character recognition algorithm.

† 三菱電機株式会社情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation
†† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

1. はじめに

ペン入力のモバイルコンピュータやPDAが業務用や個人用に普及してきたが，だれにでも容易に手書き文字データを入力できることが普及の大きな決め手で

ある．オンライン手書き文字認識技術はそのための重要な技術であり多くの研究が行われている[1]．しかし，日常の筆記速度で書かれた文字を高精度で認識することは容易ではない．手書き文字は，字形・画数・筆順に個人ごとの癖があり，筆記速度が上がるほど続け書きが増えて標準パターンからの変形が大きくなる．さらに，中国文字や日本語の漢字は字種が多いので認識の困難さはさらに大きい．中国語・日本語のオンライン手書き文字認識の研究は，最近は続け書き文字を対象としたものが増えている[2)~6)]．

筆者らはすでに，大局的整合法とストロークアナリシス法を併用した認識方式を漢字認識用に開発し，実用化した．本稿の中国文字認識方式はこれをベースに開発したものである．以下，2章では認識対象文字セットの特徴を，3章では印刷字体の違いを，中国文字と漢字の対比で論ずる．4章では手書き文字パターンの収集と文字パターンの分析結果を漢字と比べて述べる．5章では漢字認識方式で中国文字を認識した実験結果を，6章では認識方式の改良とその認識結果を述べる．7章では単語知識処理方式とそれによる単語入力効率の向上について述べる．

2. 認識対象文字

2.1 標準セットと拡張セット

中国文字には簡体字と繁体字の2つの種類がある．簡体字は中国国家標準（GB）文字セットで定められた“字形を簡略化した文字”で6763字種ある．繁体字は，中国では簡体字が制定されるまで使われていた“字形が複雑な文字”であり，非常に多くの字種があるが，一部は現在も姓・住所などで使用されることがある．

筆者らは簡体字のうち，新聞などに日常よく使われる3942字種を“標準セット”とした．残りの使用頻度の低い2821字種と姓・住所などでも使用される繁体字の1300字種を合わせて“拡張セット”とした．中国の新聞1年分のデータベースから求めた文字の出現頻度情報などを用いて調べた結果，簡体字の出現頻度は次のとおりであった．出現頻度の高い上位175字種でデータベース中の文字の約50%，995字種で90%，1424字種で95%，2397字種で99%，標準セットの簡体字全体の3942字種で99.69%をカバーしている．したがって，使用頻度の高い字種の認識率が高くなるような設計上の配慮が必要である．もちろん，出現頻度の低い字種でも誤りなく入力できる必要があるのはいうまでもない．

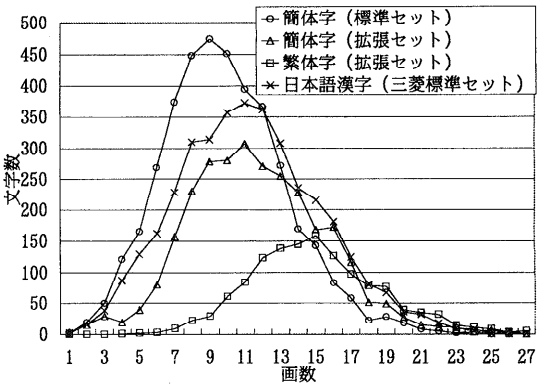


図1 画数分布（正規の画数）
Fig. 1 Categories with same number of strokes.

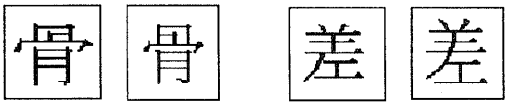


図2 同一のUnicodeで字形の異なる文字
Fig. 2 Different shapes with same Unicode.

2.2 文字の画数分布

対象とする手書き文字の画数分布を知ることは，認識方式を設計するうえで重要である．そこで，文字の正しい画数を基に，中国文字がどのような画数分布を構成しているかを調査した．**図1**に中国文字の各セットと三菱電機で使用している日本語漢字の標準セット（3693文字）の画数別の文字数分布を示す．

標準セットの簡体字は9画の文字が最も多い．拡張セットの簡体字は11画にピークがあり，漢字の分布と類似している．一方，拡張セットの繁体字は15画が最も多い．

3. 中国文字と日本語漢字との印刷字形の違い

Unicodeは中国文字と日本語漢字とに共通に使えるとされる文字コードだが，同一コードでも画数・字形・筆順の異なるものがある．**図2**に字形の異なるものの例を示す．手書き文字パターンの収集や認識方式改良の前に，このような違いを明確にしておく必要がある．

標準セットの簡体字3942字種のうち，1911字種が同一コードであったが，446字種に違いがあった．内訳は，画数違い140，字形違い186，筆順違い120．

拡張セットの簡体字2821字種では33字種が同一コードで，このうち10字種に違いがあった．内訳は，画数違い4，字形違い1，筆順違い5である．

拡張セットの繁体字1300字種では1085字種が同

一コードで，このうち215字種に違いがあった．

4. 手書き文字パターンの収集と分析

4.1 文字パターンの収集

オンライン手書き中国文字の認識方式・認識辞書の設計には大量の文字パターンが必要であるが，そのようなデータベースは入手できないので，1996年に上海で収集作業を行った．

