高精度手書きOCR技術と第6回中国国勢調査への適用

Highly Accurate Handwriting OCR Techniques and Their Application to Sixth National Population Census in China

● 鄭　大念　● 孫　俊　● 于　浩　● 直井　聡

あらまし

高精度手書きOCR(光学式文字認識)は，とくに中国語や日本語など大規模カテゴリ言語においては実用化がいまだ困難である。著者らは，このたび修正線形判別分析，部分空間に基づく類似文字判別，複数分類器の結合，相互情報に基づく適応的棄却による高度な認識アルゴリズムを開発した。本技術は中国政府により2010年の第6回中国国勢調査(世界最大規模の国勢調査)に採用された。ここに示すように，開発したアルゴリズムは住所と民族名の既知情報を組み合わせることにより99％を超える精度と低棄却率を実現できる。漢字認識技術が大規模な中国国勢調査プロジェクトで使用されたのは今回が初めてである。

本稿では，第6回中国国勢調査に適用された富士通研究開発中心有限公司の高度な手書き文字認識技術について紹介する。

Abstract

Achieving highly accurate handwriting optical character recognition (OCR) is still a challenge in real applications, especially for non-Western languages like Chinese and Japanese. We proposed an advanced recognition algorithm using modified LDA, subspace-based similar-character discrimination, multi-classifier combination and mutual-information-based adaptive rejection. As an application, our technologies were adopted by the Chinese government in the Sixth National Population Census (the largest census in the world) in 2010. As shown in the paper, by combining these technologies with knowledge about addresses and nationalities, our algorithms can achieve an accuracy of over 99% with a low rejection rate. This is the first time that Chinese character recognition technology has been used in a large-scale Chinese census project. This paper will introduce Fujitsu Research and Development Center's highly accurate handwriting OCR techniques applied in the Sixth National Population Census.

ま え が き

中国は14億人という世界最大の人口を有する国である。最新の人口統計データを収集するため中国統計局（NBSC）によって10年ごとに国勢調査が行われてきた。第6回国勢調査は2010年11月に開始された。数百万人の国勢調査員が数十日をかけて国中を回り，調査用紙に記入された大規模な国勢調査データを収集する。国勢調査で得られた統計データは，社会と経済の発展に不可欠で大変貴重なものである。

この調査用紙を処理するため，過去の国勢調査では膨大な労力と費用が費やされたが，近年では装置（スキャナ，PCなどのハードウェアデバイス）の改良やOCR（光学式文字認識）技術によりこのコストは削減されている。手作業によるデータ入力は大部分が自動スキャニングや自動帳票認識などの自動機械処理に取って代わられている。富士通研究開発中心有限公司（FRDC）の高度な手書きOCR技術は，第6回国勢調査に適用された。調査用紙の例を**図-1**に示す。用紙1枚につき数字や漢字を数百文字識別する必要がある。数字と住所，民族名，名前の認識タスク（認識対象）を**図-2**に示す。数字認識タスクには住所コード，世帯人数，生年月日などの認識が含まれる。漢字認識タスクには民族名と住所，名前の認識が含まれる。大規模な中国の国勢調査プロジェクトで漢字認識技術が使用されたのは今回が初めてである。

国勢調査の実際の環境では手書きの品質が高いとは限らない。周知のとおり，OCR研究においては手書き漢字認識（HCCR）はいまだ困難であり，[1] 国勢調査においても低品質の手書き漢字の認識は大きな問題となっていた。FRDCでは，手書きが低品質である場合にも手書き漢字認識能力を向上

