様式 C-19

科学研究費補助金研究成果報告書

平成２３年５月３１日現在

機関番号：15501
研究種目：基盤研究(C)
研究期間：2007～2010
課題番号：１９５００２１４
研究課題名（和文）　古文献の時代定位システムの開発に関する研究

研究課題名（英文）　A study on Identifying Ages of Japanese Historical Documents

研究代表者
中田　充　（NAKATA　MITSURU）
山口大学・教育学部・准教授
研究者番号：６０３０４４６６

研究成果の概要（和文）：古典文学作品に書かれた手書き文字を対象とした文字認識手法を提案し，それに基づいた認識プログラムを試作した．本手法では，特徴グラフを用いて文字の構造を表現する．認識対象文字と認識用辞書に含まれる既知文字（辞書文字）の類似性を計算し，認識対象文字を最も類似性の高い辞書文字として認識する．次に，源氏物語中に書かれた文字を対象として評価実験を行った．その結果，一文字毎に切り出された文字を対象とした場合の認識率は 76.6%であり，続け字を含む縦一行を対象とした場合の認識率は 54%であった．

研究成果の概要（英文）：We have proposed a character recognition method for consecutive handwritten characters written in Japanese historical documents and implemented a prototype recognition program based on this method. In our method, structures of the characters are represented by feature graphs and dictionary storing known characters is prepared for character recognition. The similarities between feature graphs of target characters and dictionary characters are calculated, and the target character is recognized as the dictionary character with the highest similarity. The computational experiment using our proposed method has been done for the characters written in The Tale of Genji. As the result, we have the recognition rates 76.6% for non-consecutive target characters and 54% for consecutive target characters.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	900,000	270,000	1,170,000
2008 年度	900,000	270,000	1,170,000
2009 年度	800,000	240,000	1,040,000
2010 年度	700,000	210,000	910,000
年度	0	0	0
総　計	3,300,000	990,000	4,290,000

研究分野：情報科学，人文社会情報学
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：手書き文字認識，異体字認識，グラフ理論，特徴グラフ，適合度

１．研究開始当初の背景

古文献の時代定位や真偽判定は，大きくいえば領土紛争や歴史問題における主張の根拠として重要であり必要とされている．また，古文献の成立年代や筆者の特定は，歴史認識の再構築につながるため，学術的にも求められる．これらのニーズに応える技術を確立するために，我々は和文の古文献を取り上げ，

その時代定位や筆者の推定などを可能とするシステムを開発することを目指している．

日本語古文献の写本が書かれた年代を推定する場合，それに書かれている変体仮名の使用傾向（出現頻度情報）に注目する方法が考えられる．例えば，ある古文献の写本の中に，仮名「た」と「ま」の特徴的な変体仮名が書かれており，これらが一文中に同時に出現することは中世以降の文にはあまり見られないと考えられる場合，この写本は中世以前のものであることが推測できる．

我々が最終的に目指しているシステムでは，まず，推定の対象である古文献の原画像中の文字を認識し，その中の仮名文字を元になった漢字(字母)に変換する．その理由は，変体仮名は，一般的に文字コード化されていないため，現在のコンピュータ上で扱うには字母に変換しなければならないからである．次に，対象の古文献における字母の出現頻度情報を求め，それを成立年代や筆者が既に分かっている文献と比較して，その古文献の成立年代や筆者を推定する．

古文献を対象とした文字認識については，これまでにも画像処理技術やニューラルネット等の手法を用いた研究が行われているが，これらの研究は，手動での文字切り出しを前提とする場合には実用的な文字認識が可能であるものの，それを前提としない場合では認識精度に問題がある．また，古文献における仮名字母に関する研究は，特定の文献や限定された時代の文献に関して行われているものの，系統的に時代ごとの字母出現頻度をあつかった研究はなされていない．

２．研究の目的

本研究は研究希望期間内に，次の(1)-(6)の手法・技術を開発して，古文献の時代定位と筆者の特定を行うシステムを実現することを目的とする．

(1) グラフ理論を用いた一連の文字の構造情報に関するモデル化手法：
本研究では，文字認識の基礎情報として，“文字の形に関する情報”（文字構造情報）を用いる．そのために，崩し字・つづけ字で書かれた一連の文字の構造を，グラフ理論を用いて表現するモデル化手法を確立する．

