様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２８　日現在

機関番号：　５５２０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３０４１６

研究課題名（和文）高精度な古文書文字認識器を用いた古文書読解支援システムの構築に関する研究

研究課題名（英文）A Study on Developing a Reading Support System for Japanese Historical Documents by using Accurate Historical Character Recognizer

研究代表者

加藤　聡（Satoru, Kato）

松江工業高等専門学校・情報工学科・准教授

研究者番号：４０３４２５４７

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：古文書読解支援システムの認識部に用いる認識手法について検討し，ノイズ等への頑健性についてはSOMテンプレートと呼ばれる手法，認識精度と計算量のバランスについてはマルチテンプレート法がそれぞれ優れていることを明らかにした．さらに，それらの手法を用いたシステムのプロトタイプをタブレット端末上に実装し，ユーザインタフェース等についての検討を行った．また，古文書読解支援システムの認識部における文字パターン学習の高速化について，PCクラスタと呼ばれる並列計算機環境の構築やGPGPUと呼ばれるベクトル計算の高速化手法の導入などを行い，学習にかかる時間を大幅に削減することができた．

研究成果の概要（英文）：In this study, several methods of Japanese historical character recognition are examined. It is revealed that the SOM based method called "SOM template" has excellent noise resistance, and multi-template method is good for the balance of trade-off between accuracy and computation time. A prototype of the reading support system for Japanese historical documents is constructed on a tablet terminal. Furthermore, PC-cluster and GPGPU computing method is used for speeding up the SOM learning process and we confirmed that this approach is effective in a large scale of the SOM learning process.

研究分野：　パターン認識

キーワード：　古文書文字認識　自己組織化マップ　マルチテンプレート　古文書読解支援システム　並列処理　PCクラスタ　GPGPU

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・読解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[1]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[2]，対象字種数が極めて少ない．

一方，本研究開始当初は，特徴量として方向線素特徴量[3]を採用した識別器を構築し，平仮名 45 字種のデータに対して約 77%（第 5 位認識率で約 95%）の認識率を得ている[4]．さらにこの識別器を，平仮名を対象とした「古文書読解支援システム」（図 1 参照）に応用している．

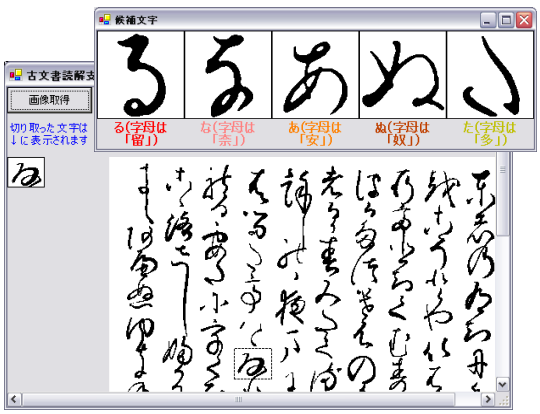


図１　古文書読解支援システム（申請時）

２．研究の目的

本研究では，上記の研究成果を踏まえて，より高精度で応答性の良い文字認識器を構築して古文書読解支援システムに搭載することが目的である．そのためには，文字画像から認識に有効な特徴量を取り出すための特徴抽出法や，実際に文字の認識を行う文字認識手法をどのようにするかが重要となる．

特徴抽出法としては，現代文字の認識手法の中でも精度の良い認識が可能であるとされる方向線素特徴量を採用する．また，文字認識手法としては，機械学習アルゴリズムである多層パーセプトロン（MLP）やサポートベクターマシン（SVM）を用い，古文書文字認識の高精度化を目指す．これらの認識手法は，認識対象の字種の増加に伴って，学習に多大な計算時間を要する．そこで，自己組織化マップ（SOM）を用いて，認識対象文字の特徴空間を適切にクラスタリングすることにより，認識器における学習・認識処理双方の高速化について検討する．

さらに，改良された古文書文字認識器を，申請者が開発中の古文書読解支援システムに組み込み，総合的な評価を行う．具体的な内容は以下の(1)および(2)に示す通りである．

(1) 自己組織化マップ（SOM）による，文字認識器の高精度化と学習高速化

MLP や SVM は高精度な文字認識を実現できる手法として注目されているが，認識対象の学習に時間がかかることが難点である．SOM を用いて学習文字サンプルの特徴空間を事前にクラスタリングすることにより，MLP や SVM の学習に必要な文字サンプル数の削減や，認識時における文字候補の絞込みが可能となり，MLP や SVM における学習・認識処理を高速化できる．ここでは，認識精度を保ちつつどの程度特徴ベクトルの次元数を削減できるか，すなわち認識精度と学習・認識処理の高速化とのトレードオフについて明らかにする．