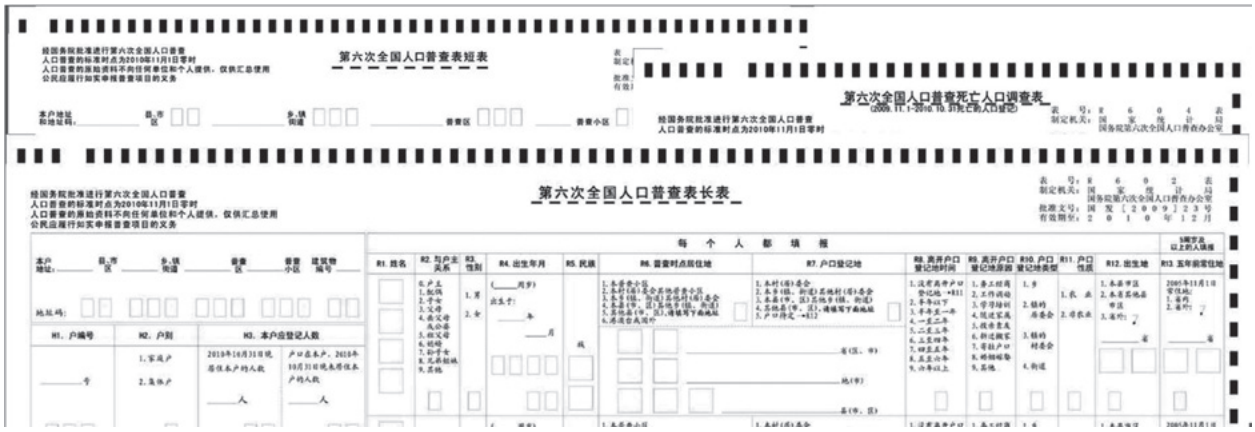


図-1　第6回国勢調査で使用された調査用紙
Fig.1-Several forms used in Sixth National Population Census.

（a）数字認識　（b）住所認識　（c）民族名認識　（d）名前認識

図-2　第6回中国国勢調査における手書き文字認識タスク
Fig.2-Handwriting character recognition tasks in Sixth National Population Census.
(a) numeral recognition; (b) address recognition; (c) nationality recognition; (d) name recognition.

させる技術として，以下の四つの技術を開発した。
・修正線形判別分析（MLDA：Modified Linear Discriminate Analysis）に基づく文字認識
・部分空間に基づく類似文字判別
・複数分類器の結合
・相互情報に基づく適応的棄却

本技術により，文字認識能力を向上させ，さらにタスクの既知情報と組み合わせることにより，著者らが開発した認識アルゴリズムは，国勢調査の厳しい要件を満たした。

本稿では，第6回国勢調査に適用された手書きOCR技術について紹介する。最初に手書き文字認識システムと技術の詳細を説明する。その後，本認識技術の評価を報告する。

国勢調査における手書き文字認識システム

● システムの概要

国勢調査の帳票処理システムでは，**図-3**に示すとおり以下の三つの大きな段階がある。
・調査用紙スキャニング
・帳票認識
・人の目による確認

調査用紙スキャニングでは，富士通のスキャナを使用して非常に高速で用紙を用紙イメージに自動変換する。つぎに，調査認識モジュールで様式の位置を特定し，手書き文字をすべて認識する。高精度を実現するためには目視確認の段階が必要となるが，認識結果が疑わしい場合のみ目視確認に回されるため，大幅な省力化となる。この疑わしい認識結果を検出する技術が認識棄却である。したがって，文字認識と棄却はいずれも重要性が高い。技術面での目標としては，人の目による確認を極力少なくしつつ，文字認識の誤り率を最低限に抑えることが重要である。

● 手書き数字認識

国勢調査用紙では，ほとんどの記入情報は数字である。数字認識精度は国勢調査データの大部分の品質において極めて重要である。本手書き数字認識では，サポートベクターマシン（SVM）[(2), (3)]に基づく数字分類器と複数分類器の結合[(4)]を応用し，認識能力を向上させている。最終的には99.9％を超える精度を実現し，これで国勢調査のデータ品質を確保している。

● 手書き住所認識

手書きの住所はすべて，省名単語と都市名単語，県名単語の三つのレベルの語で構成される。例えば「河南（省名単語）- 信阳（都市名単語）- 淮滨（県名単語）」などである。

最新の中国標準住所テーブルによれば，2863の標準住所がある。住所テーブルとユーザの要求により，省，都市，県を示す用語としてそれぞれ35語，345語，2895語（標準の県名単語，一部旧県名）を収集し，語彙（いい）とした。つぎに，省と都市，県の語彙から漢字51文字の省名辞書，350文字の都市名辞書，1162文字の県名辞書を作成した。住所の文字はすべてあらかじめ作成した四角い欄に記入するものとし，文字同士の区分の問題を回避した。

記入ミスは非常に多く，54 093件の手書き住所サンプルでは7.8％の記入ミスがあった。このようなミスには住所が不完全なものや省名の誤り，都市名の誤りや抜け，県名の誤りや抜けなどがある。

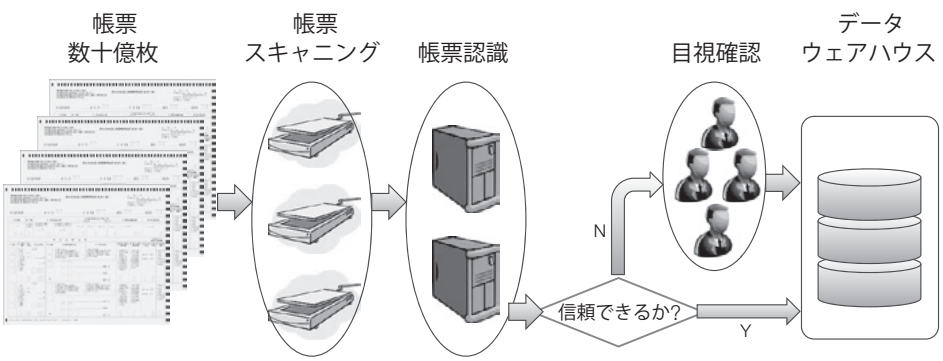


図-3　国勢調査における帳票処理のフロー
Fig.3-Flowchart of form processing in population census.

