2006－CH－72 (3)
2006／10／27

Estrangeloの古文書デジタル化と半自動認識の支援システムについて

Tan Siang Chin[1] 坂田年男[1] 貞廣泰造[2]

1 九州大学芸術工学府、九州大学芸術工学研究院
2 熊本県立大学総合管理学部

概要: 古代シリア語のEstrangeloの古文書のデジタル化と翻訳のための半自動システムの提案を行った。本システムの一つの特徴は統計の世界でよく知られたフリーの統計ソフトRをベースにしていることである。その意義は誰でも使用できる事と、解析のどの時点でも統計的な手法に容易にアクセス可能な点にある。また、システムのGUI化をRと連携できるTcl/TkまたはJavaを用いて行うことにより一層使用しやすいシステムとして提供することができる。また、Rの不得意な部分をCなどで補うことも考えている。二つ目の特徴はEstrangeloの語幹を手がかりに単語の特定を行う点にある。これはradicalを手がかりにする漢字認識と同じ発想である。三つ目の特徴は偏微分方程式方による画像修復法や機械学習・統計学的手法による分類手法など最新のパターン認識手法がシステムに組み込まれていることである。本報告では全体の構想とシステムのデジタル化の部分の考察を行い、GUI化や単語認識過程の部分は検討中であり、構想を述べるにとどめた。最後に、今後の検討課題についても言及した。

Semi-automatic recognition of Estrangelo

Tan Siang Chin[1] Toshio Sakata[1] Taizo Sadahiro[2]

1 Graduate School of Design, School of Design in Kyushu University
2 Department of Administration, Kumamoto Prefectural University

Abstract: A semi-automatic system for digitalization and recognition of Estrangelo is proposed. The first feature of the system is that it is based on the statistical free software R. This means that the system is available for any scholars usage without any cost and the statistical methods could be accessed at any time for their analysis. This system is also developed based on GUI-system by using Tcl/Tk languages, which makes it a more user-friendly system. The second feature uses the word roots concept, which is similar to the case of Chinese character recognition by using the radicals. The third feature uses the most recent highly developed pattern recognition methods and image analysis such as machine learning methods and image recovery method by using partial differential equations.

1 導入

1.1 Estrangelo

紀元前１１世紀頃から使用されてきたAramaic言語の一形態であるSyriacの中の最古のものがEstrangeloである。Estrangeloは経典言語として重要な位置を占め、その古文書を発掘しデジタル化することには意味がある。Estrangeloの22個のアルファベットは右から左に書かれるが、アラビア語同様に単語の中で各文字は連続して書かれるのが普通である。また、単語の中での位置（語頭、語中、語尾）によって字形が変化するなどアラビア文字に共通の文字法を有している。Syriacは横書きであるが、縦書きにしたものがモンゴル文字であると云われている([5],[8])。

1.2 研究の目的

本論文ではEstrangelo文字(古代Syriac語形のひとつ)研究に資するための古文書をデジタル化しかつ翻訳できる簡易ワープロER^2の構想とその研究の一端について解説する。ただし、本研究は緒に付いたばかりであり、継続中であり完成品ではないことをお断りしておきたい。
本研究が提案するER^2は統計学のサークルで世界的に有名なフリーソフトであるRをシステムベースに使用することで多様な統計処理に随時移行できることに利点がある。またフリーソフトRを基盤にしていることは研究者に対するコスト的面での寄与が大きいと考えられる。さらに、このようなソフトでは全体をGUI化することが使用に当たっての負担感を少なくするのに役立つので、RとJavaまたはRとTcl/Tkの組み合わせでGUI化することも視野に入れている。また、最終的にはどのようなGUIシステムが研究者にとって最も使いやすいかを様々なGUIモデルを構築し、比較検討することも目標としている。なお、本システムは辞書を取り替えれば汎用性を持つシステムであることを注意しておきたい。

1.3 研究手法の特徴

古文書の自動的デジタル化プログラムについては最近のいくつかの論文において、特に日本の古文書について議論されている([9],[10])。そこでは、文字の重なり、続きの自動認識など、人ならば簡単な判断がコンピューターには難しい問題の克服に議論の焦点が当てられている。しかし、本研究のもうひとつの特徴はセミオートマティックなシステムの構築にあり、不必要で完全な自動化を目指さず、人の判断を適度に導入した簡易システムを構築することを目指している。「どこまで人の手を借りるか」の基準として「作業が苦痛でない程度」で、自動化の負荷が大きい場合とした。ただし、セミオートマチックなプログラムを作成した上で、できる限り意味のある自動化の研究は継続すべきことは当然である。なお、本研究手法の妥当性を検証するためにPeshittaのwebsiteにある画像Aramaic Lectionary(:about A.D.550, Pierpont Morgan Library New York所蔵)を用いた。