標準セット3942字種は400人分（15～65才の男性259人，女性141人），拡張セット4121字種は50人分（20～60才の男性37人，女性13人）を収集した．そのうちの半分を認識システムの設計に，残りの半分を評価テストに使用した．前者を設計データ，後者を評価データと呼ぶ．ただし本稿では，使用頻度が高く，かつ収集した文字パターンが多い標準セットについてのみ評価結果を論ずる．

4.2 画数変動の分析

中国の人の文字筆記速度は日本人の漢字筆記速度よりもかなり速い．平仮名・カタカナがないことが一因だと思われる．このため，**図3**に示すように，複数のストロークが続け書きされることが多く，記入者によるが文字パターンの画数は正規の画数より小さくなることが多い（本稿では，このように複数のストロークが続け書きされた文字を“続け字”，続け書きにより形成されたストロークを“続け字ストローク”と呼ぶ）．

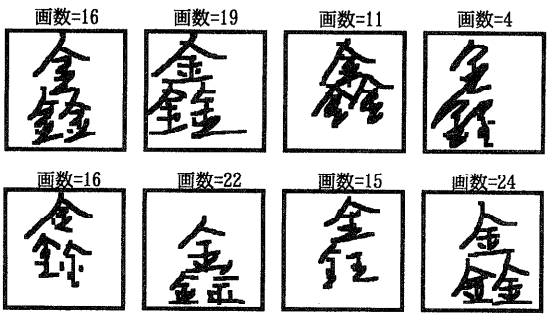


図3 手書き文字サンプルの例
Fig. 3 Examples of a sample category.

図4に，標準セットの設計データ（簡体字200人分）について，正規の画数との差に対する文字サンプルの分布を漢字データ（三菱電機および東京農工大[7),8)]）との対比で示す．正規の画数で書かれたのは30%（漢字は60%）であり，日本人に比べて続け書きがいかに多いかを示している．

4.3 筆順変動の分析

図5は，上記の正規の画数の文字サンプルについて，画数ごとに，正しい筆順で書かれた文字と間違った筆順で書かれた文字の比率を示す．平均で約20%が間違った筆順で書かれており，筆順変動の比率は画数が増すほど増えている．また，パターンを調べてみると同じ字種でも何通りもの筆順で書かれていた．この傾向は日本語漢字（東京農工大データ）でもほぼ同様である[9)]．

図6は筆順変動の内訳を画数別に示す．図中の筆順距離は“筆順が昇順でない個所の数”によって定義し

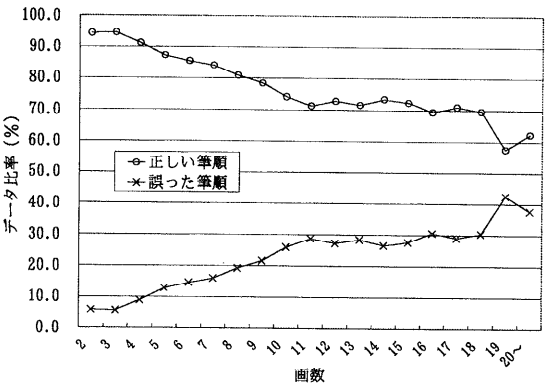


図5 筆順変動分布
Fig. 5 Correct and incorrect stroke order.

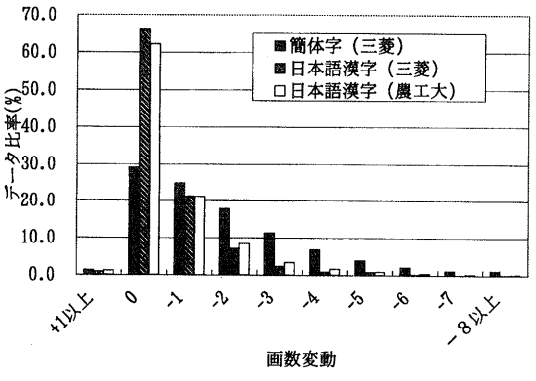


図4 画数変動分布
Fig. 4 Difference from correct number of strokes.

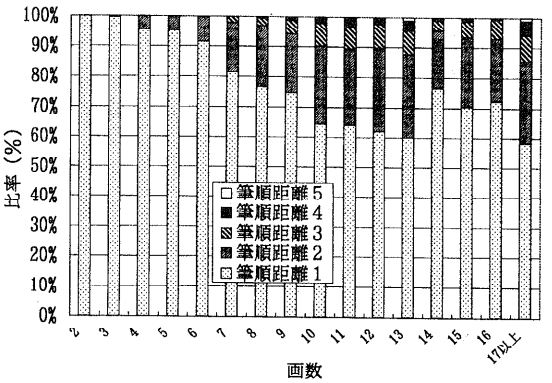


図6 筆順距離別の比率
Fig. 6 Breakdown of incorrect stroke order.