したがって，住所を示す3レベルの語はそれぞれ別個に処理する必要があるが，そうすれば3レベルの語の組合せから最も可能性の大きい住所の結果を推定することができる。

FRDCによる手書き住所認識解決法を**図-4**に示す。省名，都市名，県名の文字を並行して認識する。各文字について，関連辞書にあるすべての候補の認識信頼度を文字信頼度リストに格納する。各レベルについて，関連語彙にあるすべての候補の単語信頼度を計算して単語信頼度リストに格納する。ここで単語信頼度は文字信頼度の平均と等しい。つぎに，住所テーブルにあるすべての候補の住所信頼度を計算して住所信頼度リストに格納する。住所信頼度は，その省名単語，都市名単語，県名単語の信頼度の加重合計と定義する。重みは3レベルの記入文字数に比例する。住所信頼度は降順にソートし，最大信頼度を持つ住所コードを最終的な認識結果として出力する。

低棄却率で高認識率を実現するには，棄却ルールが大変重要である。第1位の住所候補を容認するか棄却するかを判断するため，県名単語の信頼度をユーザ定義閾値と比較する。信頼度が閾値よりも小さい場合は認識結果を棄却し，そうでない場合は容認する。この棄却ルールを，住所信頼度基準と第1位住所の単語信頼度最低値基準という，ほかの二つの棄却ルールと比較したところ，県名単語信頼度基準の棄却ルールが最も優れていた。

手書き住所の典型的なミスを**図-5**に示す。図-5（a）は標準の完全な住所で，本認識システムにより上位5位の住所と第1位の住所の県名単語信頼度が表示されている。図-5（b）は県名単語が抜けている不完全な住所である。図-5（c）は省名「湖北」と県名「洪湖」の間の都市名「荆州」が抜けている不完全な住所である。図-5（d）は，都市名が正しくは「东营」でなく「烟台」の住所不一致を示す。図-5（e）は県名を間違って都市名としている住所の誤りで，「青州」は県名，「黄楼镇」はその県にある町の名前である。図-5（f）は訂正のある住所で，この場合，訂正後の都市名「祈州」がなくても住所が正しく認識されている。図-5（b）～（f）より，本認識システムがこの実際の適用において典型的なミスを十分に処理できることが分かる。

● 手書き民族名認識

中国では14億の人が56の民族に分類されている。この民族名は「族」（Zu）という接尾辞を除いて1～4文字で構成される。例えば「汉族」（Han），「蒙古族」（Mongolian），「乌兹别克族」（Uzbek）などである。スペースを節約するため，民族名

