Vol.2013-CVIM-188 No.21
2013/9/3

文書画像検索による対応テキスト抽出を用いた翻刻文校正支援システム

服部 雄輝[1,a)]　寺沢 憲吾[2,b)]

概要：本稿では，文書画像に対する画像検索を活用し，対応テキストの抽出を利用して古文書翻刻文の校正を支援するシステムを提案する．翻刻は，古文書などの古い字体で書かれた原文を，現代の活字体に直すことを指す．歴史的価値のある文化財資料を電子化し活用するデジタルアーカイブの取り組みにおいて，その内容をテキストデータ化する際に翻刻が行われる．古文書は毛筆手書きによる表記が多く，自動認識は実用段階ではないため，翻刻作業の大部分は手作業になる．しかしその場合，文書中の同じ語に対する表記の違いが生じる可能性があり，翻刻文の校正を行う必要がある．目視による作業は作業量が膨大となるため，校正を支援するシステムへの需要が存在する．ただし，テキスト検索を用いた場合は，作業者が翻刻文の異なる表記を網羅する必要があるため，校正漏れが生じやすい．本提案システムでは，検索の対象を文書画像とし，文書画像と翻刻文の各行を互いに対応付けた上で，画像検索を行う．この結果から，画像の各位置に対応するテキストを抽出することで，翻刻文の表記が異なる語の発見を容易にし，校正作業の負担を軽減する．抽出結果の表示手法には KWIC (Keyword In Context) を採用し，ユーザーの翻刻文校正作業を補助するシステムの実装を行った．

キーワード：デジタルアーカイブ, 翻刻支援, 文書画像検索, 対応語抽出, Keyword In Context

1. 序論

文書をデジタルデータ化して扱う機会は，近年増加の一途をたどっている．その代表的な例として，デジタルアーカイブの取り組みが挙げられる．デジタルアーカイブは，歴史的価値のある文化財をデジタルデータとして記録する取り組みを指す．日本では，2003 年に発表された国家戦略 “e-Japan 戦略 II” によって，デジタルアーカイブの活用の推進が掲げられた頃から，その取り組みがより広がってきている．その一例として，2005 年より，歴史的価値のある国家の公文書を収蔵する機関である国立公文書館が，インターネットによるデジタルアーカイブの一般公開 [1] を開始した．デジタルデータ化された文化財には，有機物に発生する風化や経年劣化，自然災害による被災の心配がほとんどなく，さらにデータ情報の複製や伝送が容易になることで，有益な文化財を様々な場面で活用することを可能とする．以上のことから，デジタルアーカイブは経年と共に，ますますその重要性が高まっている．

デジタル化される文化財の中でも，本研究では古文書に着目している．古文書は，イメージスキャナに代表される画像入力装置を用いることで，文書内容を高精細な画像データ（本研究では “文書画像” と呼ぶ）として扱うことが容易に実現できる．他方で，文書の内容をテキストデータに変換することは難しい状況にある．特に古文書に関しては，印刷技術の普及していない時代の書物であり，その大半の文書は手書きによるものである．そのような文書に対して光学文字認識（OCR）による文字の認識を行うとなると，文字の見た目，文法などの違いにより，現在確立されている OCR では汎用的な読み取りを行うことができない．よって，古文書に対する OCR は，すぐに使用できる実用的な段階には至っているとは言えない．以上より，古文書の内容をテキストデータ化する作業を自動化することは難しい．

1.1 翻刻作業

デジタルアーカイブにおいてテキストデータを作成する際，一般的に行われているのは “翻刻” と呼ばれる作業である．翻刻とは，古い字体で書かれている文書を，現代で用いられる活字体に書き起こす作業のことを指す（**図 1**）．

[1] 公立はこだて未来大学大学院
Graduate School of Future University Hakodate
[2] 公立はこだて未来大学
Future University Hakodate
[a)] g2112029@fun.ac.jp
[b)] kterasaw@fun.ac.jp