(2) 仮名の字母情報データベースの構築：
変体仮名（崩し字で書かれた仮名や現在の仮名とは字母が異なる仮名を含めた全ての仮名）について，(1)の手法で作成したモデル（グラフ）と字母情報を格納した「字母情報データベース」を構築する．

(3) 原文画像からつづけ字で書かれた一連の文字のグラフを構成する技術の確立：
スキャナ等で取り込んだ古文献の原画像を一行ずつに分割し，(1)の手法に基づいて，行単位の文字構造情報を示すグラフを自動生成する技術を確立する．

(4) グラフによりモデル化された仮名の構造を用いた仮名字母変換技術の確立:
(3)の技術で生成した一行の文字構造を表すグラフの中に，(2)のデータベース中のグラフと同様の構造を発見するためのアルゴリズムを，部分グラフの同定理論を用いて設計する．これにより，一連の文字に含まれる変体仮名一文字を認識して字母に変換する技術を確立する．

(5) 字母出現頻度情報データベースの構築:
(4)の技術を成立年代や筆者が明らかとなっている古文献に適用し，文献別，時代別，筆者別などの視点から字母出現頻度情報を分析し，データベース化する．

(6) 古文献の時代定位や筆者を推定する技術の確立:
推定対象の古文献に対して(4)の技術を適用し，その結果を(5)のデータベースと比較することにより，文献の時代定位や筆者を推定する機能を実現する．

3. 研究の方法

研究の目的に挙げた(1)～(6)について，以下の手順に従い，研究を進めていき，最終的に古文献の時代定位システムを構築する．

(1) グラフ理論を用いた一連の文字の構造情報に関するモデル化手法

① 文字の構造を表現するための枠組みを，グラフ理論を用いて提案する．

② ①の枠組みを用いて変体仮名辞典や代表的な古文献影印本に書かれている文字の構造をグラフ化し，その枠組みを検証する．

③ ①，②に基づいて，古文献に書かれた文字のモデル化手法を開発する．

(2) 仮名の字母情報データベースの構築

① 古文献データベースシステムに関する調査を行い，収集した仮名を(1)の手法を用いてグラフ化する．

② 表現した仮名のグラフと字母の対応情報を格納する字母情報データベースの構造を設計し，仮名のグラフとその字母情報を登録することで字母情報データベースを構築する．

(3) 原文画像からつづけ字で書かれた一連の文字のグラフを構成する技術の確立

① 取り込んだ古文献の画像データを，行単位に分割する技術を開発する．

② 文字の筆使いにおける交差する点（交点）や端点を，画像から求める技術を実現する．さらに，画像のかすれや汚れに対応する技術を実現する．

③ 求めた交点や端点からグラフを作成する技術を確立する．

④ ①～③の技術を用いた一行のつづけ字に対応するグラフを自動生成するソフトウェアを開発する．

(4) グラフによりモデル化された仮名の構造を用いた仮名字母変換技術の確立

① (3)で作成したグラフに，(2)のデータベース中のグラフと同様の構造が含まれるか否かを判定するアルゴリズムを，同型部分グラフの算出アルゴリズムを用いて設計する．
② ①のアルゴリズムに基づいて，原画像中の仮名を字母に変換する文字認識ソフトウェアを実現する．

(5) 字母出現頻度情報データベースの構築

① 時代毎，筆者毎に幾つかの代表的な古文献に対して，(3)と(4)のソフトウェアを適用し，字母出現頻度情報を得る．

② 字母出現頻度情報を時代別，文献別，筆者別に整理分類し，データベース化する．

(6) 古文献の時代定位や筆者を推定する技術の確立

① 推定対象の古文献の字母出現頻度情報を(5)の時代別，筆者別等の字母出現頻度情報と比較し，文献の時代定位・作者推定を行うアルゴリズムを設計する．

② そのアルゴリズムに基づいて，時代定位を行うソフトウェアを実現する．

４．研究成果

研究の目的に挙げた(1)～(4)について，それぞれ以下のような成果を挙げた．

(1) グラフ理論を用いた一連の文字の構造情報に関するモデル化手法に関する成果

古文献中の文字の構造を表現する枠組みとして，“特徴グラフ”を提案した．特徴グラフは，文字の構造を表現するグラフであり，頂点とその接続関係を表す辺から構成される．また，頂点の座標，隣接頂点同士の位置関係などの情報を含む．特徴グラフを用いることで，単に文字の構造を表現するだけではなく，グラフ理論の様々な手法や枠組みを文字認識や文字の整理等に利用可能となった．