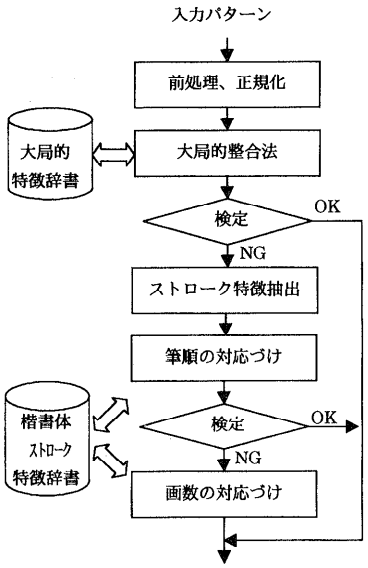


図7 認識処理のフローチャート
Fig. 7 Flowchart of Japanese Kanji recognition algorithm.

た．ほとんどの文字が筆順距離4以下で書かれており，認識方式の改良にあたっては，過度の筆順変動に対応する必要がないことが分かる．

5. オンライン漢字認識方式による認識実験

5.1 オンライン漢字認識方式

図4に示したように，日本語漢字はその60%以上が正しい画数で筆記されており（楷書体），これらの文字を高精度で読み取ることは非常に重要である．筆者らは文字パターンの2次元の形状を特徴として用いる大局的整合法と，ストローク特徴を用いて筆順の対応付け・画数の対応付けを行うストロークアナリシス法を併用した認識方式を開発し，特に楷書体パターンを高精度で読み取ることが可能なことを示した[10]．

図7に，この認識方式のフローチャートを示す．まず，タブレットにペンを降ろすときに発生することが多い“連続する同一座標点”を入力パターンの座標点列から除去する前処理，文字の幅と高さの大きい方を64画素のサイズになるようにする大きさの正規化を行った後，サンプル点の補完処理を行う．その後，筆順，画数に依存しない“文字の全体的な形状”を用いた大局的整合法により，候補文字の絞り込みを行う．次に絞り込んだ候補文字に対し，筆順変動を吸収する筆順の対応付けと，続け字などの画数変動を吸収する画数の対応付けを行う．

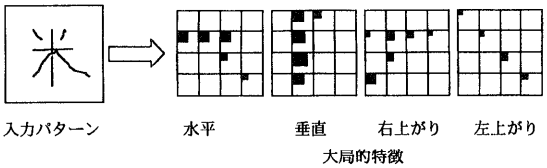


図8 入力パターンと大局的特徴の例
Fig. 8 Global features.

5.1.1 大局的整合法

前処理・正規化後の入力パターン中の各ストロークを一定間隔に分割し，分割した部分ストロークの方向成分（水平，垂直，右上がり，左上がり）を抽出する．次に，文字領域を 4×4 領域に分割し，各領域内に存在する入力パターンの方向成分の個数（4×4 領域 $\times$ 4方向 $=64$ 次元）を“大局的特徴”として抽出する[11]（**図8**）．得られた64次元の入力パターン特徴と，大局的特徴辞書中の各文字の標準パターン特徴からCity Block距離を求め，距離の小さいものから順に200個を候補文字として絞り込む．大局的特徴の標準パターンは，各文字について1つとし，設計データ200人分の平均パターンを用いている．

ここで，第1位の候補文字に関しては類似度を算出する．64次元の入力パターンベクトルを A，第1位の候補文字の標準パターンベクトルを T とすると，類似度 $S(A,T)$ は次式で求められる．

$$S(A,T)=\frac{\sum_{i=1}^{64}a_i\cdot t_i}{\sqrt{\sum_{i=1}^{64}{a_i}^2}\sqrt{\sum_{i=1}^{64}{t_i}^2}}.$$

類似度が所定の閾値より高い場合には，正解文字である可能性が高いと判断し，以降の処理は行わずに最終認識結果として出力する．比較的字形が整った文字は，筆順にかかわらず，大局的整合法だけで高速に認識できる．

5.1.2 ストローク特徴抽出

前処理・正規化後の入力パターンの字形情報は大局的整合法で抽出しているので，ストローク特徴としては，より識別能力が高い局所的な特徴を抽出する．具体的には，(1) ストローク形状（方向別の直線，折れ線，点など），(2) ストロークの運筆方向（ストロークの始点から終点への方向），(3) ストロークの大きさ（ストロークの外接矩形の幅と高さ），(4) ストローク間の相対的な位置関係（現在のストロークから次のストロークへの方向）を抽出する．