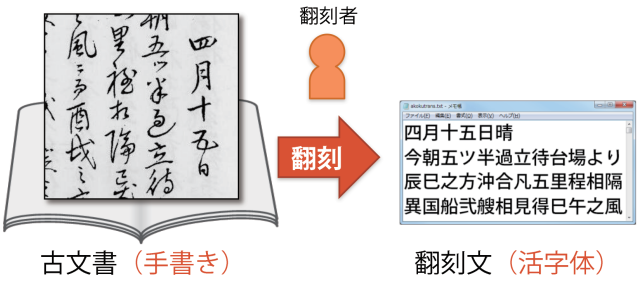

図 1 文書の翻刻
Fig. 1 The transcription of a document

この作業を行うことで，読み取りのために知識が必要な古文書がより理解しやすくなり，現代における歴史資料の活用の促進につながる．翻刻作業に従事する作業者は，行書体や草書体のような古い字体に対し，ある程度の知識が必要である．翻刻対象となる文書内の文章量が多かったり，歴史的価値のある文化財になると，複数人で文書の翻刻作業を分担して作業する，いわゆる分担翻刻が行われることもある．

一方で，このような翻刻作業は手動であるがゆえに問題も発生しうる．翻刻したい語を活字体で表記する際，表記できる字が複数存在することで，複数箇所に現れた語の間で表記の相違が生じることがある．具体例を挙げると，原文で「國」と書かれている文字において，現代の表記に沿って新字体の「国」で表記するか，文書の表記を尊重した旧字体の「國」で表記するか，といった表記の相違が存在する．翻刻方針や解釈にもよるが，特に翻刻を複数人で分担している場合，品質の高い翻刻テキストデータの作成のためには，この相違が複数個所にある語で発生した際，最終的にいずかに統一する必要がある．よって，品質の高い翻刻文を作成するためには，最終的な翻刻文を校正する必要がある．

しかし，作業者にとって翻刻文の校正作業は負担のかかる作業である．前述のように，表記の違いが問題となった場合，作業者はそれぞれの表記箇所を確認しなければならない．しかし，そのような表記のぶれは，通常のテキスト検索だけでは発見することができない．異なる表記のパターンをあらかじめリスト化し，それらを網羅的に検索する方法で，ある程度なら対応が可能であるが，この場合もリストから外れるような表記のぶれを検出することはできない．旧字体と新字体の違い程度であれば，ある程度精度の高いリストを作ることができるが，それでも膨大な古文書のバリエーションすべてを網羅できるわけではない．また，翻刻作業者の古文書読解の習熟度によっては，原文に表記されている字そのものの解釈を誤ってしまい，間違った状態で翻刻文に表記してしまう可能性も考えられ，この場合はリスト方式による対応は不可能である．従って，いずれの場合でも最終的には，校正作業者が目視で校正が必要な箇所を探さなければならないのが現状である．

以上のことを解決するために，本研究では翻刻文の表記の比較校正を支援するシステムの開発を目的とする．本研究における提案システムによって，校正作業者の作業効率を高め，高品質な翻刻テキストデータの作成を支援する．

1.2 研究目的

本研究では，古文書の文書画像データの情報を翻刻文に関連付けることで，翻刻の校正作業をより効率化するシステムを提案する．

これまでの翻刻文の校正支援は，翻刻文の情報のみを頼りに，校正箇所の提案を行っていた．本提案システムにおいては，翻刻文のみならず，文書画像の情報も共に活用し，テキストと画像の二つの情報を利用した翻刻の校正支援を実現する．

古文書の画像データは，原本に忠実に情報を保存できるデータとして作成が容易であるため，デジタルアーカイブの取り組みにおいてはデジタル画像データ化が一般的であり，実際に画像データとして利用できる史料は多い．そのため，翻刻文校正のプロセスにおいて，そういった画像データを有効に活用することで，校正者の支援に役立てるのが目的である．

提案するシステムでは，文書画像検索の手法を適用し，画像中の文字の見かけに基づいた検索を行うことで，同一の語を画像中から抽出する．さらに，あらかじめ文書画像と翻刻文のテキストデータとの間に，行ごとに対応関係を付加し，該当する行から検索された語の翻刻テキストの抽出を行う．その結果として，翻刻されたテキストの表記が統一されていなくても，翻刻文の対応する部分を抽出することを可能にする．校正作業者は，複数の表記をし得る語を，文書画像中から選択するのみで抽出を行えるため，テキスト検索を用いた従来の校正作業に比べて，効率の向上および作業時間の短縮が期待できる．