(2) 仮名の字母情報データベースの構築に関する成果

字母情報データベースのスキーマを設計し，データの登録，検索，変更等が行えるシステムを設計・構築した．データベースに登録できる情報は，文字画像，字母，平仮名，書名，備考である．次に，変体仮名辞典に掲載されている変体仮名 2195 文字の画像データから生成した特徴グラフとその他の文字情報を登録し，字母情報データベースを構築した．

(3) 原文画像からつづけ字で書かれた一連の文字のグラフを構成する技術に関する成果

特徴グラフを作成する手法を提案し，Java を用いて実装した．特徴グラフは，細線化処理を施した文字画像の，端点，交差点，変曲点を検出し，それらに対してクラスタリング処理，曲頂点の追加処理，直線近似処理などを適用することで作成される．また，実装したプログラムを古文献に書かれた 100 種類程度の手書き文字に適用し，作成された特徴グラフを目視にて確認することで，作成手法の有用性を確認した．

現状では，一文字単位の文字の構造は的確に表現可能となっている．しかし，縦一行に

書かれた一連の文字列のうち，極端に大きさの異なる文字を含む箇所については，無駄な頂点を含む，あるいは，必要な頂点を含んでいないなどの問題がある．この解決は今後の課題である．

(4) グラフによりモデル化された仮名の構造を用いた仮名字母変換技術に関する成果

特徴グラフを用いて表現された文字の構造情報に基づいた文字認識技術を確立するために，まずは“適合度”と呼ばれる指標を定義し，その算出アルゴリズムを提案した．適合度は，二つの特徴グラフの類似性を表す数値であり，二つの特徴グラフを比較して，対応すると考えられる（同じであると考えられる）頂点の個数に基づいて計算される．次いで，そのアルゴリズムを Java を用いて実装し，適合度が正しく算出されるか評価実験を行った．

次に，適合度に基づいた，一文字を対象とした認識アルゴリズムを提案し，実装した．さらに，字母情報データベース（認識用辞書）に源氏物語と変体仮名の手引きから切り出された約 3000 文字の情報が登録した上で，それを用いた認識実験を行い，提案手法の評価を行った．実験では，93 文字の平仮名に対して認識実験を行い，76.6%の認識率（棄却数は 16 文字）を得た．この認識率は実用には不十分であるものの，関連する他の研究と比較して極端に低いものではないと考えられる．また，提案してきたアルゴリズムには 2 種類の方法があったが，評価実験の結果グリッド方式と呼ばれる方法がよりすぐれていることが明らかとなった．

次に，続け字を含む複数文字を対象とした認識アルゴリズムを提案した．それに伴い，認識用辞書の構造を見直し，辞書文字をその構造の類似性に着目してグループ（クラスター）化し，クラスターに属する辞書文字が必ず持つ構造を表す代表グラフを定義した．その後，全てのクラスターの代表グラフを構成する連結成分（基礎成分）からなる基礎成分表を提案した．新たな認識用辞書は，各辞書文字の画像と特徴グラフ，辞書文字の集合であるクラスターとその代表グラフ，そして基礎成分表から構成される．代表グラフを導入することで，形状が少し異なる文字が辞書中に多数存在した場合でも，認識処理に掛かる時間を必要以上に増加させないことが可能となる．

また，提案したアルゴリズムの概要は，①認識対象文字列の画像から特徴グラフ G を作成する，②特徴グラフ G から基礎成分表に含まれる基礎成分と同型の部分グラフ R を求める，③同型部分グラフ R の周囲の領域を辞書文字のサイズに応じてトリミング（切り出し）する，④切り出した領域に含まれる特徴グラフ G の部分グラフと辞書に含まれる文字の特徴グラフとを比較し適合度を計算する， ⑤切り出した領域に含まれる部分を最も適切な適合度を持つ辞書文字として認識する，というものである．

このアルゴリズムを実装したうえで，認識実験を行った．実験において，認識対象文字列は 7 文字前後の異なる手書き文字で構成されている 5 つの文字列である．また，認識用辞書は 30 のクラスターを持つ．実験結果より，19 文字を正しく認識することができ，認識率は 54%となった．この認識率は，実用を考えると十分ではない．その原因として，異なる文字にもかかわらず，その形状が非常に似通っており，人間でも正しい認識が難しい文字や，縦一行に書かれた一連の文字列のうち，極端に大きさの異なる文字が考えられる．今後は，これらの文字を正しく認識できるように，文字認識アルゴリズムを改善する必要がある．

なお，続け字を含む縦一行の文字列を対象とした文字認識プログラムの認識精度が十分ではないため，それを前提とした(5)の字母出現頻度情報データベースと(6)の古文献の時代定位や筆者を推定する技術については，未だ実現できていない．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

① Mitsuru Nakata, Shuichi Nishida, Ryuzo Fukuda, Qi-Wei Ge and Makoto Yoshimura, ″A Method of Recognizing Handwritten Characters in Japanese Historical Documents by Using Feature Graphs″, *INFORMATION*, Vol.13, No.3(B), pp953-966, 2010, 査読有り.