(2) 古文書読解支援システムへの組込みと評価

高速・高精度化した古文書文字認識器を，開発中の古文書読解支援システム（図 1）に組込む．これは，くずし字に関しての初心者が読解困難な文字に対して，読みの候補文字を複数個（5 個程度）提示することにより，古文書の読解を支援するシステムである．このシステムに対し，読みの候補文字が表示されるまでの時間や，表示される候補文字の妥当性などを評価し，その有効性を明らかにする．

３．研究の方法

本研究の目的を達成するため，以下の計画に基づいて研究を進めることとした．

(1) 古文書文字特徴量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[1]で公開している古文書データベースや他の出典などから，より多くの古文書文字データを集め，これらを元に文字特徴量（方向線素特徴量）のデータベースを作成する．

(2) 認識手法の実装と性能評価

多層パーセプトロン（MLP），改良型マハラノビス距離，サポートベクターマシン（SVM）を用いた認識手法をそれぞれ実装し，認識精度を比較する．また，文字画像の部分的な欠損などによるノイズに対して頑健な手法と

して，自己組織化マップ（SOM）を用いた認識手法を検討し，同じく SOM による文字特徴空間のクラスタリングを併用した場合の，学習時間や認識時間と認識精度のトレードオフについて，評価実験を通じて確認する．

(3) 古文書読解支援システムへの認識器の組込み

これまでに実装した認識手法を古文書読解支援システムのプロトタイプシステムに組込む．現状のプロトタイプではデスクトップ PC での使用を前提とした実装となっているため，ユーザの利便性を考慮して，タブレット端末向けにシステムを構築しなおす．

(4) PC クラスタ等を用いた学習処理の高速化

一般的に，機械学習をベースとした認識器は，ある程度以上の認識精度を得ようとすると大量のサンプルを学習させる必要があり，学習処理に多大な時間を要する．そこで，並列分散計算機環境である PC クラスタや，ベクトル演算を高速に行うことができる GPGPU の手法などを応用し，本研究で必要となる文字認識のための学習処理時間を削減する．

４．研究成果

(1) 古文書文字データベースの作成

古文書文字データベースの作成に関しては，文献等からくずし字のサンプルを集め，平仮名のくずし字画像データベースを作成した．現代の平仮名は「ゐ」と「ゑ」を含めると 48 字種が用いられているが，古文書に用いられている変体仮名は必ずしも一音につき一文字とは限らず，例えば同じ「か」の音を表すにも「加」を字母とするものや「可」を字母とするものなどがある．本研究では，字母の異なるくずし字を 1 字種とし，合計で 51 字種分の文字画像データベースを構築した．各字種の字母および収集したサンプル数を表 1 にまとめる．

(2)① 認識手法の比較検討

各種の認識手法における認識精度等の比較については，マルチテンプレート法によるもの，改良型マハラノビス距離[5]を用いたもの，サポートベクターマシン（SVM），の 3 手法を対象にして行った．実験開始時には前節で述べた古文書文字データベースが作成途上であったため，[1]による既存の古文書文字データベースから 61 字種分のデータを用いて認識精度を比較した．結果を表 2 に示す．純粋に認識精度だけで判断すれば，改良型マハラノビス距離が最も適しているように思われるが，古文書文字認識の場合，辞書として用いることができる文字サンプル数が字種によっては極端に少なくなるため，マハラノビス距離のような統計的パラメータに基づく認識手法は必ずしも適さない場合がある．また，SVM による認識についても，計算量の観点から CPU パワーの面で劣るタブレット端末には不向きであることが予想される．一方，マルチテンプレート法はシンプルで計算量も少ない手法でありながら，他の手法と比較して認識精度はそれほど劣っていない．以上のことから，認識精度と計算量のバランスを考慮した場合，マルチテンプレート法は有効な認識手法であると思われる．