2. 関連研究

本章では，本研究で提案を行う翻刻支援および対応テキスト抽出に関して先行する研究事例，あるいは本提案の中で実際に研究内容を活用している研究事例を紹介する．

2.1 文書画像と翻刻文の関係性

コンピュータ上における古文書については，早くからデジタルデータ化された古文書をどのように扱うか検討がなされており，北村 [2] は，電子化された文書画像（イメージ）と翻刻文（テキスト）の関係性について述べている．北村は，イメージを HCI (Human-Computer Interaction) とし，テキストはコンピュータ処理のための内部表現データとした上で，イメージとテキストの対応関係を結ぶことによって古文書をコンピュータ上で活用する考え方を提案している．この考え方に基づき，実際にテキストの検索結

果をイメージを通して閲覧できるシステムを実装しており，検索を行ったユーザーにはイメージによって結果が提示されるため，あたかもイメージをテキストデータであるかのように扱うことを実現している．

本研究で提案するシステムにおいても，文書画像と翻刻文を対応付けて扱う点は同様である．しかし，内部表現データであるテキストの表示を行うための HCI として文書画像を扱うだけでなく，本研究では画像検索を採用することによって，文書画像も内部表現データの一つとして利用している点が異なっている．

2.2 翻刻支援に関する研究事例

翻刻作業の支援，という分野は，どのようなアプローチから翻刻支援を行うかというか点で様々な手法が存在するため，その分野の枠組みは広く，実際に先行する研究においても様々なアプローチから翻刻を支援する研究が行われている．

古文書を対象に翻刻支援を行う代表的な取り組みとして，毛筆手書きの文書画像に対し文字切り出しと文字認識を行うことで，翻刻を支援する取り組みが行われている [3]．このような取り組みは，翻刻の中でも文字読解の支援に焦点を当てたものである．学習による古文書文字データベースの作成，既存知識を活用した不明文字の候補抽出などの手法を活用して，古文書の文字がどのような文字であるかを認識することで，翻刻者の文字読解を支援している．

一方で，翻刻作業のプロセスの一部分にシステムが介入することによって，翻刻を支援する例も多く存在する．本研究で述べる提案システムは，このアプローチによる翻刻支援に該当する．このようなシステムは読み取りの対象が古文書に限定されないため，古文書の翻刻に限らず，各分野の史料の翻刻に向けた様々な翻刻支援システムが提案されている．

複数の作業者による翻刻を支援する例として，文書画像に対するアノテーション（注釈情報）を集合知として活用する SMART-GS[4][5] が挙げられる．このシステムは複数の作業者が協働して翻刻を行うためのプラットフォームとして作用しており，難読な史料の翻刻作業を効率化することに成功している．

また，村川ら [6] は，経典の翻刻を支援するためのシステムを，翻刻文のテキスト管理にバージョン管理ソフトウェアを用いることで実現している．複数の作業者による翻刻文をバージョン管理ソフトウェアを利用して管理することで，品質の高い翻刻文の作成を支援している，さらに，複数の翻刻文同士で，文字単位の差異を比較検討することが容易になり，正しい翻刻文を採用するための翻刻支援という側面でもバージョン管理ソフトウェアが作用している．

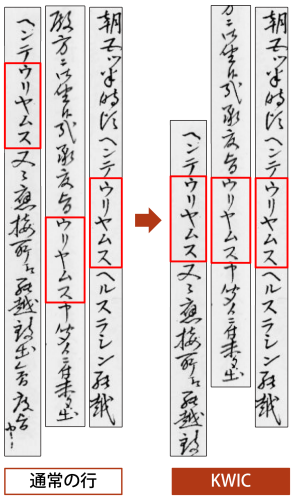

図 2 KWIC (Key Word In Context) による表示例
Fig. 2 The example of KWIC view (Key Word in Context)

2.3 表示手法

本研究で提案するシステムでは，文書内の語の検索結果を表示する手法として，KWIC (Key Word in Context) を採用した．これは対象となるキーワードを画面の中央に整列させて表示させることで，前後に書かれている文を閲覧しながら文書中のキーワードを確認するための表示手法である（**図 2**）．実際に翻刻文の校正を行う際に，語の前後の文脈から翻刻文の表記を正しいかを判断することもあり，修正対象となる語をキーワードとした KWIC を行うことで，古文書の翻刻作業においても KWIC の表示手法が有効に作用すると考えられる．本研究においては，文書画像と翻刻文テキストのそれぞれにこの表示手法を採用し，さらにお互いの表示位置を同期することによって，文書画像と翻刻文の内容の比較の際にも，翻刻者の負担を軽減させられるようにしている．