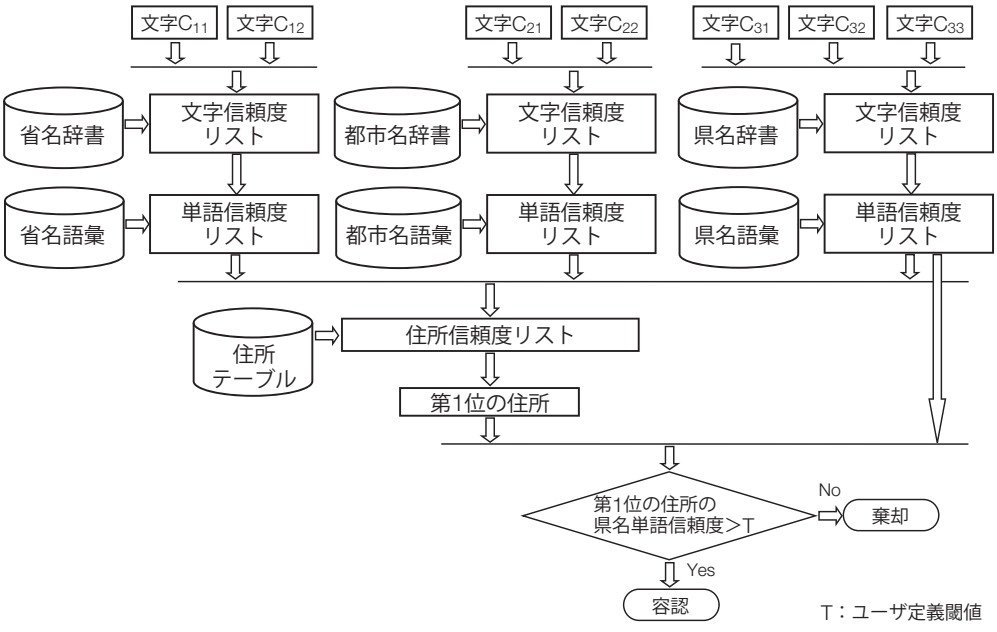
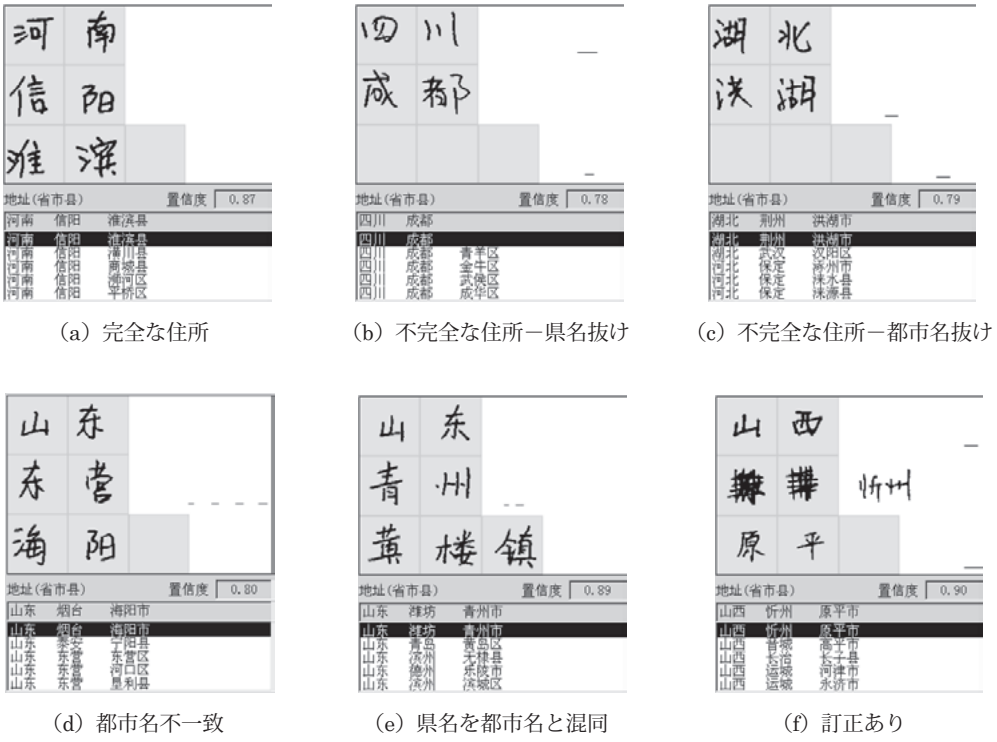


図-4　手書き住所認識の解決法
Fig.4-Solution for handwriting address recognition.



（a）完全な住所　（b）不完全な住所－県名抜け　（c）不完全な住所－都市名抜け

（d）都市名不一致　（e）県名を都市名と混同　（f）訂正あり

図-5　手書き住所と認識結果の例
Fig.5-Examples of handwriting addresses and their recognition results.

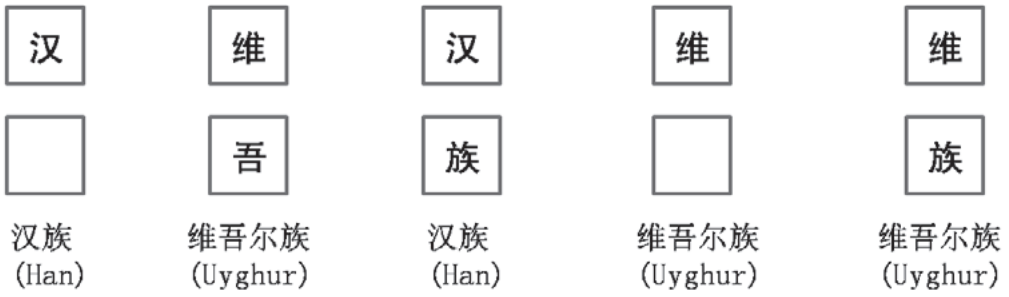
フィールドは2文字分のみとなっている。したがって，識別する文字集合を100文字未満にすることができる。文字集合が小さいことは識別精度向上に役立つが，省略と別名認識により識別難度も上がる。例えば,「蒙古族」（Mongolian）と「维吾尔族」（Uyghur）の省略形はそれぞれ「蒙」,「维」となり,「摩梭族」（Mosuo）は「纳西族」（Naxi）の別名である。接尾辞「族」（Zu）は書かないという条件があるが，調査員が誤って書いてしまうことがよくあるため，いずれも対応しなければならない。対応済みの形式を**図-6**に示す。