② Ryuzo Fukuda, Mitsuru Nakata, Qi-Wei Ge and Makoto Yoshimura, ″Discussion on Recognition Method for Consecutive Handwritten Japanese Characters by

Feature Graph", *Proc. of ITC-CSCC2010*, pp404-407, 2010, 査読有り.

③ Shuichi Nishida, Mitsuru Nakata, Qi-Wei Ge and Makoto Yoshimura, "Discussion on Handwritten Character Recognition by Feature Graphs", *Proc. of the Fifth International Conference on Information(info2009)*, pp128-133, 2009, 査読有り.

④ Shuichi Nishida, Mitsuru Nakata, Qi-Wei Ge and Makoto Yoshimura, "A Method of Handwritten Character Recognition by Feature Graphs", *Proc. of ITC-CSCC2009*, pp863-866, 2009, 査読有り.

⑤ 稲木奈穂, "古文献を対象とした文字認識のための文字構造 DB の構築", 平成 20 年度山口大学教育学部卒業論文, 2009, 査読無し.

⑥ Masaki Hayashi, Shuichi Nishida, Mitsuru Nakata, Qi-Wei Ge and Makoto Yoshimura, "A Method of Generating Feature Graph for Handwritten Character Recognition of Japanese Historical Documents", *Proc. of ITC-CSCC2008*, pp305-308, 2008, 査読有り.

〔学会発表〕(計 6 件)

① 福田竜三, 松田 直子, 中田 充, 葛 崎偉, 吉村 誠, "日本語続け字を対象とした特徴グラフに基づく文字認識アルゴリズムの改良", 電子情報通信学会 CST 研究会, 2011/1/21, 山口県下関市海峡メッセ下関.

② 林正紀, 西田秀一, 倉持真理子, 二宮亜佐美, 中田充, 葛崎偉, 吉村誠, "手書き文字認識のための特徴グラフ生成アルゴリズムの改善", 電子情報通信学会 CST 研究会, 2009/1/29, 神奈川県産業振興センター会議室.

③ 西田秀一, 林正紀, 倉持真理子, 二宮亜佐美, 中田充, 葛崎偉, 吉村誠, "手書き文字認識のための特徴グラフの類似性判定アルゴリズム", 電子情報通信学会 CST 研究会, 2008/11/7, 大阪, 大阪大学 吹田キャンパス 銀杏会館.

④ 西田秀一, 林正紀, 中田充, 葛崎偉, 吉村誠, "手書き文字認識のための特徴グラフのマッチングアルゴリズムの提案", 電子情報通信学会 CST 研究会, 2008/6/3, 愛知, 名古屋大学 野依記念学術交流館.

⑤ 林正紀, 齋藤名美, 中田充, 葛崎偉, 吉村誠, "古文献を対象とした文字認識のための文字構造グラフの生成法", 電子情報通信学会 CST 研究会, 2008/1/29, 徳島, 徳島大学工業会館 2 階メモリアルホール.

⑥ 崔日男, 齋藤名美, 中田充, 葛崎偉, 吉村誠, 古文献時代定位システムのためのグラフ理論による文字認識モデルの考察, 電子情報通信学会 CST 研究会, 2007/8/31, 島根, 島根大学総合理工学部多目的ホール.

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況 (計 0 件)
○取得状況 (計 0 件)

〔その他〕
ホームページ等 なし

６. 研究組織

(1)研究代表者
中田 充 (NAKATA MITSURU)
山口大学・教育学部・准教授
研究者番号：60304466

(2)研究分担者
葛 崎偉 (KATSU KI)
山口大学・教育学部・教授
研究者番号：30263750

吉村 誠 (YOSHIMURA MAKOTO)
山口大学・教育学部・教授
研究者番号：70141116

(3)連携研究者
なし