表 1　平仮名くずし字データベース
（各字種の字母およびサンプル数）

No.	字母	サンプル数	No.	字母	サンプル数
1	安	154	27	祢	100
2	以	156	28	乃	142
3	宇	145	29	波	94
4	衣	123	30	者	182
5	於	160	31	比	148
6	加	101	32	不	138
7	可	213	33	部	128
8	幾	187	34	保	107
9	久	191	35	末	142
10	計	102	36	美	129
11	己	148	37	武	86
12	左	155	38	女	101
13	之	188	39	毛	141
14	寸	82	40	也	125
15	世	110	41	由	99
16	曽	138	42	与	134
17	太	48	43	良	207
18	多	207	44	利	192
19	知	137	45	留	193
20	川	153	46	礼	154
21	天	147	47	呂	128
22	止	178	48	和	104
23	奈	184	49	為	56
24	仁	53	50	恵	70
25	尓	129	51	遠	104
26	奴	81			

表 2　各種の文字認識手法における
認識精度の比較

認識手法	認識精度 （パラメータ）
Multiple Template	93.3 % (25 templates/class)
Modified Mahalanobis	96.1 % (bias = 0.02)
SVM (RBF Kernel)	95.6 % (C=4, γ=0.125)

(2)② ノイズに頑健な認識手法の検討

一般的に，機械学習アルゴリズムに基づく文字認識手法は，欠損のない文字画像に対しては比較的良好な認識精度を得ることができる．しかしながら，古文書にはノイズや欠損のある文字画像があり，そのことが認識率の低下を招くおそれがある．そこで，装飾文字の認識に用いられる，SOM テンプレートを用いた文字認識手法を前節で構築した古文書文字データベースを用いた文字認識に適用し，従来手法との比較を行った．

テスト用画像に人工的に欠損を加えた場合と加えない場合とで認識精度を比較したところ，SOM テンプレートを用いた認識手法は欠損に対する頑健性が確認できたものの，総合的な認識精度は従来手法に及ばないことが分かった（図 2 参照）．

変体仮名には字母が同じであっても，字体のくずし方にいくつかのバリエーションが存在する場合がある．バリエーションの違いを無視して同一字種として取り扱ってしまったために認識精度の低下を招いてしまった可能性が考えられたため，SOM テンプレート作成時に使用される文字サンプル群をあらかじめクラスタリングしておき，一つの字種に対して必要に応じて複数のテンプレートを使用できるようにすることで，SOM テンプレートを用いた古文書文字認識手法の総合的な精度向上を試みた．

具体的には，平仮名のくずし字を対象として，各字種のくずし字画像サンプル群（一字種あたり 50～200 サンプル）に対し，ウォード法による階層的クラスタリングを施し，その結果によって字種ごとにサンプル群を複数（2～4 個）のクラスタに分割し，それぞれのクラスタに対して SOM テンプレートを作成した．

この手法を，従来法において極めて認識精度の低かった字種に対して適用したところ，平均して 20%の認識精度の改善が見られた（図 3 参照）．しかしながら，全体的な認識精度を見ると，(2)①で述べた手法と比較して劣っている．したがって，提案手法はノイズが大きく(2)①の各手法がうまく適用できない場合に，補助的に併用することが望ましいと考えられる．

(3) 古文書読解支援システムの試作

古文書読解支援システムを手軽に使用できるようにするためには，従来のデスクトップ型 PC ではなく，タブレット端末等を用いることが望ましい．本研究において実装の対象としたのは 9 インチ程度の Android タブレットである．タブレット端末は，拡大・縮小，スクロール，領域選択など，画面に触れて直観的に操作できるため，マウス操作よりも利便性が高く，デスクトップ PC やノート PC と比較して可搬性も非常に優れている．