民族名文字列の長さを考慮し，1文字民族名文字列と2文字民族名文字列という二つのケースを別々に扱う。民族名は，1文字民族名単語の語彙と2文字民族名単語の語彙の二つに分類する。民族名フィールドに1文字のみ記入されている場合，1文字民族名語彙とマッチングする。結果は1文字の認識のみに基づく。2文字記入されている場合，2文字民族名語彙を使用して民族名のマッチングを行う。2文字目の認識候補に「族」（Zu）がある場合は，1文字民族名に「族」（Zu）を加えたものを民族名マッチングに追加する。民族認識の解決法を**図-7**に示す。実際の国勢調査の環境で民族名認識の精度は99.9％を上回った。

●　低品質手書き漢字認識

前述の手書き漢字認識（HCCR）では，既知情報を適用して本技術による解決法の実現可能性を確保したが，高精度文字認識技術という基盤は共通している。HCCRのカテゴリ数が多いため，高精度文字認識は容易ではない。さらに実際の国勢調査環境では続け字やノイズなどが起こりやすい。国勢調査員は数百万人おり，その手書き文字も千差万別で，手書き品質の統一は不可能である。これらはすべて，制御できない品質要因である。低品質の手書き漢字の認識成績を向上させることは不可欠である。実際は，国勢調査におけるHCCRの最終的な認識成績は1文字認識と文字認識棄却という二つの要素によって決まる。これらの要素を両方とも強化するため，以下の四つの主要技術を開発し，適用した。

・修正線形判別分析（MLDA）に基づく文字認識
・部分空間に基づく類似文字判別
・複数分類器の結合
・相互情報に基づく文字認識の適応的棄却



（a）標準の1文字民族名　（b）標準の複数文字民族名　（c）1文字民族名に接尾辞を加えたもの　（d）民族名の省略形　（e）民族名の省略形に接尾辞を加えたもの

図-6　民族名の記入形式
Fig.6-Nationality writing format.
(a) Standard single-character nationality; (b) Standard multi-character nationality;
(c) Single-character nationality plus suffix; (d) Nationality abbreviation; (e) Nationality abbreviation plus suffix.

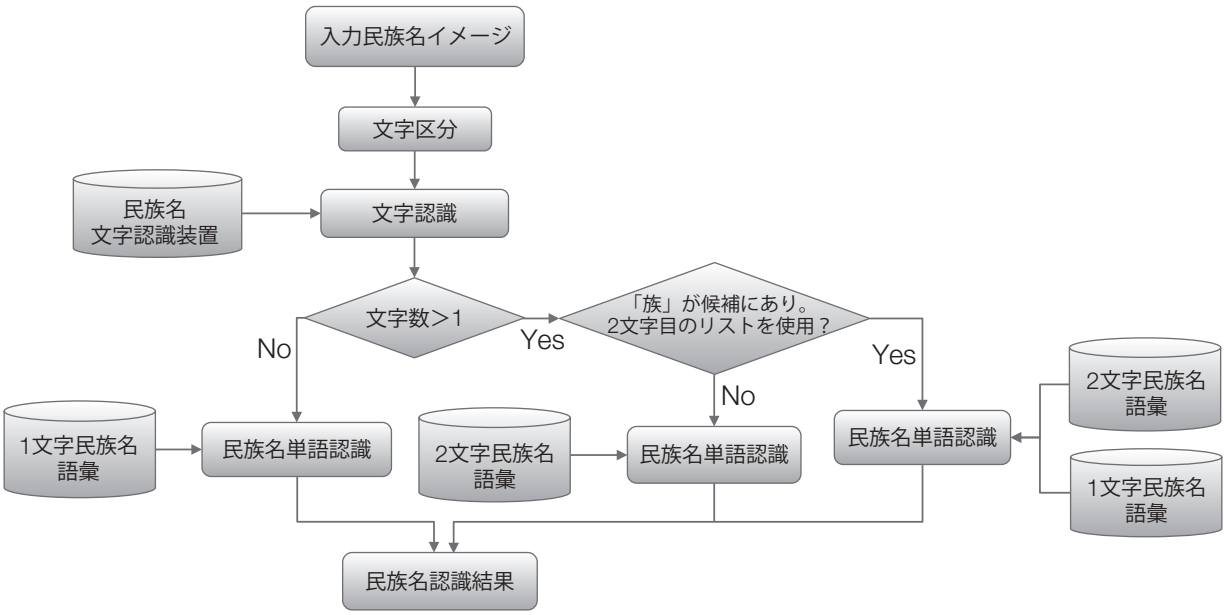


図-7　民族名認識の解決法
Fig.7-Solution for nationality recognition.