5.1.3 筆順の対応付け

抽出されたストローク特徴を用いて，大局的整合法により絞られた200個の候補文字について，ストローク特徴辞書中の標準パターンのストローク列と最も一致するように，入力パターンのストローク列を入れ換えることにより，筆順の対応付けを順次行う．辞書の標準パターンは，200人分の設計データ中の画数・筆順が正しいサンプルを用いて作成している．具体的には，画数・筆順が正しいサンプルを“ストローク形状列”で記述し，最も頻度の高いストローク形状列を標準ストローク列としている．さらに，標準ストローク列のクラスタに属するサンプルパターンを用いて，その他の標準ストローク特徴を求める．すなわち，ストロークの運筆方向とストローク間の相対位置関係については，最も頻度の高い方向を抽出する．ストロークの大きさについては，幅，高さ別に平均値を算出している．

一般に，入力パターンのストローク列と標準パターンのストローク列との対応付けをすべての組合せに対して行うと，対応付けの計算量が膨大になり実時間では処理できない．しかし，図6に示したように実際に生じる筆順変動は一定の変動範囲（筆順距離）に収まるため，実用上はすべての筆順変動を考慮した対応付けは無意味である．そこで，筆順距離を基に対応付けの組合せを制限することにより，実時間での処理を可能にしている．

本処理で筆順が対応付いた場合は，最終認識結果として出力する．

5.1.4 画数の対応付け

筆順の対応付け処理では，正しい画数で筆記された入力パターンに対する認識を行ったが，入力パターンが正しい画数で筆記されていない場合は，認識結果が得られない．そこで，入力パターンのストローク列と標準パターンのストローク列とをDP（Dynamic Programming）マッチングを用いて対応付ける．DPの評価値計算には，ストローク特徴を用いる．本処理により，入力パターンの画数と標準パターンの画数が異なる場合の対応付けが可能となる（**図9**）．対応付けの精度向上および処理時間短縮のために，ここでは筆順変動は考慮せず，さらに以下のような制約を設けている．

(1) 画数増加に関する制限
収集した文字サンプルの画数変動分布（図4）にも示すとおり，画数増加はわずかであり，これをプラス1画までに制限しても実用上問題はない．これにより計算量の削減が可能になる（図中の点線内および点線上のノードのみ計算）．

(2) 対応付けの制限
たとえば，入力パターンの1つのストロークが標準パターンの2つのストロークに対応付けられたときには，この2つのストロークのいずれかが入力パターンの別のストロークにさらに対応付くことは実際にはありえない．このように，ありえない対応付けを制限することにより，対応付け精度の向上が可能になる．

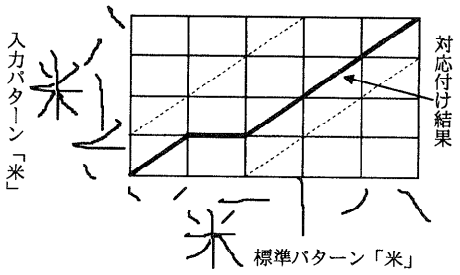


図9 DPマッチングによる画数対応付け（ストローク単位）
Fig. 9 Stroke number DP-matching.

表1 従来方式における分類率
Table 1 Recognition rates for learning samples and unknown samples (Kanji recognition algorithm).

設計データ		評価データ	
認識率	第10位分類率	認識率	第10位分類率
83.0	96.7	87.8	97.6

表2 処理方式別の対象パターン比率と分類率（設計データ）
Table 2 Recognition rates for learning samples using stroke order matching and stroke number DP-matching.

	処理対象データ比率	認識率	第10位分類率
筆順対応づけ方式	38.5	97.0	98.0
画数対応づけ方式	61.5	74.2	95.9

5.2 簡体字の認識実験とその結果

表1に標準セットの簡体字（3942字種）の設計データ・評価データ，各200人分に対する認識結果を示す．予想どおり，漢字認識方式（以後，従来方式と呼ぶ）では十分な認識性能が得られない．そこでその課題を明確にするために，筆順対応付け方式および画数対応付け方式の性能を個別に評価した．図7の大局的整合法で認識されなかった設計データのうちで，筆順の対応付けの段階で対象となったパターンの比率と分類率を**表2**の“筆順対応付け方式”に示す．また，画数の対応付けの段階で対象となるパターンの比率と分類率を表2の“画数対応付け方式”に示す．表2から分かるように，筆順対応付け方式の性能は，ほぼ実用的になっているが，画数対応付け方式の性能が十分でなく，続け字に対する画数対応付け方式を改良する必要があ