2 研究の構想

古文書のデジタル化・翻訳は、
スキャンした古文書ページの入力 → 単語の切り出し → 単語の判定 → 文字と意味の書き出し

の過程を繰り返すが、各過程でアルゴリズムと人の手の適度な協調作業としてシステムを構築する。具体的には次の工夫を行う。これらは従来の手法とかなり異なるアイデアと手法を用いたシステムと成っていることに注意したい。

(1) スキャンした古文書の汚れやページ内の傷を修復する前処理を行う。そのために、通常はスキャン後の画像にアルゴリズム的に修復を掛けるが、本研究ではページ内の汚れや傷をスキャンする前に、色をつけて他と区別し易くしてからスキャンする。明らかに周囲と異なる色をつけることにより、そこをマークされた領域として偏微分方程式法による画像修復アルゴリズムを掛けることが可能となる。

(2) 単語の切り出しのために、通常の方法と同じく行ヒストグラムや列ヒストグラムなどを利用して、行間隔を検出し、それから縦方向のヒストグラムを利用して単語領域を抽出する。しかし、本研究では古文書をターゲットとしており、行方向の文字の重なりや列方向の単語の重なりが含まれる。そこで、どのような手法であろうと修正すべき箇所が出てくる。そこで、一度ある基準である程度の正確さで行や列の区切り線を入れた後で、Rの locator 関数を用いて修正すべき区切り線の箇所の座標を取得して修正する機能を持たせた。これで、間隙のはっきりしない古文書にある程度対応可能となった。

(3) 切り出された単語に意味（その意味を持つ辞書中の単語）を対応させるために、Estrangelo 特有の語形と文法を利用する。Estrangelo の単語はほぼ３つのパート、接頭辞＋語幹＋接尾辞から成り立っている。これを利用して、単語を一度、接頭辞、語幹, 接尾辞に分解して、対応するそれぞれの辞書を参照する。これらの組み合わせで単語の意味をマッチさせる。これら３つのパートへの分解を半自動化するために、接頭辞、接尾辞のパターンを別画面に表示しておいて、ラベリングに不都合なつながりがあればそこをマウスではさみを入れた上で、ラベリングに掛けて部位を切り出す。これは漢字における部位や部首を切り出してその情報を利用して漢字の意味を検索する方法と同じ考えである（[6] 等）が、完全自動化しないことでいたずらにシステムに負荷を掛けないようにしたものである。

（４）切り出された部位を部位のパターンとマッチングさせるために辞書のサイズにあわせる（拡大、縮小）か不変特徴を用いる。前者のためのひとつの手法に偏微分方程式法による非線形な拡大縮小を組み込む。

（５）それぞれの部位の辞書との突き合せて機械学習・統計的判別法により総合的に単語を特定する。

3 偏微分方程式による画像修復

デジタルカラー画像の偏微分方程式に基づく修復理論は比較的新しい技術である。いくつかのグループによる研究があるが、本研究では Tschumperle と Deriche の手法をRでプログラム化したものを用いる。$I=(I_1,I_2,I_3)$ をカラー画像とするとき、初期画像を $I(t)=I_{original}$ として、次の偏微分方程式の解として修復画像を実現する方法である。

$$\frac{\partial I_i}{\partial t}=\begin{cases} trace(DH_i) & \text{M(x,y)=1 の場合} \\ 0 & \text{M(x,y)=0 の場合} \end{cases}$$

ここで、

$$D=\frac{1}{\sqrt{1+\lambda_+^*+\lambda_-^*}}\theta_-^*\theta_+^{*T}$$

で、H_i は I_i のヘッシアン行列で、$\lambda_\pm^*,\theta_\pm^*$ は構造行列 $G_\sigma=(\sum_{i=1}^3\nabla I_i\nabla I_i^T)*N_{0,\sigma}$ の固有値、固有ベクトルである。ここで、$*$ は合成積を意味し、$N_{0,\sigma}$ は平均０、標準偏差 σ の正規分布とする。$M(x,y)$

