文書画像からの全文検索のオンラインサービス

寺沢 憲吾　　川嶋 稔夫
公立はこだて未来大学 システム情報科学部 情報アーキテクチャ学科

本論文では，公立はこだて未来大学が函館市中央図書館と連携し，同図書館所蔵のコンテンツと本学の開発した文書画像からの全文検索技術を組み合わせた「文書画像検索システム」を開発し，インターネットを介したオンラインサービスとして一般市民向けに公開を行った取り組みについて紹介する．まず地域コンテンツを公開する際の画像による全文検索の有用性と，それを実現するための既存の方法について説明する．さらに，その技術の利便性を高めるための実装上の工夫と，インタフェースの整備についても述べる．最後に，現行システムの公開状況について述べる．

Word Spotting Online

Kengo Terasawa　　Toshio Kawashima
Department of Media Architecture, School of Systems Information Science,
Future University Hakodate

This paper introduces the online word spotting system that was released by Future University Hakodate in collaboration with Hakodate City Central Library. Word spotting is the task of retrieving a text region that has a similar appearance to a query image specified by the user. We discuss the utility of word spotting for manipulating the digital archives. After we describes the state-of-the-art techniques of word spotting, we also discuss how we have improved the connectivity of our tools. The summary of the access traffic is also reported.

１．まえがき

情報通信技術の発展に伴い，多くの文献や史料がディジタル画像化され，インターネットを通して広く一般に公開されている．これにより従来は図書館や博物館の奥深くに収納されていた貴重な文献や史料を，一般のユーザがネットワークを介して容易に入手し，閲覧することが可能となってきている．

しかし，こうしたディジタル技術が文献や史料の共有・流通を促進している一方で，公開された文献や史料を有効に使うための技術はまだ十分に開発されているとは言えないのが現状である．たとえば，文献や史料の多くは画像形式で保存・公開されており，そのままでは全文検索ができず，必要な情報へ素早くアクセスすることが難しい．

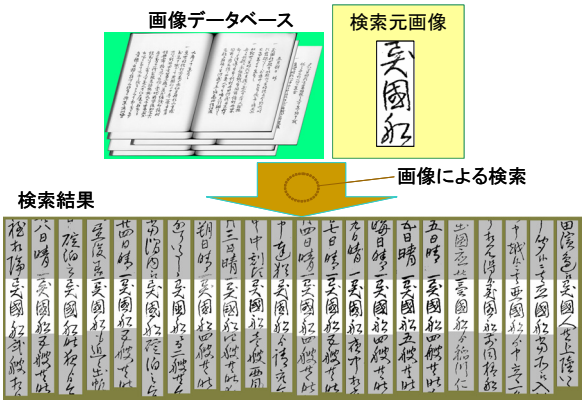
こうした問題を解消するための一つの方法として，すべての文書画像に文字認識を施してコード化（テキストデータ化）を行うことが考えられる．この方法は有力ではあるが，いくつかの理由から万能ではない．まず第一に，画像を正確にテキストデータに変換することは必ずしも容易ではない．手書きの文書画像を対象にOCR（光学文字認識）技術を用いてテキストデータ化する方法は，郵便番号や住所などのように文字種や語彙が限られている場合にはすでに実用化されているが，語彙の限定されていない一般の手書き文書や歴史的文書を対象とする場合には適用が困難である．対象が活字文書であったとしても，OCR の精度は 100%ではないため，正確なテキストデータを作成しようとするならば専門家の手による確認および修正作業が不可欠である．極めて重要度が高い史料に対してならばこの方法でもよいかもしれないが，地域の図書館などに所蔵されているあらゆる文書をこの方法でテキストデータ化することは，コストの問題から見て現実的ではない．

さらに，OCR は言語や書体に依存した技術であるため，時代が古いものや特殊な文字あるいは用法を含むもの，また十分なサンプルが確保できないものに対しては，市販の OCR では対応できない．対象とする文献に対する特別な OCR を開発することは，サンプルが十分にあれば理論的には可能であるが，やはりコストの問題で現実的ではない．ましてサンプルが十分にない場合や未解読文字を対象とする場合は，そもそも原理的に不可能である．