上記の各技術の採用により，FRDCのアルゴリズムに基づく認識は従来方式のみを使用する方法よりも優れていることが分かった。新しい技術をどのように利用して最終的な認識成績を上げるかを**図-8**に示す。認識段階でより多くの文字が正しく認識されるだけでなく，棄却段階では誤って認識されて棄却される文字の数も増加した。技術の詳細を以下に示す。

（1）MLDAに基づく文字認識

文字分類においては通常，LDA（線形判別分析）[5]を特徴選択に適用する。これで文字の特徴の次元数を数百から減らすことができる。重複する情報を除去すると計算効率が上がる。さらに，判別情報を特徴選択で強化し，これで分類精度を向上させる。従来のLDAではグローバルなクラス間分散の最大化を最適化目的関数として使用したが，これにより，変形後に分離が困難になるクラスができる。したがって本技術では，文字の類似を共分散の計算の要素として採用した。これで特徴変形後に類似文字が互いに遠ざけられやすくなる。すなわち，新しい特徴空間で類似文字間の高可分性が維持され，分類性が向上する。

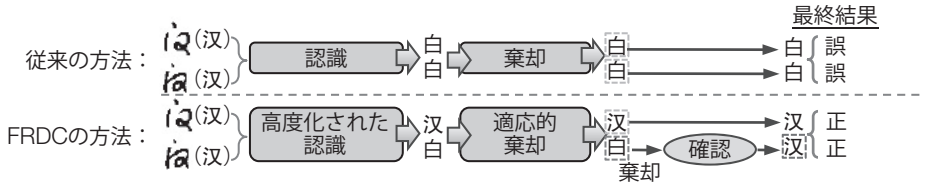


図-8　手書き漢字認識能力の向上
Fig.8-Improved handwritten Chinese character recognition.

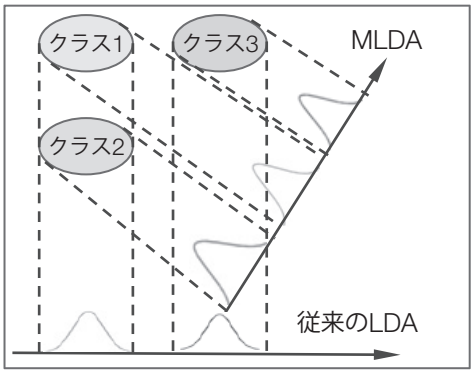


図-9　修正線形判別分析
Fig.9-Modified linear discriminant analysis.

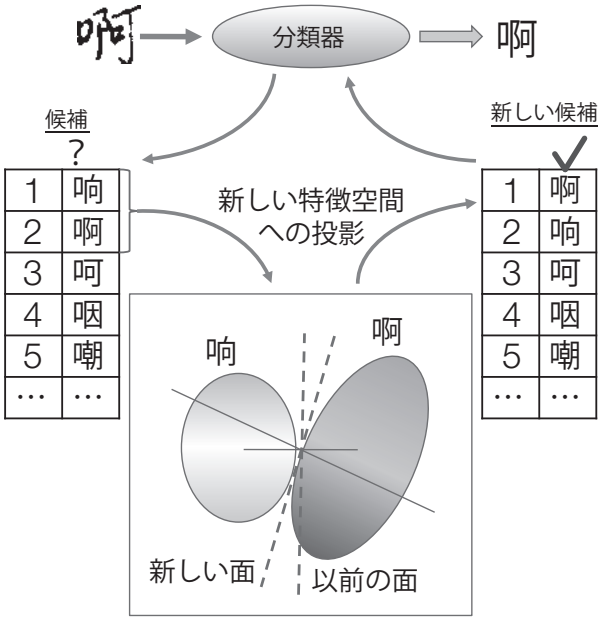


図-10　部分空間に基づく類似文字判別
Fig.10-Subspace-based similar-character discrimination.