は修復領域を指定するマスク関数である。Tschumperle と Deriche の手法ではマスク領域の設定が必要であるが、これをあらかじめ人手で色分けしておけば、マスク領域の検出が容易になる。ただし、実験の結果みると文字抽出のためにこの手法を適用するには、Mask 関数の設定にもう一工夫必要であることがわかった。文字の欠けている輪郭の推定境界を（手動または自動で）作成し、文字境界の内側と外側を区別し、まず最初に内側のみに Mask を掛けて文字を修復し、その後外側のみに Mask を掛けて修復すべきと考えられる（図１、２）。さもないと文字の内側と外側が融合してしまう（文字の一部分が消えてしまう）場合が起こり得るようである。また、スピードの点で C との連携あるいは Tschumperle と Deriche による手法以外の手法 ([3] 等) との比較検討も必要である。

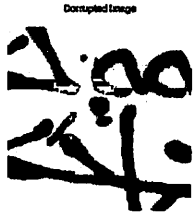

図 1: 傷を文字部位に制限したイメージ

図 2: 傷修復後のイメージ

4 行の検出

古文書の性格上（手書きであるゆえに）、行ヒストグラムが完全には分離しないのが通常である。それでも、手法としては行の検出は基本的には重なりのあるヒストグラムをどのように分離するかに尽きる。本件では、反転画像の数値行列に対して行和をとり、その最大値を max、最小値を min とするとき行和が (max-min)/a+min 以下の行を行の隙間にある点の候補とし、その集合にクラスタリングを行い、クラスター中央値を行線を引くべき位置と定めた。このとき、クラスター個数を与えるために行数を目視で確認して入力する形式をとった。a は自動で 2,3,4,5 と動かし、それぞれの a に対して、作成される行間隔がもっとも一様な大きさになるような a を求めて、この a に対して行線を引いた。行幅の一様性の基準は行幅の分散がもっとも小さいこととした。クラスタリングを簡単にプログラムに組めるのは R をベースにしているおかげである。

5 単語の仮切り出し

5.1 仮切り出し

単語の仮り切り出しは行線をすでに求めているので、各行独立に行う。反転画像の数値行列の列和をとりその平均値を mean とするとき、和が mean/10 以下の列を単語の区切りの候補とし、隣り合う候補点と 5 ピクセル以上はなれている候補点を単語の区切りとした (図３)。これは連続する候補点（区間）

からは1個しか選ばないための工夫である。候補点が密集しないためにクラスタリングではなくこのような基準を採用した。

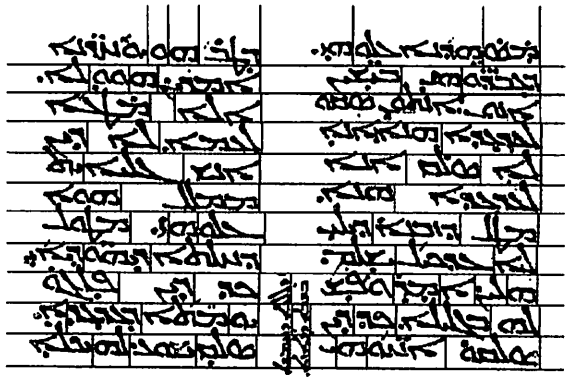

図 3: 単語の仮り切り出し

5.2 本切り出し

単語の間の隙間線を仮切り出しの手順で引いた場合、修正すべき箇所が必ず含まれる。例えば前または後に空白を持つ部位を含む単語では単語の途中で区切れてしまう。そこで、R の locator 関数を用いて、削除すべき区切り、追加したい区切りの座標を画面上で取得し、修正するプログラムを加味した。さらに、行ごとに単語切り出しを行うとき、ある単語では上下の行にまでまたがっているかまたは上下の単語が食い込んできている。そのような単語を画面で選択し、修正するには上下の行のそれぞれ半分の幅と本体の行をまとめた行を同時に考え、そこで labeling を行い連結成分を切り出す。連結成分の面積を S、その連結成分が上の行の半分に含まれる面積を S_1, 下の行半分に含まれる面積を S_2 とし、$max\{S_1/S, S_1/S\} < 0.4$ のときその連結成分を当該単語の一部とし、それ以外は当該単語の一部とはみなさず除去することで修正する（図4，5）。

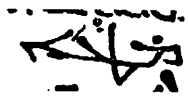

図 4: 上下の行線を越えた単語

図 5: 成功裏に抽出された単語

しかし、この手法では、上下の行の文字と接触してしまっている単語は接触している上下の行の中の部位が当該単語の連結成分として扱われるために、連結成分の面積が大きくカウントされ切り出せない可能性がある（図６，７)。