実際に先行する研究においても，古文書データの表示に KWIC を採用している例があり，北村 [2] による古文書テキストの検索においては，検索結果の KWIC の表示形式による出力を行うことが可能である．

また，KWIC の表示手法は，手書きの文書から活字体への翻刻とは異なり，日本語から英語のように，ある言語を他の言語に翻訳する作業の支援システムにも用いられている．その研究例として，対応する訳語の抽出を行い，KWIC の手法で可視化した Bilingual KWIC [7] が挙げられる．他の言語への翻訳の場合，用語を正しく表現するために，原文では同じ言葉に対しても，訳語で訳し分けを行うことがある．この研究における KWIC は，ユーザーによってそのような複数の対訳パターンを確認するための閲覧手段として採用されている．実際のシステムでは，原言語で書かれたテキストと，翻訳対象の言語で書かれたテキストの双方を文単位で対応付け，原言語で指定したキーワードが対象言語のテキストのどの部分にあたるかを抽出

し，その結果を KWIC を用いて表示している．

このような翻訳における支援の考え方は，本研究で対象としている，翻刻文の校正支援にも共通している．異なる点としては，先行研究例では原言語のテキストと翻訳されたテキストを対象に，語の抽出と表示を行っているが，本研究では元の文書の画像と翻刻されたテキストを対象にしている部分が異なる．つまり本研究では，[7] のようなテキスト同士による翻訳支援を，画像とテキストという扱いの異なるデータで適用し，翻刻支援にも活用しようとしている．

2.4 文書画像検索エンジン

本研究で提案するシステムでは，文書画像に採用する画像検索エンジンとして，寺沢ら [8] による文書画像検索システムの取り組みにおいて実装された．ワードスポッティングに基づく画像検索エンジンを採用している．具体的な手法としては，行単位に切り出された文書画像に対して，行内の部分領域に存在する画素情報から特徴量を抽出し，得られた特徴量ベクトルの系列に対して検索対象とのマッチング処理を行うことで，類似画像の検索を実現している [9]．実際のシステムの稼働実績として，インターネット上で文書画像を対象に検索を行うことのできるウェブサイトが公開 [10] されている．

この検索エンジンを使用する際は，あらかじめ文書画像データから，特徴量データを作成しておく必要がある．本研究において取り扱う画像データは，この特徴量データがすでに作成されているものとする．

3. 提案システム

本章では，提案する翻刻文校正支援システムの概要，およびその構成と実装内容について述べる．

提案システムの利用にあたっての構成は，下記の三つの段階に大別することができる．

(1) 文書画像と翻刻文の対応付けデータの作成（入力段階）
(2) 画像検索結果に基づく対応テキストの抽出（処理段階）
(3) KWIC を用いた結果の提示（出力段階）

以下，これらの各段階について詳述する．

3.1 文書画像と翻刻文の対応付けデータの作成

システムを使用するための前処理として，文書画像のデータに，校正対象となる翻刻文を行単位で対応付けを行った情報をデータとして作成し，提案システムに入力するための準備を行う．具体的には，画像データの 2 次元平面上における各行の領域に対し，対応する翻刻文データの行番号の値を振る．これにより，画像とテキストの各行の位置情報を結びつけ，各々のデータを双方向に利用することができるようになる．