第二の理由として，林ら[1]の指摘するように，文献研究の現場ではある文字の読みを定めること（翻刻）自体が研究行為の一部であり，その解釈が必ずしも一意に定まらず，複数の意見があることも珍しくないということが挙げられる．すべての画像にテキストデータを対応させる方法では，こうした場合に対応することができないばかりか，無理に行えば有害にもなり得る．

こうした背景を踏まえ，我々は文書画像から文字の見かけに基づいて画像検索を行うことに

(c) Information Processing Society of Japan



（図１）画像による全文検索（ワードスポッティング）の例

よる全文検索システムをかねてから開発し，改良を続けてきた [2,3]．（図１）はその例を示したもので，全 182 ページの画像データベースから，クエリ画像「異国船」を検索した場合の結果が表示されている．このシステムにより，ユーザは画像内の任意の文字列をクエリとして，全文検索を行うことができる．このような文字の見かけに基づく全文検索の技法は「ワードスポッティング」と呼ばれ，Manmatha et al. [4] の研究をその端緒として，各国で広く研究が行われている．前述の林ら[1]は，特に人文学の研究という視点から，文献の電子化において画像化を基本とするという画像化主義文書処理を提唱しており，画像化文書と，それに対する検索を取り入れることで，コード化に伴う現実的困難の多くは大幅に解消できるとしている．我々の研究のアプローチもこの方針に沿ったものとなっている．

もちろんこれらのアプローチは，文字認識によるコード化を不要だと主張するものではない．文字認識が可能な場合には文字認識を施して，画像データとテキストデータを連動させつつ保持することは理想的である．地域の図書館が膨大に所蔵しているような地域史料の有効活用のためには，文字認識の利便性と画像による検索の汎用性とを踏まえた上で，相互に補いつつ活用していくことが必要である．

本論文では，公立はこだて未来大学が函館市中央図書館と連携し，同図書館所蔵のコンテンツと本学の開発した文書画像からの全文検索技術を組み合わせたシステムを開発し，インターネットを介したオンラインサービスとして一般市民向けに公開を行った取り組みについて紹介する．第２章でワードスポッティングに関する関連研究について概観した後，第３章で今回採用した検索システムのしくみについて述べる．第４章では，検索エンジンを他のソフトウェアと連携して使用することができるようにすることの重要性を指摘するとともに，それを実現するために我々が採用した実装上の工夫について述べる．第５章で今回公開したシステムのユーザインタフェースについて述べ，第６章で現行システムの公開状況について述べた後，最後に今後の展望を述べてまとめとする．

Query: supermarket

Mr. O'Sullivan was hanging out in the super-market yesterday.

（図２）検索単語が行をまたいで出現する例．単語"supermarket"は，行切出しを前提とする手法でのみ見つけることができる

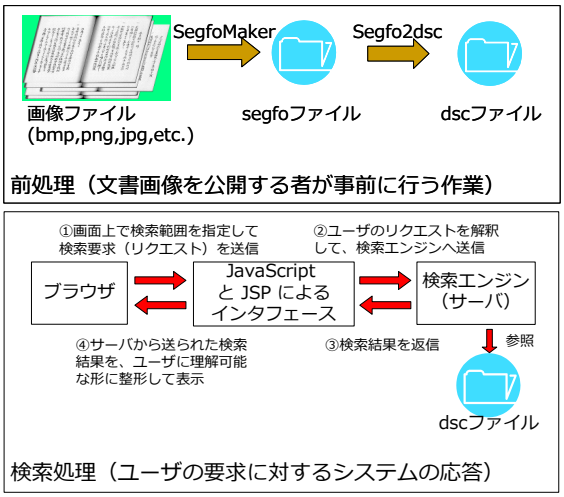
２．関連研究

画像による全文検索（ワードスポッティング）に関しては，文献や史料の特性に応じてさまざまな研究事例がある．Manmatha et al. [4] の研究はこうした研究の端緒とも言えるもので，英語の手書き文献に対して，単語毎に切り出された画像同士の間でマッチングを行うことで，全文検索が可能であることを示している．その後，特徴量やマッチングの識別手法などを改良することによって精度を高める研究が行われ，Rath and Manmatha[5] では，ジョージ・ワシントンの手稿文を対象に，72%を超える精度で検索が可能であることが示されている．