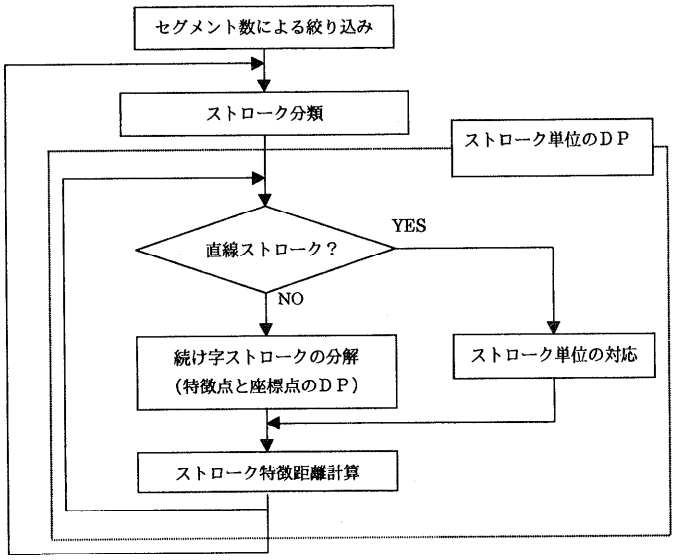


図10 ストローク分類による続け字対応付け方式
Fig. 10 Flowchart of stroke analysis for connected strokes.

ることが分かる．

6. 認識方式の改良とその結果

6.1 認識方式の改良

日本語漢字における続け字パターンの高精度認識方式として，筆者らはストロークをセグメントに分割し，方向コード特徴とストローク特徴を用いる認識方式を開発した[12]．この方式は，入力パターンを筆順どおりに一筆書きにし，屈曲点で区切られたセグメント列を抽出し，標準パターンのセグメント列とのDPマッチングにより対応付ける．その結果を用いて入力パターンを楷書体のストローク列に分割し，ストローク特徴で標準パターンとの距離を求め，最終的な認識結果を得る．

この方式は，日本語漢字における続け字の認識率向上に効果的であった．しかし，入力パターンを一筆書きにして対応付けを行うので，本来ストローク単位で対応付けが可能なストロークにもセグメントレベルの対応付けが行われる．このため，続け字に多い字形の変動があるとセグメント列も変動しやすく，対応付けに失敗するという欠点があった．そこで，続け書きされたストローク（続け字ストローク）に対してのみ，入力パターンの座標点情報を用いてストロークの分解を行うことにより，字形変動に強い続け字対応付け方式を新たに開発した．続け字対応付け方式のフローチャートを図10に示す．

6.1.1 セグメント数による絞り込み

画数変動を考慮したストローク対応付け方式においては，計算量が非常に大きくなるため，対応付けを行う対象文字の絞り込みが必要である．そこで，画数変動が発生したパターンにおいても安定なセグメント（ストロークの端点あるいは屈曲点で区切られた線分）の数を用いた絞り込みを行う．具体的には，設計データから求めた各文字のセグメント数の分布情報から，各文字のセグメント数の変動範囲を推定して辞書に用意し，入力パターンのセグメント数が変動範囲に収まっている場合に対象文字として選択する．

6.1.2 ストローク分類

入力ストロークを，a）直線ストローク，b）続け字ストローク，c）あいまいストロークの3種類に分類する．ここで，直線ストロークは，辞書の単一のストロークとの対応付けのみ可能とし，続け字ストロークは辞書の2つ以上のストロークとの対応付けのみ可能とし，あいまいストロークは両方の可能性があるストロークとする．次に，入力パターンのストローク列と辞書の標準パターンのストローク列との対応付けを，ストローク特徴を用いたDPにより行う．ただし，ストローク分類結果に基づき，直線ストロークおよび続け字ストロークの対応付けには上記の制限を設ける．また，直線ストローク以外の場合は以下の続け字ストロークの分解を行う．

6.1.3 続け字ストロークの分解

図11に続け字ストロークと判定された入力ストロークに対する，標準パターンの複数のストロークとの対応付けの例を示す．具体的には，入力ストロークの前処理後の座標点列情報と，標準パターンの複数ス

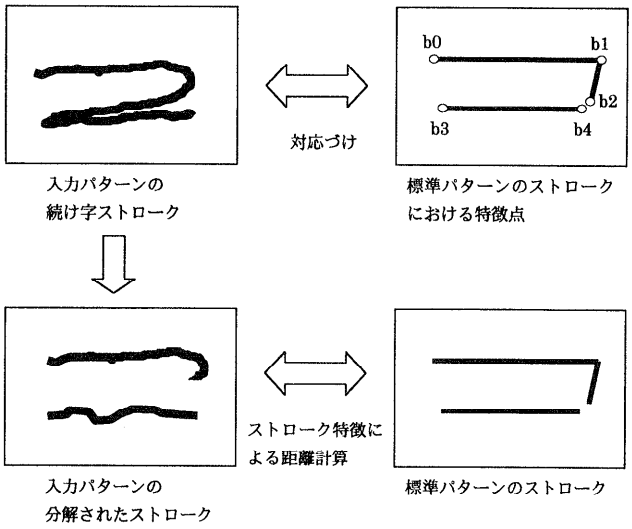


図11 DPマッチングによる続け字ストロークの分解
Fig. 11 Example of stroke analysis for connected strokes.