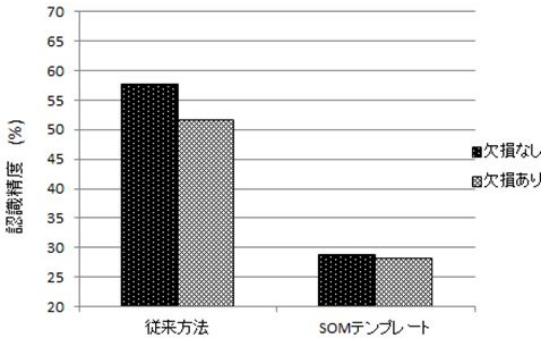


図 2　SOM テンプレートと従来手法との認識精度の比較

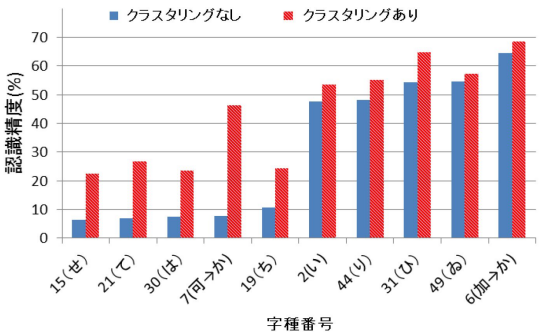


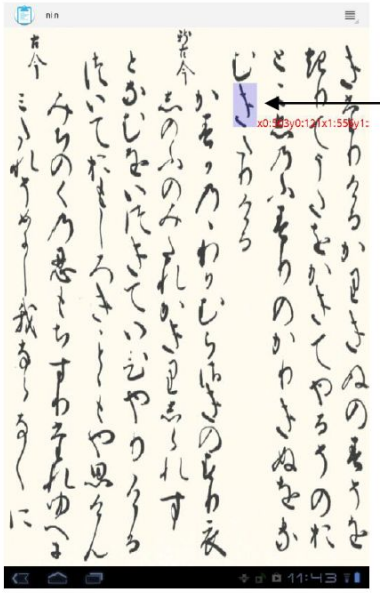
図 3　文字サンプルのクラスタリングによる認識精度の改善

実装したタブレット版古文書読解支援システムの画面を図 4 に示す．メイン画面にはあらかじめスキャナ等で取り込まれた古文書画像が表示されており，ユーザは読みが分からない文字がある場合には，タッチパネルでの操作によって不明文字の領域選択ができる（図 4(a)参照）．システム側は選択された領域に含まれる文字画像を文字認識部に送り，文字認識部は認識結果に基づいて，いくつかの認識候補をユーザに提示する（図 4(b)参照）．

なお，本研究において実装されたシステムは，基本的なユーザインタフェースのデザインは完成しているものの，今のところ試作の域を脱していない．したがって，ユーザにとっての使いやすさ等に関する評価実験を行いつつ，実用化に向けてシステムの改良を続けていく必要がある．

(4) PC クラスタによる SOM 学習の高速化

SOM の学習に必要な計算時間は学習サンプル数に比例して増大するため，PC クラスタのような並列・分散計算機環境を用いることで学習時間の増大を抑えることができる．さらに，SOM の学習アルゴリズムにはベクトル演算が頻出するため，3 次元グラフィクス生成用のプロセッサである GPU を汎用の数値計算に用いる，GPGPU と呼ばれるアプローチも有効であると考えられる．そこで，PC クラスタ



(a)　古文書画像の表示と不明文字の領域選択

(b)　選択された不明文字に対する候補文字の提示

図 4　Android タブレット版
古文書読解支援システムの画面例

の各計算ノードに GPU を搭載して，計算ノード単位で GPGPU による数値計算の高速化を可能とした GPU クラスタを構築し，その上で SOM の学習アルゴリズムを実装した．

GPGPU の環境として，本研究では，NVIDIA 社が自社の GPU 向けに提供している CUDA8.0 を使用することとした．PC クラスタを構築するためのメッセージ通信ライブラリとしては，MPI（Message Passing Interface）を用いることが一般的である．本研究では，MPI の実装の一つである OpenMPI を使用した．計算ノードとなる PC のスペック等は表 3 に示す通りである．本研究では，試作として 4 台の計算ノードを準備した．

GPU クラスタ上に実装した SOM に対して，入力データの個数を変化させて SOM の学習にかかる時間を計測した．図 5 は，1 台の計算ノードによる SOM 学習の実行時間（競合層サイズ 64×64，入力データ数 4200，反復回数 200）を比較したものである．GPU を使用しない場合（"CPUonly"）と比較して，GPU を使用した場合（"GTX980"および"GTX570"）では，学習にかかる時間が著しく短縮されていることが分かる．また，図 6 は，計算ノードを 1 台から 2 台に増加させたときの，入力データ数と学習時間の関係をまとめたものである（競合層サイズ 64×64，反復回数 200）．計算ノードに搭載する GPU は，いずれも GTX980 である．GPU による高速化の度合いが大きいため，計算ノードを増やした場合には，各ノードの計算量がある程度多くなければ，ノード間通信のオーバヘッドにより並列化の効果が得られないことが分かる．