ただし，これら[4,5]で提案された方法は，単語が分かち書きされているため画像を単語ごとに分離すること（単語切出し）が可能であること，文字がベースラインと呼ばれる仮想的な線の上に記述され，標準的な高さ（エックスハイト）や上への突出（アセンダー），下への突出（ディセンダー）などが定義可能であることなどといった，英語などの西洋言語ならではの特性に強く依存したものになっている．一方で我々は，単語切出しが難しい日本語の毛筆文書にも適用可能なワードスポッティングの方法を[2]において提案した．ここでは，日本語のような単語切出しが難しい文書においても行ごとの分離（行切出し）は比較的容易であるいう点に着目し，特徴量記述と DTW (Dynamic Time Warping) の手法を組み合わせることで全文検索が可能であることを示した．その後，特徴量やマッチング手法などに関する改良を加え，[3] では幕末の毛筆手書き文書である「亜国来使記」に対して 97～99%，ジョージ・ワシントンの手稿文に対して，79%の検索精度を達成した．

単語切出しを前提とする[4,5]の研究，単語切出しは前提とせず行切出しを前提とする[2,3]の研究は，いずれも切出し結果が正確であること

(c) Information Processing Society of Japan



（図３）公開した文書画像検索システムの概要

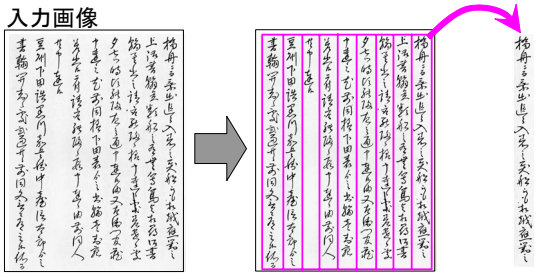
を前提としている．これらに対し，単語切出しも行切出しも必要としない研究（たとえば[6]）も存在する．この方法は画像内のあらゆる位置から類似部分を探すため，一見最も汎用性が高いように感じられるが，必ずしもそうではない場合も存在する．たとえば（図２）のように探したい単語が行をまたいで存在する場合，我々の手法[2,3]ではこれを見つけ出すことができるが，[6]の方法ではこれを見つけ出すことができない．

このように，ワードスポッティングにはさまざまな方法が提案されており，どの方法も万能ではない．特定の言語や文献に対する検索精度をひたすら高めることが必要であれば，それに適した方法を用いるべきだし，逆に多少の検索精度を犠牲にしても汎用性を求めるのであれば，それに合わせた手法を用いることも可能である．いずれにしても，適用しようとする文献の性質や，検索システムを公開する者ならびにそれを利用するユーザのニーズに応じて，どの手法を用いるかを検討していくことが必要である．

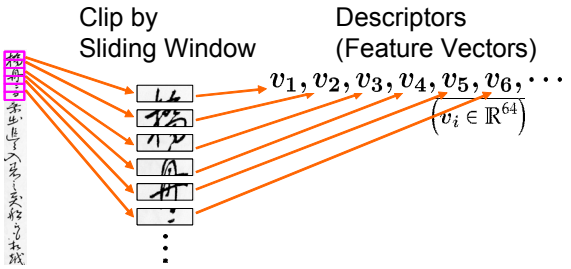
３．画像に基づく全文検索のしくみ

この章からは，函館市中央図書館所蔵のコンテンツの一部に対する全文検索のオンラインサービスとして，公立はこだて未来大学が一般市民向けに公開したシステムについて紹介する．まずこの章では，このシステムに用いた全文検索の手法について述べる．

