Vol.2018-CH-116 No.16
2018/1/27

# 歴史的文書画像に対する内容解析への取り組み

寺沢 憲吾[1,a)]

**概要**：くずし字・つづけ字を含む日本語古典籍や，語法が現代文書と異なる近代文書など，通常の OCR（光学文字認識）の適用が困難であるような文書に対し，その内容の解析と理解を目指した著者らの取り組みについて紹介する予定である．中でも，主にワードスポッティングと，それを用いたテキスト化，頻出語・重要語の自動抽出などについて詳しく述べる．

# An Attempt to Analyse and Recognize Historical Document Images

Kengo Terasawa[1,a)]

## 1. はじめに

文書をデジタル化して保存することはすでに一般的となっており，歴史的文書画像もその例外ではない．我が国でも，国立国会図書館，国立公文書館，国文学研究資料館などにおいて，総計 400 万点以上の文献資料がデジタルアーカイブとして蓄積されている．

しかし，蓄積された文献資料の多くはテキストデータ化されていない．これは，従来の OCR（光学文字認識）技術は活字印刷文書に対してはある程度精度の高い認識が可能となってきている一方で，手書き文字，なかでも特にくずし字・つづけ字を含む日本語古典籍や，語法が現代文書と異なる近代文書などに対しては，十分な精度での認識が困難なためである．現状では，こうした文書画像のテキスト化は専門家が手作業で行うよりほかなく，そのため，テキスト化は主に重要度の高い一部の文献に限って行われ，その他の全国各地に数多く存在する文献資料のほとんどは，画像データとして保存されるにとどまっている．こうした文書画像に対しては，全文検索を行うことはできないし，自然言語処理の手法を用いた内容解析などの手も及んでいない．

このような，従来手法ではテキスト化が困難である歴史的文書画像のデジタルアーカイブを有効活用するためには，

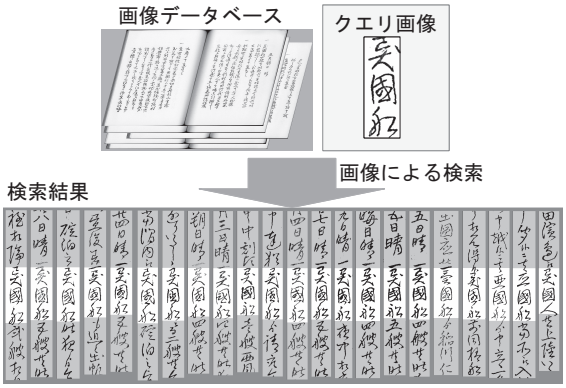


**図 1** ワードスポッティングの概念図

従来手法とは異なる新しい文字認識の手法を開発してテキストデータ化するか，あるいは文字認識に頼らない情報検索の手法を開発するかのいずれかが必要となる．本講演ではそれらに関する著者らの取り組みについて紹介する．以下は紹介を予定している取り組みの事例である．

## 2. ワードスポッティング

ワードスポッティングとは，文書画像全体の中から，ユーザーによって指定されたクエリと同じ単語（あるいは文字列）が出現している位置を特定し，出力するものである（図 1）．ワードスポッティングは，テキストデータによる全文検索の代替手段として有用であるほか，これをインデックス作成に応用することもできる．さらに，読者が解読できない文字列に遭遇した際に，それと同一の文字列が

[1] 公立はこだて未来大学
Future University Hakodate, Hokkaido 041–8655, Japan
[a)] kterasaw@fun.ac.jp

別の文脈で使われている箇所を参照することにより解読の手助けとするという用途もある．実際，我々の開発したシステムは歴史学の文献研究用ツール SMART-GS[1], [2] に組み込まれ，この用途で活用されている．

ワードスポッティングの研究は，英語の手書き文書画像を対象とした Manmatha らによる研究 [3], [4] が有名である．一方で我々は，単語と単語の間にスペースを置かない日本語文書に対しても適用可能なワードスポッティングの手法を提案した．研究初期は [5], [6] のように主成分分析による特徴量を採用していたが，その後 SIFT や HOG のように勾配分布に基づく特徴量を用いる方法がより有効であることを発見し [7]，現在もその方法を主に使用している．勾配分布に基づく特徴量に関しては [8] や [9] でも述べられている．また，我々は開発したシステムを「文書画像検索システム」と名付け，函館市中央図書館所蔵の資料の一部を対象に全文検索をオンラインで利用できるようにして公開した [10], [11]．また，前述した SMART-GS[12] を含め，企業や他機関からの依頼を受けてライセンス提供し，すでに商用化に至ったもの [13] のほか，現在進行中のプロジェクトが複数ある．