図-9に示すように，従来のLDA変形後は二つのクラスが混ざっているが，MLDA投影軸では三つのクラスがすべて分離可能である。MLDAの分類精度のほうが確実に優れている。

(2) 部分空間に基づく類似文字判別

文字認識誤りを調べると，分類の誤りのほとんどは類似文字間で発生している。一般的には，最終的な識別精度は類似文字判別力によって決まる。通常の文字認識装置では，文字集合全体に対して全般的な高い認識成績を期待する。周知のとおり漢字認識は大規模カテゴリの問題であるため，集合全体をグローバルに考えると類似パターンのすべてを処理することができなくなる。認識から潜在的類似文字パターンを特定し，これらの類似文字パターンのための特殊分類器を設計できれば，文字集合全体に対する大規模カテゴリ認識を類似文字間の小規模カテゴリ認識に変換できる。その結果，類似文字の認識能力の向上を図ることができ，最終的な文字認識成績が上がる。

正しいクラスラベルのほとんどは認識結果の最初の2候補にある。これに基づき，最初の二つの認識候補にある疑わしい類似文字ペアを処理するペアワイズ類似文字分類器を設計した。とくに，類似文字ペアが最初の2候補にあることが分かった場合，この類似文字ペアのペアワイズ分類器を適用する。これで，大規模カテゴリ認識が2クラス問題に変換される。これらの二つの認識装置をそれぞれグローバル認識装置，ローカル認識装置とした。ローカル認識装置はローカル判別情報を採用するため，対応する類似文字ペアにおいて常にグローバル認識装置より優れた成績を示す。最終的に，ローカル認識装置の結果を組み合わせ，グローバル認識装置の分類誤りは**図-10**に示すように修正される。

(3) 複数分類器の結合

文字分類精度と信頼性を更に向上させるため，複数特徴分類器の結合を使用して，より強力な分類器を作成する。一般的に，これらの特徴間の相

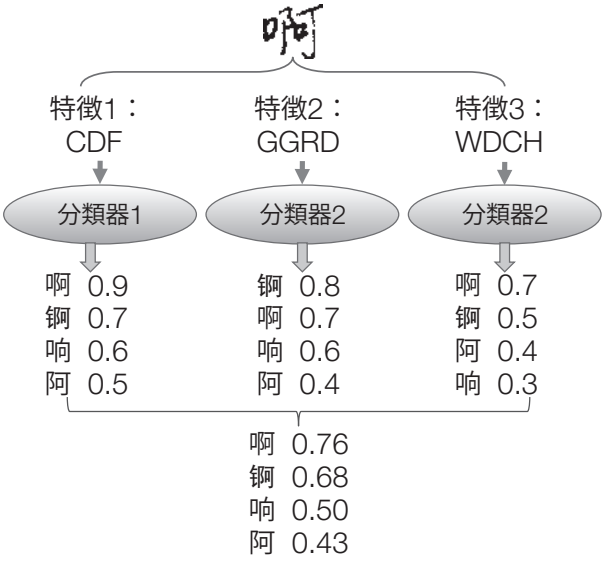


図-11　複数分類器の結合
Fig.11-Multi-classifier combination.

補性が高いほど結合後の認識成績が高くなる。理論多様性の原則により，輪郭方向特徴（CDF），[6] 疑似グレー階調方向（GGRD），[6] 加重方向コードヒストグラム（WDCH）[7] の特徴を選択した。加重投票ルールを結合に適用すると，分類精度だけでなく汎化能力も大幅に向上する。これは実際の適用において大変重要である。**図-11**は複数分類器の結合の例である。

（4）相互情報に基づく適応的棄却

国勢調査のデータ精度の要求は非常に厳しい。自動文字認識結果の精度は，そのままでは要求を満たすことができない。人の目で確認すれば誤り率は非常に0に近くなるが，そのコストは容認できないレベルである。疑わしい認識結果のみを特定して確認すれば，データ精度とコストの間の適切なバランスが得られる。棄却の目的は，疑わしい認識結果を判別することである。棄却量は人の目による確認の仕事量に対応する。したがって，棄却の目標は棄却をなるべく少なくしつつ誤り率を最低限に抑えることである。

今回，相互情報に基づく適応的棄却法を提案した。[8] すべてのサンプルに同じ棄却閾値を使用する従来の棄却法と異なり，この手法ではサンプルのローカル情報を利用して棄却閾値を最適化する。最初の二つの認識候補に同じクラスのペアがあるサンプルは同じサンプルカテゴリとして扱った。

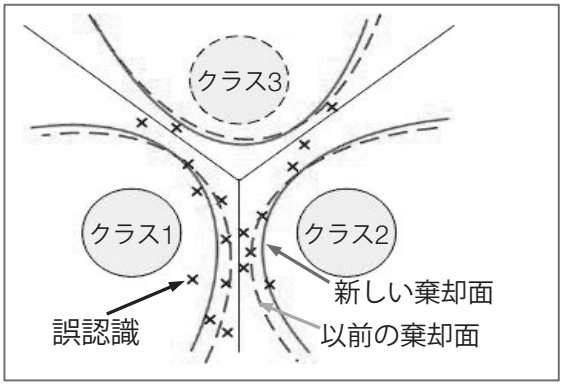


図-12　適応的棄却
Fig.12-Adaptive rejection.