今回公開したシステムでは，文献[3]で公表した検索手法を用いている．この手法は，行切出しを前提とし，行の書記方向に合わせて部分領域を切り出すウィンドウ（小窓）を移動させつつその部分領域を特徴量で記述し，そうして得られる特徴量ベクトルの系列に対し，CDP（連



（図４）行分離処理（行切出し）



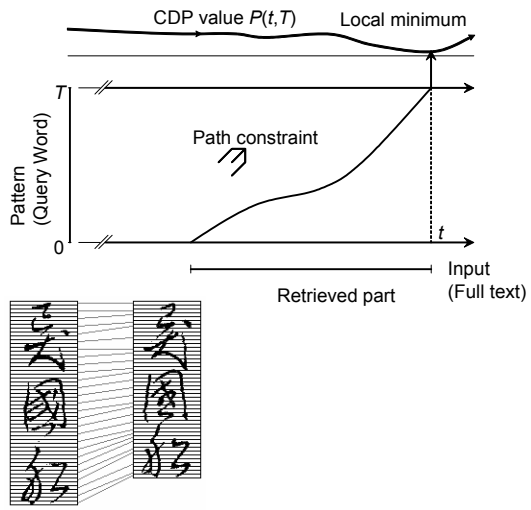
（図５）特徴量記述

続 DP，Continuous Dynamic Programming）[7] を用いてマッチングを行うことで，全文検索を実現している．以下にその概要を述べる．

本システムは，システム公開者が事前に行っておく作業である「前処理」と，前処理によって作られた特徴量ファイルを用いて実際にユーザが検索を実行する「検索処理」に分けることができる（図３）．「前処理」は公開する文献１編につき１回ずつ発生する作業であり，ある程度は自動化されているが，現状では最終確認は人間が目視により行うこととしている．一方の「検索処理」は，ユーザの検索要求１回ごとに発生する処理であり，完全に自動的に行われる．検索処理の実行時間は文献の長さに依存して変化するが，今回公開した文献に対しては数秒から十数秒程度である．

まず前処理について述べる．前処理の第一段階は，入力画像に対し，行分離処理を施すことである（図４）．この処理は，入力画像に対し，画素値の書記方向への射影を計算し，そのピーク位置を行の境目とすることで自動的に行分離位置の推定を行うものである．しかし，この推定結果は文献の性質によっては誤りを含むことがあるため，その場合，作業者が目視で確認した後に個別に修正することにしており，そのための専用のソフトウェア（SegfoMaker）を用意してある．自動推定に関する研究を進めて精度を高めていくことも可能だろうが，この作業は１文書に対して公開前に１回だけ行えばよい処理であるので，現在のところそうした高精度化