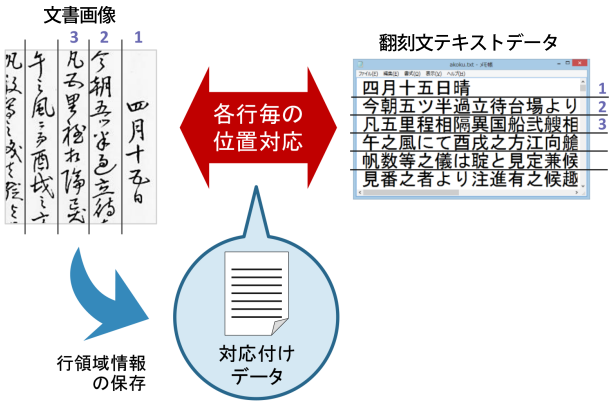

図 3 位置情報の対応付けデータ
Fig. 3 Position mapping data

本研究で扱う対応付けデータとして，文書画像とその内容のテキストデータ間で位置の対応付けを行う，トランスクリプトマッピングのためのデータ書式 [11] を採用した．トランスクリプトマッピングとは，文書から書き起こされたテキスト（本稿における翻刻文に該当）と，画像データのそれぞれの位置情報を関連づける手法のことを指す．この書式は XML 形式によって記述され，位置情報データとして，画像データにおける位置と，テキストデータにおける位置を記述することによって，双方のデータの対応する箇所を記録する．また，行領域や単語領域，文字領域のような，さまざまな構成情報単位で対応を記録することが可能であり，本研究ではこのうちの行領域を用いて記録を行う．

位置情報の対応付けデータの概略図を**図 3** に示す．実際のデータの作成にあたっては，まず文書画像データ中にある各行領域の位置情報を得て，対応付けデータに記述する．画像検索エンジンで用いられる特徴量データには，行領域の情報が含まれているため，ここではそのデータを，本提案システムにインポートして使用することができるように実装を行った．そのため，画像検索のためのデータを準備しておくことで，特別な処理を行わずとも，本提案システムで使用することのできる行領域のデータが取得できる．得られた各行領域に対し，行番号の値を振ることで，翻刻されたテキストの行と対応付けを行う．

文書の翻刻テキストデータは，画像の改行に合わせ，一行ごとに翻刻文が記述されたテキストファイルとして用意する．対応付けデータは行単位で管理されるため，必ず元文書の行に倣って改行を行う必要がある．また行の途中で，テキスト側で任意に改行を挿入したりしてもならない．テキストファイルの仕様としては一見すると厳しい条件のように思えるが，実際の古文書解読の現場においては，原典となる文書の各行に番号を振り，参照箇所は行番号で指定するなど，行ごとに区分して情報を記述することはごく一般的に行われている．

このように作成された翻刻文のテキストデータは，先ほど作成した対応付けデータにて，対応するテキストとして

明示する，その結果，画像の各行領域に設定した行番号と，翻刻文テキストの行番号が結びつき，文書画像と翻刻文の各行が対応づけられる．

ここまでが対応付けデータの作成手順となり，本提案システムではこのデータを読み込ませることで，画像と翻刻文を関連付けた状態で翻刻文の校正を行うことができるようになる．

3.2 画像検索結果に基づく対応テキストの抽出

提案システムにおける実処理の段階として，校正のために翻刻文の表現を確認したい語の範囲を画像中から選択し，その範囲の画像をクエリ（検索対象）とした画像検索処理を行う．検索結果と対応付けデータを照らし合わせることで，その行のテキストを抽出することが可能となる．

校正者は，読み込ませた古文書の画像データと翻刻文を比較しながら，校正の必要のある語を確認していく．翻刻文の中で，複数の表記がされていると思われる語を発見したら，画像中からその語が書かれている範囲を選択し，その範囲内をクエリとした類似画像検索を行う．ここでの検索には，対応付けを行ったテキストの情報は用いられず，画像から抽出された特徴量の情報のみで検索が行われる．その結果，検索対象の語の見かけに基づき，類似する部分を文書画像中から発見することができ，必然的に文書中に複数個所出てくる同一の語が検索結果の上位に現れる．

ここで，3.1 で作成した対応付けデータが使用される．画像検索の結果として，選択したクエリに近い見かけの画像が，「画像データ中のどの座標範囲に存在するか」という形で示される．この座標範囲を対応付けデータと照合することで，その検索結果が含まれている行の位置が分かる．さらに行位置の情報を利用して，対応づけられた翻刻文のテキストから該当行を抽出することで，画像検索の結果に基づいた対応翻刻テキストの抽出を実現している．

実際の利用シーンを想定した具体例として，文書中に「異國船」という言葉がある例を考える．この言葉は文書中の複数の箇所で出てくるが，画像データ上では「異國船」と一貫して表記されているの対し，翻刻作業を分担した影響で，翻刻文では「異国船」という新字体表記と「異國船」という旧字体表記が混在しているものとする．さらには箇所によっては，「黒国船」のように，見た目は似ているが全く異なる言葉で翻刻してしまっているとする．この時，テキスト検索のみで旧字体の「異國船」を検索した場合，新字体の「異国船」は発見できない．仮にそれに気づき，新字体で再検索したとしても，「黒」のような，そもそも違う文字であるものはテキスト検索で拾うことができない．