## 3. 特徴量の疑似量子化と頻出語・重要語抽出

ワードスポッティングにおいて高次元ベクトルで記述される画像特徴量に対し，LSPC という疑似量子化手法を導入することで，計算コストを大きく削減することに成功した [14]．また，この技法を用い，文字切出し可能な文書のみが対象ではあるが，テキストデータ化されていない文書からの頻出語抽出をも可能にした [15]．さらに出現頻度以外の評価基準も導入して重要語を抽出することまでもすでに可能となっており，現在はデータ量を増加させて実験を行っているところである．

## 4. データベースの整備による文字認識精度の向上

人手での文字切出しと文字認識とでは必要な専門性の水準に差があることに着目し，文字切出しは人力で担い，ワードスポッティングで認識候補を出力し，最後に高度な専門家の目で校正するという方針で翻刻を目指す手法の開発に携わっている [16]．以前は自前で用意できる字形データベースのサイズに限りがあることが研究の障壁となっていたが，近年では国文学研究資料館と国立情報学研究所との協働によるデータベースの整備が進んでおり [17]，これを活用した文字認識精度向上への期待が高まっている．実際に我々もこれを活用した研究をすでに開始しており [18]，また一方で電子情報通信学会 PRMU 研究会主催によるアルゴリズムコンテスト [19] も開催されるなど，この分野の研究は活発化している．

## 5. おわりに

本稿では，歴史的文書画像に対する内容解析へ向けた，著者らの取り組みについて述べた．このような人文学と情報科学が融合した研究を実施するにあたっては，実際の人文学の研究の現場からの多様な声を聞くことが肝要と考えている．本研究会の場に限らず，忌憚のない声をお寄せいただければ幸いである．

### 参考文献

[1] 林晋，永井和，宮崎泉：文献研究と情報技術—史学・古典学の現場から—，人工知能学会誌, vol. 25, no. 1, pp. 24–31 (2010).
[2] Hashimoto, Y., Aihara, K., Hayashi, S., Kukita, M. and Ohura, M.: The SMART-GS Project: An Approach to Image-based Digital Humanities, *Proc. Digital Humanities 2014*, pp. 475–476 (2014).
[3] Manmatha, R., Han, C. and Riseman, E.M.: Word spotting: a new approach to indexing handwriting, *Proc. CVPR'96*, pp. 631–637 (1996).
[4] Rath, T.M. and Manmatha, R.: Word image matching using dynamic time warping, *Proc. CVPR'03*, vol. 2, pp. 521–527 (2003).
[5] Terasawa, K., Nagasaki, T. and Kawashima, T.: Eigenspace Method for Text Retrieval in Historical Document Images, *Proc. ICDAR2005*, vol. 1, pp. 437–441 (2005).
[6] 寺沢憲吾, 長崎健, 川嶋稔夫：固有空間法と DTW による古文書ワードスポッティング，電子情報通信学会論文誌，vol. J89-D, no. 8, pp. 1829–1839 (2006).
[7] 寺沢憲吾, 長崎健, 川嶋稔夫：勾配分布特徴量による高精度手書き文字検索，画像の認識・理解シンポジウム (MIRU)2006 講演論文集，pp. 1325–1330 (2006).
[8] Rodríguez, J.A. and Perronnin, F.: Local gradient histogram features for word spotting in unconstrained handwritten documents, *Proc. ICFHR2008*.
[9] Terasawa, K. and Tanaka, Y.: Slit Style HOG Feature for Document Image Word Spotting, *Proc. ICDAR2009*, pp. 116–120 (2009).
[10] 寺沢憲吾，川嶋稔夫：文書画像からの全文検索のオンラインサービス，人文科学とコンピュータシンポジウム「じんもんこん 2011」, pp. 329–334 (2011).
[11] http://records.c.fun.ac.jp/
[12] https://osdn.jp/projects/smart-gs/
[13] http://www.ozorashapub.net/
[14] Terasawa, K. and Tanaka, Y.: Locality Sensitive Pseudo-Code for Document Images, *Proc. ICDAR2007*, vol. 1, pp. 73–77 (2007).
[15] 細谷拓史，寺沢憲吾：LSPC を用いた活字文書画像における頻出文字列抽出，情報処理学会第 75 回全国大会 (2013).
[16] 山本純子，大澤留次郎：古典籍翻刻の省力化：くずし字を含む新方式 OCR 技術の開発，情報管理，vol. 58, no. 11, pp. 819–827 (2016).
[17] http://codh.rois.ac.jp/
[18] 釜谷勇輝，山田雅之，目加田慶人，長谷川純一，檜山幸夫，東山京子，中貴俊，宮崎慎也，寺沢憲吾，川嶋稔夫：近代公文書自動解読のための手書き字形データセット構築，平成 29 年度電気・電子・情報関係学会東海支部連合大会，B5-9 (2017).
[19] https://sites.google.com/view/alcon2017prmu/