表-1　手書き文字認識評価結果

手書き文字認識タスク	認識率（%）	棄却率（%）	データセットサイズ（サンプル数）
数字認識	99.98%	0.62%	1.0 M
住所認識	99.6%	4.1%	0.5 M
民族名認識	99.92%	2.7%	1.5 M

相互情報に基づく相関関係を使用し，利用できるサンプルカテゴリを選択する。つぎに，カテゴリに依存する棄却パラメータを，最大エントロピーに基づき最適化する。これで棄却閾値は特定のすべてのサンプルに適応し，最適な棄却結果が得られる。**図-12**は，適応的棄却法と従来の棄却法の差を示すものである。新しい方法によれば，従来の棄却法と同じ棄却率で，より多くの誤認識サンプルが棄却される。

評　　価

前述のアルゴリズムと技術を検証するため，大規模データセットで評価した。評価結果の詳細を**表-1**に示す。手書き数字認識はHCCRと比較して小規模カテゴリ認識であるため，良い認識成績が得られる。非常に低い誤り率と棄却率を同時に実現している。住所と民族名の認識結果から，タスクの既知情報が大きな利点となっていることが分かる。周知のとおり，HCCRでは1文字認識は困難であるが，既知情報を利用すると非常に高精度の住所と民族名の認識が実現可能になる。全体として，すべての認識タスクが顧客の要求を満足した。

むすび

本稿では，第6回国勢調査に適用されたFRDCの高度な手書き認識技術を紹介した。低品質の手書き漢字を国勢調査の実際の環境で処理できるようにするため，FRDCは手書き漢字認識能力を向上させる一連の技術として，MLDA，部分空間に基づく類似文字判別，複数分類器の結合，相互情報に基づく適応的棄却を開発した。FRDCのアルゴリズムは，住所と民族名の既知情報を組み合わせることにより低棄却率で99％を上回る精度を実現し，国勢調査の厳しい要求に応えた。大規模な中国の国勢調査プロジェクトに漢字認識技術が使用されたのは初めてである。これらの成果はすべて国勢調査の効率向上とコスト削減に大きく貢献するものである。

最後に，このプロジェクトにおける中国統計局（NBSC），Nikoyo（NKY）のご支援とご協力に感謝する。

参考文献

(1) M. Cheriet et al.：Character Recognition Systems：A Guide for Students and Practitioners. Wiley-Interscience，2007.
(2) V. Vapnik：The Nature of Statistical Learning Theory. Springer，New York，1995.
(3) C. Burges：A tutorial on support vector machines for pattern recognition. In “Data Mining and Knowledge Discovery.” Kluwer Academic Publishers，Boston，1998.
(4) C. Liu：Classifier combination based on confidence transformation. *Pattern Recognition*，Vol.38，No.11，p.11-28（2005）.
(5) R. O. Duda et al.：Pattern Classification. Wiley，2000.
(6) C. Liu et al.：Handwritten digit recognition：investigation of normalization and feature extraction techniques. *Pattern Recognition*，Vol.37，No.2，p.265-279（2004）.
(7) F. Kimura et al.：Improvement of handwritten Japanese character recognition using weighted direction code histogram. *Pattern Recognition*，Vol.30，No.9，p.1329-1337（1997）.
(8) Y. Zhu et al.：Rejection Optimization Based on Threshold Mapping for Offline Handwritten Chinese Character Recognition. The 12th International Conference on Frontiers in Handwriting Recognition，2010，p.72-77.

著者紹介

鄭　大念（Zheng　Danian）
富士通研究開発中心有限公司（FRDC）
情報技術研究部（ITL）所属
現在，文書画像処理と光学式文字認識の研究開発に従事。

于　浩（Yu　Hao）
富士通研究開発中心有限公司（FRDC）
情報技術研究部（ITL）所属
現在，情報処理やスマートグリッドなどの研究開発に従事。

孫　俊（Sun　Jun）
富士通研究開発中心有限公司（FRDC）
情報技術研究部（ITL）所属
現在，文書画像処理と光学式文字認識の研究開発に従事。

直井　聡（なおい　さとし）
富士通研究開発中心有限公司（FRDC）
総経理