一方で，画像検索によって見かけの類似度を基準にした検索を導入した場合，文書画像中の「異國船」と書かれている箇所を一か所指定して画像検索を行うのみで，翻刻文の記述に関わらず，全ての箇所を検索できる．つまり，「異国船」のような新字体で表記されている箇所や，「黒国船」のような翻刻の誤りも含めて，一度の検索で一括して拾い上げることができる．

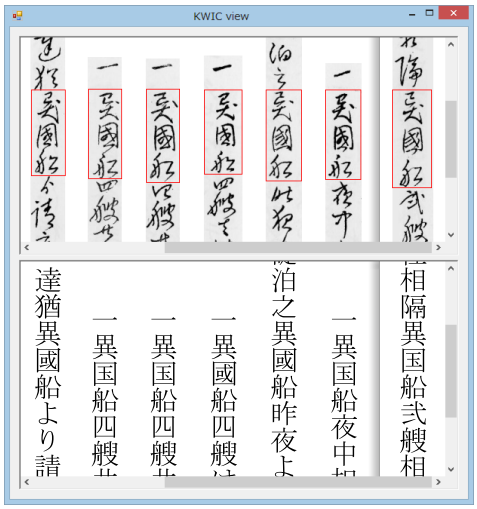

図 4 KWIC による検索結果表示
Fig. 4 Search result view with KWIC

3.3 KWIC を用いた結果の提示

処理された画像検索の結果およびテキスト抽出の結果は，KWIC の表示手法を用いて一覧表示される．この一覧表示では，校正者が選択した語の画像と翻刻文それぞれの記述を，比較閲覧できるように提示される．インターフェース上では二つのカラムで構成され，上のカラムに画像の検索結果が，下のカラムにテキストの抽出結果が，それぞれ KWIC 形式で表示される．スクロール位置は互いに同期され，対象となった語の画像上の表記と，実際の翻刻文の表記の違いを上下に並べて比較することができる．

実際の提案システムにおいて，KWIC による検索結果を表示している例が**図 4** である．上下に分かれたカラム内で，それぞれの一番右の行が検索クエリの語であり，そのすぐ左隣の行は，画像検索において類似度が一番高かった語を表す．以後，左に行くほど，クエリとの間の画像間の類似度は下がっていく．この例は，先述した「異國船」の例であり，抽出した翻刻文のテキストには，新字体による「国」の表記と，旧字体による「國」の表記が並んでいることが分かる．この結果を用いて，校正者は該当する行の文書画像や，前後の文脈を判断しながら，翻刻文の表現を校正することが可能になる．以上のような KWIC による情報の提示により，翻刻文の校正支援を実現する．

画像検索の結果では，画像データ上の位置情報が検索結果として返されるため，このデータをそのまま使用している．検索結果として出力されたの各該当語の位置を同一の位置に揃えた上で，行全体の画像を横に並べて表示していくことで KWIC を実現している．

3.3.1 テキスト位置の概算

検索結果の画像に対応するテキストの KWIC は，該当する語の位置を中央に揃えて表示するために，抽出したテキスト行の中で該当する語のある箇所を概算する処理を行っている．これは，行単位での対応付けは行われているが，行内にある各文字までは対応付けが行われていないため，概ねどの位置に該当する語があるのかを取得する必要があるためである．

例えば，仮に文書の内容が活字体の活版印刷による記述であれば，各文字の大きさは基本的に同じである．よって，画像中の行をその文字数の数だけ分割することで，各文字の位置を特定することができる．しかし，今回対象としている手書きの文書画像上においては，各文字の大きさにはばらつきがある．特に日本語の古文書の場合，つづけ字による記述で発生する各文字の形状の変化や，「くの字点」のような大きな繰り返し記号など，字の大きさが大きく変化する事例も多い．そのため，KWIC を実現するためのテキスト位置は，文字画像の各文字が同じ大きさと仮定して概算を行うが，その際に生じる概算位置のずれは，校正者の知覚で補正する必要がある．もっとも，一行の中の文字の大きさが，極端にばらばらな状態になっているとはあまり考えられない．一般的な古文書の例として，本研究で検証のために使用している古文書では，一行あたりの平均文字数は約 16 文字であった．そのため，表示のずれが発生した場合でも，そのずれの範囲は数文字程度に収まる．よって，ここで生じる表示のずれが致命的な作業効率の低下を招くことにはならないと考えられる．