以上のことから，より大量のサンプルを学習させる場合に，本研究における並列化のアプローチが有効であることが示唆された．

表 3　PC クラスタにおける
計算ノードの諸元

CPU	Intel core2Duo E7500 (2.93GHz,2core)
GPU	NVIDIA GeForce GTX570，GTX980
TPC/IP	1000Base-T Ethernet
OS	Linux （xubuntu 16.04LTS）

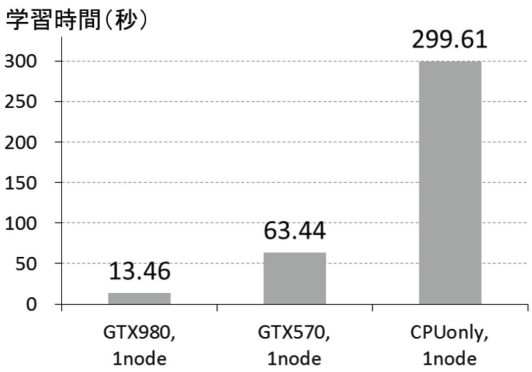


図 5　CPU と GPU における SOM 学習の
実行時間の比較

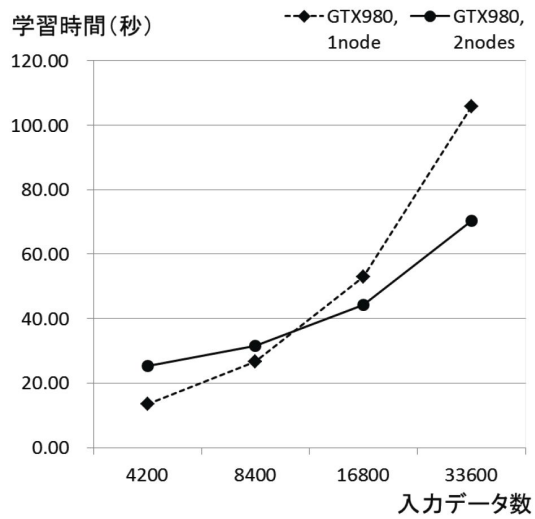


図 6　計算ノード数による SOM 学習の実行時間の比較

＜引用文献＞
[1]　“古文書翻刻支援システム開発プロジェクト”, http://ys.nichibun.ac.jp/~shoji/hcr/index.html, (2018 年 6 月 27 日アクセス確認)
[2]　“ニューラルネットワークを用いた古文書個別文字認識に関する一検討”, 和泉勇治, 他 5 名, 情報処理学会研究報告, Vol.2000, No.8, pp.9-15 (2000)
[3]　“改良型方向線素特徴量および部分空間法を用いた高精度な手書き文字認識システム”, 孫　寧, 他 2 名, 電子情報通信学会論文誌, Vol.J78-D-II, No.6, pp.922-930 (1995)
[4]　“方向線素特徴量を用いた古文書文字認識とその応用に関する検討”, 加藤聡, 他 3 名, 平成 19 年電気学会電子・情報・システム部門大会講演論文集, pp.1136-1141 (2007)
[5]　孫寧, 安倍正人, 根元義章,「改良型マハラノビス距離を用いた高精度な手書き文字認識」, 情報処理学会研究報告「グラフィクスと CAD」, No.1994-CG-072, pp.169-176 (1994)

５．主な発表論文等

〔学会発表〕(計 4 件)

①加藤聡, GPU クラスタにおける SOM の実装に関する基礎的検討, 第 33 回ファジィシステムシンポジウム, 2017 年

②加藤聡, 堀内匡, 古文書読解支援システムのタブレット端末への実装に関する研究, 電子情報通信学会総合大会, 2017 年

③加藤聡, 森田浩旭, 辞書サンプルのクラスタリングによる SOM テンプレートを用いたくずし字認識の精度向上に関する研究, 第 31 回ファジィシステムシンポジウム, 2015 年

④加藤聡, 淺野涼太, SOM テンプレートを用いた古文書文字認識に関する研究, 第 30 回ファジィシステムシンポジウム, 2014 年

６．研究組織

(1)研究代表者
加藤　聡（KATO, Satoru）
松江工業高等専門学校・情報工学科・准教授
研究者番号：４０３４２５４７

(2)連携研究者
堀内　匡（HORIUCHI, Tadashi）
松江工業高等専門学校・電子制御工学科・教授
研究者番号：５０２９４１２９

(3)研究協力者
森田　浩旭（MORITA, Hiroaki）
淺野　涼太（ASANO, Ryota）