様式 C-19

科学研究費補助金研究成果報告書

平成２３年６月２４日現在

機関番号：55201
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：20560389
研究課題名（和文）　階層型識別器を用いた高精度古文書文字認識とその応用

研究課題名（英文）　Japanese Historical Character Recognition using Hierarchical Classifiers and Its Applications

研究代表者
堀内　匡（HORIUCHI TADASHI）
松江工業高等専門学校・電子制御工学科・准教授
研究者番号：50294129

研究成果の概要（和文）：
本研究では，認識対象の古文書文字の字種数を限定したうえで，認識部を大分類部と細分類部に分けた階層的な識別器を用いた高精度の古文書文字認識を実現した．さらに，古文書文字認識の応用として，高精度の認識手法を用いて，初心者が読解困難な文字に対する読みの候補文字を複数個提示することにより古文書読解を支援するシステムを構築した．

研究成果の概要（英文）：
In this research, we developed the Japanese historical character recognition system using the hierarchical classifiers. The hierarchical classifiers consist of a rough-classifier and a set of fine-classifiers. Moreover, we developed the reading support system for Japanese historical documents. In this reading support system, the user selects the unknown character by mouse operation and the character recognition system outputs five candidate characters.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,700,000	510,000	2,210,000
２００９年度	1,000,000	300,000	1,300,000
２０１０年度	800,000	240,000	1,040,000
総　計	3,500,000	1,050,000	4,550,000

研究分野：工学
科研費の分科・細目：電気電子工学・システム工学
キーワード：パターン認識，機械学習，古文書文字認識，古文書読解支援

１．研究開始当初の背景

古文書の翻刻（古文書を読んで活字に直すこと）は，歴史研究において不可欠な基礎的作業であるが，国内には翻刻されていない古文書が数多く存在する．古文書の翻刻作業は専門家に頼らざるをえないにも関わらず，翻刻すべき古文書数に比べて専門家は非常に少ないのが現状である．そこで，知能情報技術を用いて古文書の翻刻作業を支援するようなシステムを開発できれば，歴史研究において有用な道具になると考えられ，近年研究が進められている．代表的な研究として，古文書翻刻支援システム開発プロジェクトがあり，プロジェクトの一環として，和泉らは漢字16字種からなる古文書文字データに対して約96%の認識率を得ているが，対象字種数が極めて少ない．一方，申請者らは，特徴量として方向線素特徴量，認識手法としてマハラノビス距離に基づく最近傍識別法を採用した古文書文字認識システムの構築を進めている．

２．研究の目的

本研究では，上記の研究成果を踏まえて，限定した字種数の古文書文字を対象として，階層型識別器を用いた高精度の認識手法およびその応用に関して検討する．具体的には以下の二つを研究の目的とする．

第一の目的として，認識部を大分類部と細分類部に分けた階層的な識別器を用いた高精度の古文書文字認識を実現する．さらに，第二の目的として，上記で得られた認識手法を用いて，初心者が読解困難な文字に対する読みの候補文字を複数個提示することにより古文書の読解を支援するシステムを構築する．

３．研究の方法

(1) 平仮名文字データベースの作成

古文書翻刻支援システム開発プロジェクトが公開している古文書データベース HCD シリーズから多数の字種の古文書文字データを集めて本研究で用いるデータセットを準備した．

(2) 認識手法の実装と性能評価

大分類部に自己組織化マップを採用し，細分類部に多層パーセプトロン，サポートベクターマシン，マハラノビス距離に基づく最近傍識別法を導入した各認識手法を実装し，上記のデータセットに対して認識実験を行い，認識率を比較した．

(3) 古文書読解支援システムの構築と評価

上記で実装した認識手法を組み込んだ古文書読解支援システムを構築し，ユーザによる実証評価を実施し，今後の課題を検討した．

４．研究成果

(1) 平仮名文字データベースの作成

古文書翻刻支援システム開発プロジェクトが公開している古文書文字データベース HCD1 および HCD1a～1e からサンプル数が 200 以上ある字種の古文書文字データを集め，本研究で使用するデータセットを作成した．

(2) 認識手法の実装と性能評価

まず，大分類部に自己組織化マップ，細分類部に多層パーセプトロンのモジュール群を用意したモジュール型ニューラルネットワークによる階層型識別器を実現し，交差検証法による文字認識実験を行い，認識性能を評価した．その結果，61 字種の古文書文字に対して約 94%という高い認識精度を得られることを明らかにした．

また，細分類部の認識手法として，サポートベクターマシンを用いた手法，マハラノビス距離に基づく最近傍識別法を用いた識別器を実現し，交差検証法による文字認識実験を行い，認識性能を評価した．その結果，61 字種の古文書文字に対して，両手法ともに約 96%という高い認識精度が得られることを明らかにした．

また，細分類部の認識手法として，マルチテンプレートマッチングを用いた識別器も実現し，交差検証法による認識実験を行い，61 字種の古文書文字に対して，約 93%の認識精度が得られた．1 字種あたり 25 個のテンプレートを自己組織化マップにより学習することで，ある程度の認識精度を得ることができた．

表 1 細分類部の認識精度の比較

手法	認識精度
多層パーセプトロン	94.2%
サポートベクターマシン	95.6%
マハラノビス距離に基づく最近傍識別法	96.1%
マルチテンプレートマッチング	93.3%