トロークから得られる特徴点（端点および屈曲点）列との対応付けにより，距離が最小となるものを選ぶ．

入力パターンのストロークにおける i 番目の座標点を a_i，標準パターンのストロークにおける j 番目の特徴点を b_j，a_i と b_j のユークリッド距離を $d(a_i, b_j)$，a_i と b_j までの累積距離を $g(i,j)$ とすると，$1 \leq i \leq I$，$0 \leq j \leq J$ に対して，次の漸化式が得られる．

$$g(i,j) = d(a_i, b_j) + \min\{g(i-1,k) | 0 \leq k \leq j\}.$$

ただし，$g(0,j) = d(a_0, b_j)$.

これにより，入力パターンのストロークと標準パターンのストロークとの対応付けの距離 $g(I,J)$ を，DPにより求めることができる．ここで，対応付けの経路情報（対応付いた a_i と b_j）を利用して，入力ストロークを楷書体のストロークに分解する．図の例では，特徴点 b_2 から b_3 に至る経路に対応する入力パターンの座標点列を消去して分解されたストロークを示している．

このようにして分解されたストロークのストローク特徴と，標準パターンのストローク特徴を用いて，ストローク特徴距離を算出し，6.1.2項で述べたDPにおける評価値とする．

6.2 評価結果

表3に標準セットの設計データ，評価データ各々に対する改良方式の認識結果を示す．設計データでは，認識率は83.0%から88.0%に，第10位分類率は96.7%から97.7%に改善された．また，評価データでは，認識率は87.8%から89.7%に，第10位分類率は97.6%から98.1%に改善された．

表3 改良後の分類率
Table 3 Improved recognition rates for learning samples and unknown samples.

	設計データ		評価データ	
	認識率	第10位分類率	認識率	第10位分類率
従来方式	83.0	96.7	87.8	97.6
改良方式	88.0	97.7	89.7	98.1

表4 画数対応付け方式改良の効果（設計データ）
Table 4 Effect of modified stroke number DP-matching for learning samples.

	認識率	第10位分類率
従来方式	74.2	95.9
改良方式	82.3	97.5

設計データより評価データの認識率が高くなった原因を次のとおり考察する．まず，設計データと評価データは無作為に分割されたが，結果的には評価データに正しい画数のサンプルがやや多かった．さらに，膨大な字種を対象としながら携帯端末への適用を考慮するために，ストローク特徴辞書設計には画数・筆順の正しい楷書体パターンの使用を基本とし，画数変動・筆順変動への個別対応を行った．しかし，設計データの中にも画数・筆順ともに変動するパターンがあり，標準テンプレートに反映できない未知サンプル的要素（評価データ的要素）が残ることを避けられなかった．この2点が主な要因である．その解決は今後の課題である．

表4に，画数対応付け方式改良の効果を設計データに対して示す．画数対応付け段階での認識率が74.2%から

82.3%へと大幅に向上したので，上記の認識率88%が得られた．

7. 単語知識処理による単語入力効率の改善

前章で述べたとおり，文字認識方式の高精度化だけでは，続け字の多いオンライン手書き中国文字を効率良く入力できるとはいえない．しかし，前記の中国の新聞文字データベース（1777万文字）によれば，新聞に出現する文字は重なり度数で87.3%が中国文字であり，残りの12.7%は数字，アルファベット，記号などである．三菱電機の中国語単語辞書（77,258単語）にヒットした“2文字以上単語”の構成要素となっている中国文字は37.1%である（残りの50.2%は1文字単語または未登録単語の構成要素）．このうちで，2文字単語の構成要素となっている文字は89%，3文字以上単語の構成要素の文字は11%であり，中国語の文章では“2文字以上単語”に占める2文字単語の割合が非常に高いことが分かった．そこで，中国語単語入力効率の向上を目的に，上記の中国語単語辞書を用いた単語知識処理方式を開発した[13]．

本章では，単語知識処理方式を用いた中国語入力システムの概要，標準セットの手書き簡体字パターンデータベースを用いた評価結果について述べる．

7.1 中国語入力システムと単語知識処理方式

図12に開発した方式の概要を示す．単語知識処理は，単語情報を用いることにより，文字単独では認識できなかった文字を救済し，単語としての認識率を向上させることを目的にしている．まず，単語の先頭2文字を手書き入力し，文字認識処理により各文字を認識した後，それぞれの文字の認識候補文字を最大10文字出力する．得られた認識候補文字を入力として，最大$10 \times 10 = 100$通りの単語を作成し，それぞれの単語が単語辞書中に存在するかをチェックし，存在する場合は単語知識処理結果として出力する．単語の距離値としては，各文字の候補順位の和を使用している．

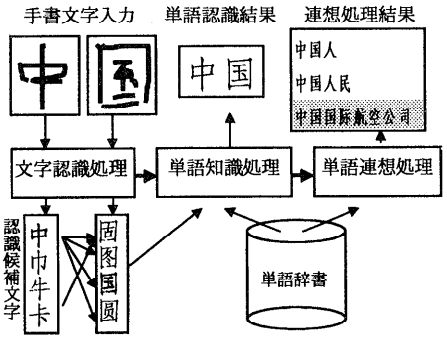


図12 中国語入力システムの概要
Fig. 12 Chinese word input system using Chinese word dictionary.