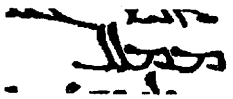

図 6: 上下の行中の単語と接触する単語

図 7: 抽出された単語の失敗例

このような場合は目視によりあらかじめマウスによりはさみを入れて分離しておくと上手く抽出できる（図８，９)。

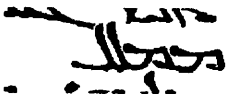

図 8: 抽出された単語の成功例（はさみを入れた図)

図 9: 抽出された単語の成功例

6　部位への切り分け

単語の接頭辞、接尾辞、語幹部位への分解は通常のラベリング手法を適用するが、前処理として、人の手で不必要な部位同士のつながりをマウスではさみを入れて切断した後にラベリングを行うことにした。これもＲの機能を使って、２点を結ぶ直線にそってはさみを入れることにした。ただし、このとき、接頭辞、接尾辞画面上で参照しながらはさみを入れることにする。そのために、接頭辞のリスト、接尾辞のリストを画面に常駐させる。これらの個数はそれぞれ２０個程度なので可能である。語幹に関しては個数は単語の異なる shape が５０程度の場合にもせいぜい４，５個なので、スライド可能な windows に表示すればある程度参照することはできる。そのためにはスライド時に画面が揺らぐ R の windows ではなく、Tcl/Tk のスライド付の windows を立ち上げることにする。

7 偏微分方程式による画像修復部位のリサイズ

接頭辞、語幹, 接尾辞へ切り分けた後、当然それらの部位の大きさは不ぞろいとなる。これを辞書の部位とマッチングさせるには大きさに無関係な特徴量を抽出するか、または大きさをそろえる必要がある。ここでは前者として、log 画像、回転とスケールに関して不変なモーメント、後者として、拡大手法は様々あるが、ここでは偏微分方程式による非線形拡大法の使用も検討する。偏微分方程式法の考え方は、初期画像としてとりあえず従来の方法で作成した拡大画像に対して拡大した点にきちんと元画像 I_{small} の対応する点以外を修復領域の点として偏微分方程式の解画像として拡大画像を実現するものである。Tschumperle と Deriche の手法では以下のマスク関数 M(x,y) を用いる点が画像修復と異なるだけで同じプログラムが使用できる利点がある。

$$M(x,y) = \begin{cases} 0 & N(x/k_x, y/k_y, \epsilon) \text{が元画像の格子点を含む場合} \\ 1 & N(x/k_x, y/k_y, \epsilon) \text{が元画像の格子点を含まない場合} \end{cases}$$

ここで、$N(x/k_x, y/k_y, \epsilon)$ は元画像における $(x/k.y/k)$ の ϵ 近傍に属する格子点の集合で,$M(x,y) = 0$ なる点での $I(x,y)$ は I_{small} の $N(x/k_x, y/k_y, \epsilon)$ 近傍における I_{small} の値による線形補間とする。

8 単語の意味の対応付け

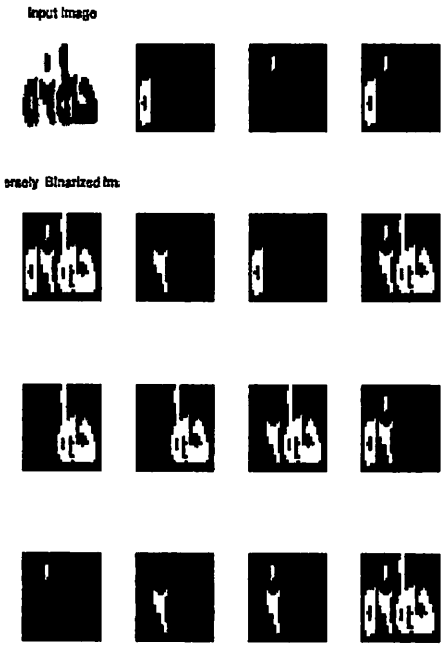

図 10: 単語の連結成分への分解と再合併による意味のある部位の探索

単語の意味づけは接頭辞、接尾辞、および語幹に分解した上で、それぞれの辞書との マッチングの結果を総合的に判断して自動的に行う (図１０)。マッチングさせる特徴量としては (a) 単語の大きさを辞書の中の単語の大きさに調整する、あるいは、(b1) 位置・尺度・回転不変モーメントや (b2)Log-polar 変換した画像の特徴量などを使用することを考えている。さらに、単語の判別には最新の機械学習・統計判別理論を適用する。また、最終判断は人にゆだねるために、辞書を画面上に 呼び出させる機能を持たせる。