(3) 古文書読解支援システムの構築と評価

上記の認識実験で最も高い性能を示したマハラノビス距離に基づく最近傍識別法を組み込んだ古文書読解支援システムを構築した．その読解支援システムの画面例を図 1 に示す．このシステムでは，初心者が読解困難な文字に対する読みの候補文字を 5 個提示している．

ユーザによる評価実験を実施したところ，ユーザが読むことが困難な文字として選択した文字に対するシステムの正答率は約 83%であった．これは，古文書文字の個別認識実験における正答率に比べると低下しているが，ユーザがマウスを用いて文字の範囲を選択しており，前後の文字がつながっている文字が多いことなどを考慮すると，一定の有効性が確認できた．

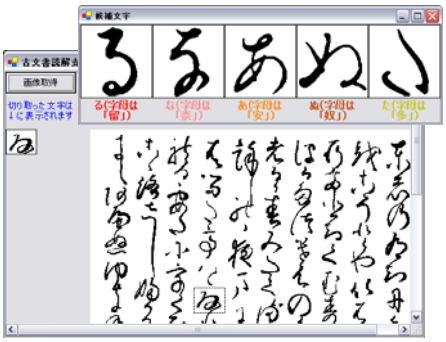

図 1 古文書読解支援システムの画面例

(4) まとめ

古文書文字認識において，大分類部と細分類部に分けた階層的識別器の有効性を示すことができた．また，古文書読解支援システムの構築により，くずし字を読むことを支援する一つの手段を実現することができた．

今後の課題として，平仮名と漢字を分けてそれぞれの認識器を用いること，より多くの字種数を対象とすること，前後の文字がつながっている場合の文字の切り出し方法を検討することなどが考えられる．

５．主な発表論文等

〔雑誌論文〕（計１件）

① Tadashi Horiuchi and Satoru Kato, A Study on Japanese Historical Character Recognition using Modular Neural Networks, International Journal of Innovative Computing, Information and Control, 査読有, Vol.7, No.8, 2011

〔学会発表〕（計 12 件）

① 伊藤崇文，堀内 匡，サポートベクターマシンにおけるパラメータ探索手法の検討，第 19 回計測自動制御学会中国支部学術講演会，pp.52-53，2010，島根県松江市
② 加藤 聡，堀内 匡，自己組織化マップと情報量規準によるクラスタ数の推定法に関する基礎的研究，第 9 回情報科学技術フォーラム，Vol.2，pp.537-538，2010，福岡県福岡市
③ 加藤 聡，堀内 匡，福島有希，改良型マハラノビス距離を用いた古文書文字認識，平成 22 年電気学会全国大会，Vol.3，pp.126-127，2010，東京都千代田区
④ 加藤 聡，堀内 匡，自己組織化マップによるマルチテンプレート学習の古文書文字認識への適用，第 11 回自己組織化マップ研究会 2010，pp.13-18，2010，福岡県北九州市
⑤ 加藤 聡，堀内 匡，古文書文字認識システムにおけるサポートベクターマシンの適用に関する研究，第 8 回情報科学技術フォーラム，Vol.3，pp.135-136，2009，宮城県仙台市
⑥ 加藤 聡，堀内 匡，サポートベクターマシンを用いた古文書文字認識に関する基礎的検討，第 53 回システム制御情報学会研究発表講演会，2009，兵庫県神戸市
⑦ Tadashi Horiuchi and Satoru Kato, A Study on Japanese Historical Character Recognition using Modular Neural Networks, The Fourth International Conference on Innovative Computing, Information and Control, B09-04, 2009, Kaohsiung, Taiwan
⑧ 加藤 聡，堀内 匡，サポートベクターマシンを用いた古文書文字認識の試み，平成 21 年電気学会全国大会，Vol.3，pp.57-58，2009，北海道札幌市
⑨ 加藤 聡，堀内 匡，古文書文字認識における認識手法に関する検討，第 13 回日本知能情報ファジィ学会中国・四国支部大会，pp.15-18，2008，広島県広島市
⑩ 堀内 匡，小須賀祐介，高橋朋之，章 忠，今村 孝，セルラニューラルネットワークを用いた古文書文字認識の試み，平成 20 年度電気・情報関連学会中国支部連合大会，pp.330-331，2008，鳥取県鳥取市
⑪ 加藤 聡，堀内 匡，古文書文字認識システムにおける認識手法の比較検討，平成 20 年度電気・情報関連学会中国支部連合大会，pp.460-461，2008，鳥取県鳥取市
⑫ 加藤 聡，堀内 匡，高橋朋之，モジュール型ニューラルネットワークを用いた古文書文字認識－正準判別分析による次元削減の導入－，平成 20 年電気学会電子・情報・システム部門大会，pp.730-733，2008，北海道函館市

６．研究組織

(1)研究代表者
堀内　匡（HORIUCHI TADASHI）
松江工業高等専門学校・電子制御工学科・准教授
研究者番号：50294129

(2)研究分担者
加藤　聡（KATO SATORU）
松江工業高等専門学校・情報工学科・講師
研究者番号：40342547

(3)連携研究者
山崎　真克（YAMAZAKI MASAKATSU）
比治山大学・現代文化学部・准教授
研究者番号：10342544