単語辞書中には，2文字単語だけでなく，3文字以上の単語も含めている．3文字以上の単語についても単語知識処理により認識率を向上させるため，先頭2文字を2文字単語として扱い単語知識処理を行う．たとえば，「中国人」という3文字単語は，「中国」という2文字単語として単語知識処理される．

次に単語連想処理では，単語知識処理により得られた単語から始まる3文字以上の単語を連想結果として出力し，3文字目以降の筆記の手間を省いている．

今回使用した単語辞書は，総単語数が77,258単語で，平均単語長は2.2文字，最長は12文字単語である．

7.2 単語入力効率の改善

文字を1文字ずつ独立して筆記した場合は，各文字の字形は前後の文字の影響を受けにくい．そこで，文字認識方式の開発用に収集した標準セットの簡体字パターンの評価データ（3942字種，200人分）を用いて，同一人物の筆記した任意の単語パターンを作成し，単語知識処理のシミュレーションを行った．単語辞書中の先頭2文字が異なる単語（63,979単語）についての評価結果を**表5**に示す．

単語認識率は第3位候補単語迄で96.1%という比較的高い値に収れんしている．エラー率（第1位候補単語が正解単語でない割合）は2.7%，リジェクト率（候補文字の2文字組みのすべてが単語辞書にヒットしなかった割合）は2.2%であった．単語知識処理をしない場合は，各文字について第10位までの認識候補文字の中から正解文字を探して入力する必要があるので，単語知識処理により単語入力効率が上がったことが分かる．また，システム設計上も候補単語の表示は数個でよいことが分かる．

単語知識処理で注意しなければならないのは，非単語（未登録単語を含む）を入力したときにリジェクトできない場合は，登録単語にヒットしてエラーになることである．しかし，オンライン文字認識を用いたモバイルコンピュータの用途は，データ入力やデータベース検索などの単語入力型の業務が多く，この場合は筆記者が単語であることを意識して記入するので，非単語が記入されることは少ない．また，未登録単語

表5 単語知識処理の効果（評価データ）
Table 5 Word recognition rates using Chinese word dictionary for unknown samples.

	1位	2位	3位	10位	エラー	リジェクト
1文字目認識率	90.0	94.1	95.5	98.0	10.0	0.0
2文字目認識率	89.9	93.9	95.5	98.0	10.1	0.0
単語認識率	95.1	96.0	96.1	96.1	2.7	2.2

が入力されないようにするには，用途別の単語辞書を設けて，適正な単語辞書容量でかつ未登録単語がないようにすればよい．

単語入力の処理時間に関する評価は，評価システムの表示画面などのヒューマンインタフェースおよび被験者の慣れなどの人的要素に大きくかかわる．新たな研究要素であるので今後の課題としたい．

8. おわりに

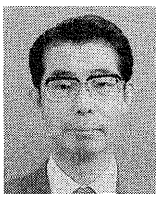
中国版モバイルコンピュータ用に，実用レベルのオンライン手書き中国語文字認識方式を開発した．上海で収集したオンライン手書き文字パターンを日本語漢字と比較して分析した結果，中国では正しい画数で書かれた文字が約30%しかなく，続け書き文字が多いことを確認した．日本語のオンライン手書き漢字認識方式をベースにして，続け書き文字の認識精度向上に焦点を絞って認識方式を改良した結果，標準セット文字データ（簡体字3942字種，400人分）の平均認識率89%，第10位分類率98%を得た．さらに，8万単語の単語情報を用いた単語知識処理を行うことにより，実用レベルの文字入力効率が得られることを確認した．今後の課題は，利用現場の情報をフィードバックして，より使いやすいシステムに改良することである．

謝辞 膨大な量の中国語オンライン手書き文字パターンの収集にご協力いただき，また有益な助言をいただいた中国上海交通大学語言文字工程研究所所長の楊恵中教授ならびにスタッフの方々に感謝いたします．また，本研究に関して，終始熱心な討議をいただいた京都大学大学院情報学研究科知能情報学専攻池田研究室の皆様に感謝いたします．