9 辞書

次の辞書を用いる。
(1) 単語の画像を収めた辞書。各単語名をつけた画像ファイルの集まり。
絞り込まれた単語の候補から目視で標的単語を決定する際に使用する。
(2) 単語の接頭辞、接尾辞、語幹の名前が入っている辞書。標的単語の各部位を決定したあと、それらを総合して単語を確定する際に使用する。
(3) 接頭辞の辞書。接頭辞の画像ファイルの集合ファイル。 目視のために接頭辞を表示させる際に使用する。
(4) 接頭辞の特徴を収めた辞書。標的単語の接頭辞の特徴と比較して接頭辞を確定するために使用する。
(5) 接尾辞の辞書。接尾辞の画像ファイルの集合。目視のために接尾辞を表示させる際に使用する。
(6) 接尾辞の特徴を収めた辞書。標的単語の接尾辞の特徴と比較して接尾辞を確定するために使用する。
(7) 語幹の辞書。語幹の画像ファイルの集合。
(8) 語幹の特徴を収めた辞書。標的単語の語幹の特徴と比較して語幹を確定するために使用する。

10 GUI画面の提案

R の操作をすべてコマンドラインから行うのはしばしば操作性が悪いと感じる場面がある。そこでメニューでいろいろな機能を選択できるような GUI-system を採用する。R で GUI 操作を可能にする、Tcl/Tk または rJava などの導入を検討した。また、Tcl/Tk との協調で spreadsheet で文字画像を自在に操る手法もあり [4]、これも選択肢の一つとして考えている。具体策は今後の検討課題であるが、いろいろな画面構成を作成し、実際の研究者に使用していただいて、最適なシステムを構築する予定である。

11 結論

Estrangelo のデジタル化と翻訳を行うための、ベースにフリーの統計ソフトRを用いたシステムの構想について紹介した。 本研究は緒に就いたばかりで、これからさまざまな機能を充実させる必要がある。具体的には関数機能の充実、辞書の充実、操作速度の検討、そのためののC言語などとの連携、適切な GUI-system の構築、完成後のパッケージ化、研究者に使用していただいて、必要な機能、使い易い GUI システムとは何かを詰める作業、実際の文献に対して使用してみて、使い勝手を検証するなどの課題が多く残されている。最新の機械学習・統計理論 ([2]), を用いた単語の意味抽出の過程については検討中であり、言及できなかったが次回に報告したい。最後に、本システムは辞書を取り替えれば、ど

のような言語でも使えるフリーの古文書のデジタル化・解読システムになることが予想されることを注意して、本稿の終わりとしたい。

12 謝辞

WEB上の辞書から印刷 → スキャンにより単語の辞書化を許可していただいたWAY-INTERNATIONALに感謝します。また、Syria語の知識を提供してくださった東京女子大学の守屋彰夫先生と佐賀大学の塚本明廣先生に感謝いたします。

13 文献

参考文献

[1] http://cran.r-project.org/

[2] Decoste, D. and Scholkopf, B.(2002). Training Invariant Support Vector Machine, Machine Learning, 46,161-190,2002.

[3] Grassauer, H. A.(2004). A Combined PDE and Texture Synthesis Approach to Inpainting. ECCV (2) 2004: 214-224.

[4] Levoy, M.(1994). Spreadsheets for Images, In Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH, pp. 139-146.

[5] http://www.peshitta.org. Aramaic で書かれたデジタル経典を含む Aramaic の総合サイト.

[6] Shi, D., Damper, R.I, Gunn, S.R.,(2003). Offline Handwritten Chinese Character Recognition by Radical Decomposition, A.C.M. Trans. on Asian Language Information Processing, vol2,No1.

[7] Tschumperle, D. and Deriche, R.(2005). Vector-Valued Image Regularization with PDEs: A Common Framework for Different Applications, IEEE Transactions on Pattern Analysis and Machine Intelligence archive Volume 27, Issue 4, Pages: 506 - 517.

[8] Thackston, W.M.(1999). Introduction to Syriac. IBEX Publishers.

[9] 坪井昭憲・八村広三郎・吉村ミツ (2005). 江戸期版本画像からの文字切り出しの試み. 情報処理学会 IPSJSIG 研究会報告,CH-66(8),pp53-60.

[10] 梅田三千雄・橋本智弘 (2002). 認識処理を援用した文字切り出しによる古文書のキャラクタポスティング. 電機学会論文集C,122 巻、11 号、pp1876-1883.