具体的なテキスト位置の概算方法は，抽出対象の存在する行内の相対位置の情報を用いている．まず文書画像の行を，対応する翻刻文の行に存在する文字数分で均等に領域を分割する．そして，抽出の対象となる領域が，分割された領域内に含まれている箇所を得ることで，分割された領域のインデックスに該当するテキスト部分の位置，概算された位置とみなす（**図 5**）．

以上の実装を行うことで，文書画像と翻刻文を利用した翻刻文校正支援システムを実現することができる．

3.4 補助機能

本研究で提案する翻刻文校正支援システムは，3.1〜3.3 で述べた内容が基本機能である．これらの機能に加えて，より校正作業を効率化するためのオプションとして，以下のような補助機能を付加している．

3.4.1 頻出文字の強調表示

画像検索において得られる情報を利用し，検索結果から抽出された翻刻文中のテキスト内で，どの文字が結果によく出現するか（頻出文字）を抽出する．得られた頻出文字は，KWIC の結果表示の中で強調して表示される．

図 5 相対位置を用いたテキスト位置の概算
Fig. 5 Text position estimation by using relative position in line

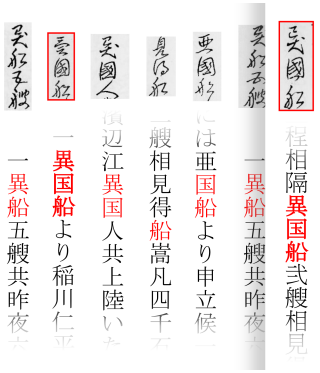

図 6 頻出文字の強調表示
Fig. 6 Highlighting frequently-appearing letters

画像検索において得られる情報を利用し，検索結果から抽出された翻刻文中のテキスト内で，どの文字が結果によく出現するか（頻出文字）を抽出する．得られた頻出文字は，KWIC の結果表示の中で強調して表示される（**図 6**）．

画像検索の結果は，各特徴量の類似度から算出された画像間距離の順に並べられるため，その結果としてクエリと同一の語が上位に並び，下位にはクエリと異なる語が並ぶ．一方で，検索対象である古文書は，ほぼ手書きで書かれているため，仮に同一の語でも字が崩されて書かれている場合などがある．その結果，画像間距離が，上位に並んだクエリと同一の語より離れている場合，クエリと異なる語の結果の中に本来クエリと同一である語の結果が紛れ込むこともある．この場合は，KWIC 表示でも意識して見ていなければ，そのような語の発見が難しい．そこで，検索結果の中で最も頻出している字を，KWIC の中で強調表示することで，クエリと異なる語の中に紛れた検索対象語を見つけやすくするのがこの機能の狙いである．

頻出文字を抽出するにあたっては，まず文字の画像検索

の各結果から，3.3.1 と同様にテキスト位置の概算を行う．ただしこのままでは，テキスト位置が正しい位置から数文字ずれることがある．そのため，本来抽出すべき文字（検索クエリとして選択した語の文字）が概算したテキストの中に入らない可能性がある．そこで，概算テキスト位置の前後の文字を広げ，抽出テキストの範囲を広く取るようにする．広げる範囲はシステムで変更できるが，デフォルトでは 2 文字に設定している．

この処理で得られた全てのテキストに対して，1-gram による，各文字の出現頻度を得る．ここで画像検索の情報を利用し，各文字の出現のカウント時に，検索された結果の対象画像と，検索クエリとの画像間距離の値から，各文字に重みを与える．

校正者がシステムで得た検索結果を件数を N とする．実装システムでは，10 件単位で検索が可能であり，検索結果のページネーションを順次進むに従って，N は増加していく．このうち，n 件目 $(1 \leq n \leq N)$ の検索結果のクエリとの画像間距離を $d_n (d_n \geq 0)$ とする．d_n が小さいほど，検索結果とクエリの間の類似度が高いことを示す．また，$d_n \leq d_{n+1}$ である．このとき，n 件目の検索結果から抽出したテキストの各文字に対して加算される重みの値 w_n を，

$$w_n = d_N - d_n + C \tag{1}$$

で与える．画像検索において見かけの類似度が高かった検索結果から抽出された文字を重要と判断し，重みを大きくしている，検索件数が増加すると，上位の重みはさらに大きくなる．C は結果の画像間距離に依存しない重みを付加する定数で，現在は 1 としている．C の加算回数は，文字の出現回数に依存するため，C の値を大きくした場合は，文字の出現回数が重要視され，C の値を小さくした場合は，検索結果のランク（上位への出現）が重要視される．