参考文献

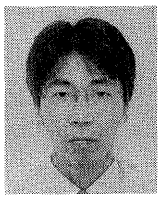
1) Tappert, C., Suen, Y. and Wakahara, T.: The State of the Art in On-Line Handwriting Recognition, *IEEE Trans. PAMI*, Vol.12, No.8, pp.787–808 (1990).
2) Chou, K., Fan, K. and Fan, T.: Radical-Based Neighboring Segment Matching Method for On-line Chinese Character Recognition, *13th ICPR*, Vol.III, track C, pp.84–88 (1996).
3) Wakahara, T., Nakajima, N., Miyahara, S. and Odaka, K.: On-line Cursive Kanji Character Recognition Using Stroke-Based Affine Transformation, *13th ICPR*, Vol.III, track C, pp.204–209 (1996).
4) Liu, J., Cham, W.K. and Chang, M.Y.: Stroke Order and Stroke Number Free On-Line Chinese Character Recognition Using Attributed Relational Graph Matching, *13th ICPR*, Vol.III, track C, pp.259–263 (1996).
5) Nakagawa, M., Akiyama, K., Tu, L.V., Homma, A. and Higashiyama, T.: Robust and Highly Customizable Recognition of On-line Handwritten Japanese Characters, *13th ICPR*, Vol.III, track C, pp.269–273 (1996).
6) Zheng, J., Ding, X. and Wu, Y.: Recognizing On-line Handwritten Chinese Character via FARG Matching, *4th ICDAR*, Vol.II, pp.621–624 (1997).
7) Nakagawa, M., Higashiyama, T., Yamanaka, Y., Sawada, S., Higashigawa, L. and Akiyama, K.: On-line Handwritten Character Pattern Database Sampled in a Sequence of Sentences without Any Writing Instructions, *4th ICDAR*, Vol.II, pp.376–380 (1997).
8) Nakagawa, M.: *TUAT Nakagawa Lab. HANDS-kuchibue d-96-02*, Tokyo University of Agriculture and Technology (1996).
9) 岡野，川又，依田：オンライン手書き文字データ（TUAT）の分析，信学会春季全国大会 (1998).
10) 川又，小川，岡野，亀代，南部，依田：大局的整合法とDPによるストロークの対応付けを併用したオンライン手書き文字認識，信学会春季全国大会 (1996).
11) 依田，小林，山本，南部：大局的特徴を併用したストロークマッチング法による手書き漢字認識の検討，信学技報，PRL82-30, pp.69–76 (1982).
12) 亀代，川又，南部，依田：方向コード特徴とストローク特徴を用いたオンライン文字認識方式，信学会春季全国大会 (1997).
13) 川又，丸山，南部，依田：中国語単語知識処理方式の開発，第55回情報処理学会全国大会論文集 (1997).

(平成10年12月25日受付)
(平成11年 6月 3日採録)



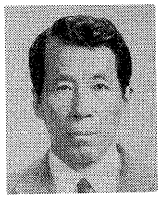
南部 元（正会員）

1941年生．1964年京都大学工学部電気工学科卒業．1966年同大学院工学研究科電子工学専攻修士課程修了．同年三菱電機（株）入社．電子計算機の内部記憶装置，文字認識システム，ペン入力PCの文字認識方式，ヒューマンインタフェース技術等の研究開発に従事．現在同社情報技術総合研究所主席研究員．この間1996年より京都大学大学院情報学研究科知能情報学専攻博士後期課程に在籍し，1999年3月研究指導認定退学．電子情報通信学会会員．



川又　武典（正会員）

1961年生．1984年山梨大学工学部計算機科学科卒業．同年三菱電機（株）入社．文字認識システム，印刷漢字認識，オンライン文字認識，筆者認識の研究開発に従事．現在同社情報技術総合研究所に勤務．



丸山　冬樹

1948年生．1971年東京工業大学電子工学科卒業．1973年同大学院電気電子専攻修士課程修了．同年三菱電機（株）入社．プロセス制御用計算機システムの開発，自然言語処理応用システムの開発，コーパス分析による英独仏日中韓の比較言語データベース構築の研究に従事．現在同社情報技術総合研究所に勤務．この間1976～1978年ミュンヘン工科大学プロセス計算機研究所客員研究員．FFT（高速フーリエ変換）専用機器開発に参加．言語処理学会会員．

依田　文夫（正会員）

1955年生．1978年東京工業大学電子工学科卒業．1980年同大学院情報工学専攻修士課程修了．同年4月三菱電機（株）入社．手書き・印刷漢字認識，画像処理，ニューラルネットワークの研究開発に従事．現在同社情報技術総合研究所に勤務．

池田　克夫（正会員）

1937年生．1960年京都大学工学部電子工学科卒業．1965年同大学院博士課程電子工学専攻単位取得退学．同年京都大学助手．1971年同助教授．1978年筑波大学教授（電子・情報工学系）．1988年京都大学教授（工学部，1998より大学院情報学研究科）．この間，1971年9月～1972年3月米国ユタ大学客員研究員，同年3月～8月マサチューセッツ工科大学客員研究員，1984年10月～11月スイス連邦工科大学客員研究員．高度の情報処理システムの構成に興味を持ち，コンピュータネットワーク，分散処理システム，マンマシンインタフェース，画像理解，文書画像理解の研究に従事．著書に，「コンピュータユーティリティの構造」（昭晃堂），「オペレーティングシステム論」（電子情報通信学会），「データ通信」（昭晃堂）等．工学博士．電子情報通信学会，人工知能学会，IEEE SM，ACM各会員．電子情報通信学会情報システムソサエティ前会長等．