(c) Information Processing Society of Japan

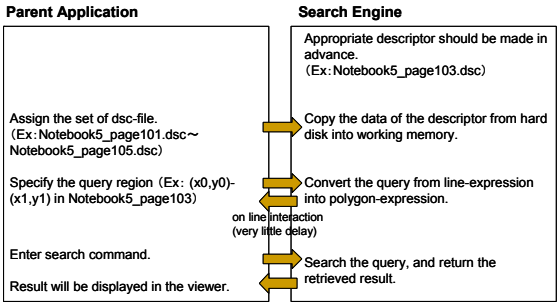


（図６）CDP の概念図．CDP を用いることで，手書き文字画像が伸縮方向に変形しても検索可能になる

はあまり追求していない（今後の研究の余地があるところである）．こうして作成された行の分離位置情報は，ファイル（segfo ファイルと呼ぶ）に格納される．

次に，こうして分離された各行の画像を，さらに細かい画像に分割して，特徴量を記述していく．画像を分割する単位としては，文字ごとに分割できれば理想的であるが，（図４）に示すような毛筆手書き文字の場合に，これを自動的に文字ごとに区切ることは極めて難しい．そこで我々のシステムでは，細長い固定長の幅のウィンドウを書記方向に移動させながら分割するという方法を採る（**図５**）．この方法はスライディングウィンドウ法と呼ばれ，ワードスポッティングに関する多くの既存研究で用いられている．分割された細かい画像を記述するための特徴量としては，画像の画素値の勾配ベクトルを計算してブロックごとに方向別に集約する，HOG[8]と同様のものを用いている．この処理のためにも専用のソフトウェア（Segfo2dsc）が用意されており，画像と segfo ファイルを与えると，特徴量を記述したファイル（dsc ファイル）が自動的に出力されるしくみとなっている．１画像あたりの処理時間は画像内に含まれる文字数によって大きく異なるが，おおむね 10 秒～2 分程度である．行切出し処理とは異なり，ユーザによる確認は必要なく，大量の画像を処理する際にもバッチ処理をしておけば単に待つだけでよい．

このようにして，文書画像全体は特徴量ベクトルの系列 $v_1, v_2, \ldots v_N$ に変換され，dsc ファイルに保存される．ここまでが前処理である．文字認識などを用いて画像内の文字部分すべてにテキストデータを割り当てる方法に比べると，必要な労力は大幅に小さい（人間による作業は，行切出しを確認するだけである．今後の研究次第ではそれすら簡略化される可能性がある）．しかも，文字認識の確認が専門家でなければできない作業であるのに対し，行切出しの確認は特に専門的知識を必要としない作業であり，実行ははるかに容易である．こうした容易性が，我々の提案する手法の大きな利点である．



（図７）親子間通信

次に検索処理について述べる．検索処理は，ユーザの検索要求に応じて，前処理で作成された dsc ファイルを読み出し，読み出した特徴量ベクトルの系列の中から，ユーザが検索したい部分に相当する部分列と最も類似している部分を探すという方法で行う．類似性の判定には，CDP（連続 DP，Continuous Dynamic Programming）[7] を用いる．CDP は DTW と同様に，手書き文字の伸縮変形を許容しつつクエリ系列と類似した系列を探索する手法であるが，局所的な伸縮倍率を 50%から 200%までに制限するようにマッチング経路の制約条件を導入することで，始点と終点が明示されていない系列データからの探索を線形時間で実行することが可能となっている．これにより，本研究のように単語切出しを前提としない場合にも検索を高速に行うことができる（**図６**）．

以上が基本的なプロセスであるが，特徴量計算や検索は，ベクトル系列の長さに比例した計算コスト（計算時間およびメモリ量）を要するので，文字切出しが可能であれば，スライディングウィンドウによるものよりも文字切出しを行った方が計算コストの面で都合がいい．今回公開対象とした文献は手書き文書１編と活字の新聞アーカイブ１編であるが，このうち活字の新聞アーカイブに対しては，文字切出しを行って特徴量を作成した．文字切出しには，商用の OCR ソフト（メディアドライブ社の活字文書 OCR ライブラリ v.7.0）を用いた．前述の通り，古い新聞に対して文字認識は高い精度を示さない（この文献では 62.8%）が，文字切出しは，使い方を工夫すればかなり高い精度（この文献では 96.3%）を達成することができる．この使

(c) Information Processing Society of Japan

い方の工夫についての詳細は文献[9]に記してある．OCR ソフトによる文字切出しの精度は100%ではないが，多少の誤りを含む場合であっても前述の CDP の効果により，正しい文字列を検出することがある程度までは可能である．

４．検索エンジンのツールチェーン化（入出力プロトコル）

前章で述べたような画像による全文検索（ワードスポッティング）は，文献研究のさまざまな局面で有用なものであると信じるが，多くの場合，全文検索を単体で用いるのではなく，他の様々な情報処理ツールと組み合わせることでより有益なシステムが構成されるものと思われる．ソフトウェアの利便性を高め，活用を促進するためには，ソフトウェアはそれ単体でのみ使われるものとして提供されるのではなく，そのソフトウェアの出力が別のソフトウェアの入力になる，というように，連鎖的に使われることが望ましい．そのようなソフトウェアはツールチェーンと呼ばれ，過去のじんもんこんシンポジウムにおいてもその整備の重要性が指摘されている[10]．

我々は，検索エンジンの開発にあたり，それを自己のシステム内のみに完結させるのではなく，他のシステムと連携して使われることが可能になるよう意識した．そうした連携を可能にする方法として，我々は標準入出力をパイプで接続して通信するという方法を採用した．この方法は，他の方法（たとえばソースコードを公開する方法や，DLL または OCX を公開する方法など）と比べ，言語やプラットフォームに依存することがないため汎用性が高く，かつ実装も容易であるという利点がある．実際に親アプリケーションが検索エンジンを呼び出して使用する場合の通信例を（図７）に示す．親アプリケーションが「データ集合 X の中から p を検索せよ」という要求をする場合，親アプリケーションは検索エンジンの標準入力に，指定の書式に従って X と p の情報を送信する．検索エンジンはその情報を受け取った後，検索を実行し，その結果を標準出力へ出力する．親アプリケーションはこの出力結果を受け取り，表示するなり解釈するなりの処理を行う．こうした標準入出力を介した通信については，独自仕様ではあるがプロトコル[11]を作成し，ホームページ[12]で公開している．

こうした工夫により，検索エンジンとそれを呼び出す親アプリケーションとは明確に分離され，それぞれの開発者はそれぞれの作業に集中できるようになった．その効果の事例として，たとえば我々はウェブサイトのインタフェース部の開発を外部に委託することができた．また，我々の検索エンジンの上にインタフェースを整

（図８）実際に公開されている「文書画像検索システム」のタイトル画面．
URL: http://records.c.fun.ac.jp/

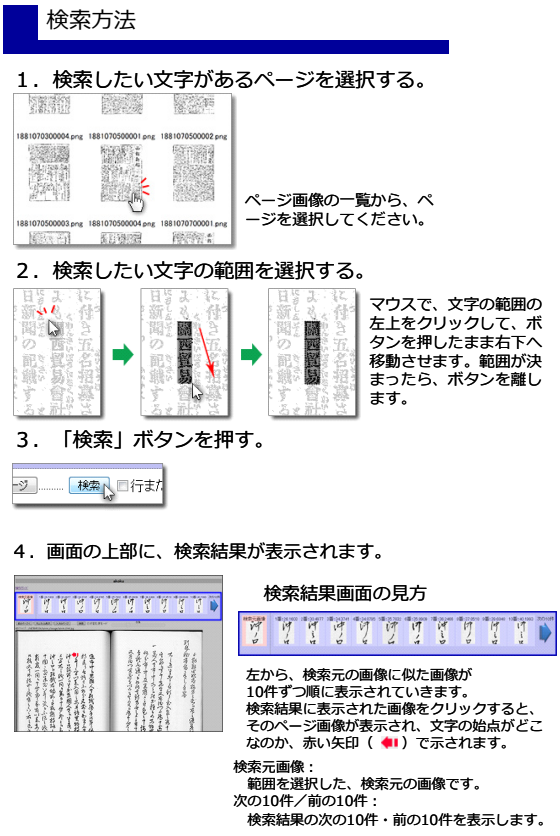
備してパッケージソフトとして売りたいとする民間企業との連携も円滑に行うことができた（現在ソフトウェア使用許諾契約を締結済みで，近日中に発売が開始される予定である）．また，実際に他のシステムに我々の検索エンジンが部品として組み込まれた例として，林らの開発した文献研究用ツールである SMART-GS[1] が挙げられる．SMART-GS は検索の他にリンクやマークアップ，コメントの添付などの機能を持つ多機能なツールであるが，それらの機能の開発は，検索エンジンの開発とは独立に行うことが可能である．

また，検索エンジンのツールチェーン化のもう一つの利点として，我々の策定したプロトコルに準拠した検索エンジンが増えてくれば，親アプリケーションが必要に応じて検索エンジンを交換しながら利用するということも可能になる．前述の通り，特定の言語あるいは特定の文献にのみ高い精度を示す検索エンジンが存在することはありえるので，必要に応じて交換可能であることは望ましい性質である．

５．ユーザインタフェース

ウェブサイトのインタフェースは，ブラウザに対して別途プラグインを要求しないなどの利点から， JavaScript と JSP で作成した．タイトル画面および検索実行画面の例を（図８），（図９）に示す．画面上で検索領域を指定し，検索ボタンをクリックすると，その要求内容が JavaScript と JSP により解釈され，検索エンジンへ送信される．検索要求を受け取った検索エンジンは検索処理を行い，その結果を出力する．JavaScript と JSP の出力処理部はそれを

(c) Information Processing Society of Japan



（図９）オンライン検索システムの操作方法

受け取り，ユーザに理解可能な形に整形して，ユーザの操作するブラウザへ情報を送信する（図９，図３）．

なお，前述の通り，検索エンジンとインタフェースは分業開発が可能になっており，実際にこのインタフェースの開発は外部に委託することで行った．

６．公開状況

平成 23 年 5 月 27 日より公開を開始した．公開開始時のコンテンツとしては，毛筆手書きの文書として「亜国来使記」，古い活字文書として「函館新聞」の２つを採用した．「亜国来使記」は全 182 ページ，「函館新聞」は 1 年分 771 ページを公開した．html サーバは専用のものを用意し，本学に設置した[13]．また，函館市中央図書館デジタル資料館[14]のページからもリンクを設定し，誘導した．

サーバは単体で運用しており，また検索処理は計算負荷が大きいため，実稼働時の負荷に対する耐性が心配されたが，公開開始以降，これまでに最高で 1 日 42 件のアクセスがあったが，トラブルなく処理できている．また，利用者の方からの問い合わせや改善要望などもすでに数件来ており，サービス提供開始の滑り出しは上々と言える．

７．今後の展望

今回公開したオンライン検索システムに関する今後の展開としては，函館市中央図書館と引き続き連携し，コンテンツを増強していく予定である．また，我々は本システムがディジタルアーカイブの利用促進の一助となることを望んでおり，各地の図書館や博物館から利用したいとの依頼があれば積極的に応じる予定である．

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] 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.
[3] K. Terasawa and Y. Tanaka: Slit Style HOG Feature for Document Image Word Spotting, International Conference on Documents Analysis and Recognition, ICDAR2009, pp.116-120, 2009.
[4] R. Manmatha, C. Han and E. M. Riseman: Word Spotting: A New Approach to Indexing Handwriting, Computer Vision and Pattern Recognition, CVPR, pp. 631–637, 1996.
[5] T. M. Rath and R. Manmatha: Word spotting for historical documents, Int. J. on Document Analysis and Recognition, vol.9, no.2-4, pp.139-152, 2006.
[6] Y. Leydier, F. Lebourgeois, and H. Emptoz: Text search for medieval manuscript images, Pattern Recognition, vol. 40, no. 12, pp. 3552--3567, 2007.
[7] R. Oka: Spotting Method for Classification of Real World Data, The Computer Journal, vol.41, no.8, pp.559-565, 1998.
[8] N. Dalal and B. Triggs: Histograms of Oriented Gradients for Human Detection, Computer Vision and Pattern Recognition, CVPR, pp.886-893, 2005.
[9] T. Shima, K. Terasawa, T. Kawashima: Image Processing for Historical Newspaper Archives, International Workshop on Historical Document Imaging and Processing, HIP, pp.127-132, 2011.
[10] 守岡知彦: データを生み出すデータのために, 人文科学とコンピュータシンポジウム論文集, pp.13-18, 2008.
[11] 寺沢憲吾, 猪村元, 田中譲: 交換可能なワードスポッティング検索エンジンの設計, 画像の認識・理解シンポジウム（MIRU）2008 講演論文集, pp.1469-1474, 2008.
[12] http://km.meme.hokudai.ac.jp/people/terasawa/WS/
[13] http://records.c.fun.ac.jp/
[14] http://www.lib-hkd.jp/rein/

(c) Information Processing Society of Japan