このような重み計算を N 件全てに対して行うことで，各文字に重みが与えられる．そして，各文字ごとに与えられた重みを合計することで，最終的な合計重みデータを得ることができる．このデータを降順にソートしたのち，判別分析法を用いて閾値を算出する．その結果，閾値より上位のグループに該当した文字を，頻出文字として判定する．

ここで得られた頻出文字は，KWIC による表示を行う際に，該当文字のみ色を変えることで強調表示を行う．また，各行で抽出されたテキストに頻出文字が全て含まれている場合は，さらに太字で強調表示される．これによって，選択した語に対して一番良く翻刻されている文字を，一覧表示内で確認することが可能になる．

3.4.2 テキストからの画像位置概算

前節で述べた，画像位置からのテキスト位置の概算の逆の処理を行う機能であり，選択したテキストから該当する画像位置の概算を行う．

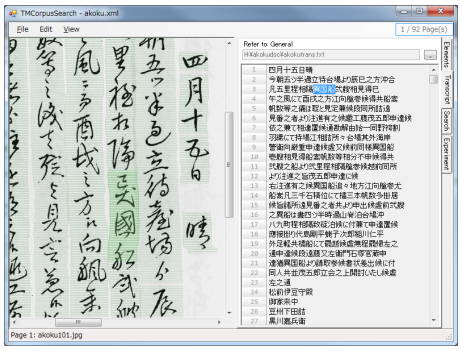

図 7 システムプロトタイプ
Fig. 7 A prototype of the proposed system

本提案システムで用いる対応付けデータは，画像からテキストの抽出を行うのみならず，双方向からの利用を想定しているため，テキストの行情報が指定されている画像の行を取得することももちろん可能である．この機能によって，従来の翻刻文校正作業のように，テキスト検索を利用して校正候補の語を見つけた場合に，検索したテキストの画像上の語の位置を確認し，校正の要否を決定することができるようになる．

4. 実装

本章では，提案する翻刻文校正支援システムの実装とその動作の検証について述べる．

4.1 システムプロトタイプの実装

本提案システムの実際の動作を確認および検証するため，実際の稼働システムに先立ち，システムプロトタイプの実装を行った（**図 7**）．開発環境には C#を使用している．また，プロトタイプにおける動作検証のための文書画像及び翻刻文として，『亜国来使記』を使用している．縦書き毛筆の草書体で書かれたこの文書は，ペリー艦隊が 1854 年に函館を来訪した際に手記で記録された応接記録であり，実際に文書画像検索システムで検索が可能な文献としてオンラインで公開されている [10].

4.2 実装システムの検証

現在のシステムプロトタイプの実装は，本稿で述べた基本機能と補助機能のそれぞれに基づいた実装が行われており，システムの実装と検証を通じて見つかった問題点などを検証する．

4.2.1 基本機能

まず，入力段階においては，採用した検索エンジンで用いられている特徴量データの読み込みをサポートしたことにより，本提案システムで使用するための行領域情報の読み込みを容易に行うことができた．一方で，システムの入力データとして採用した XML 形式による対応付けデータ

表 1 単語「異國船」検索時の頻出文字の抽出結果
Table 1 Result of extracting frequently-appearing letters in the case of searching "Ikokusen"

文字	重み総計	出現回数
船	758.56	33
異	735.22	40
艘	431.66	21
一	397.89	19
国	391.06	20
日	344.77	17
國	342.41	15
晴	336.08	13

(上位 8 件，出現回数は抽出テキスト内に限る．太線は閾値)

の記述量が多く，特に画像検索の際の特徴量データを含めた場合，831 行（画像データ数 92 枚）の文書を対象にしたデータとして，ファイルサイズが約 231MB の対応付けデータが出力された．そのデータ量の多さから，システムの利用中こそ障害は見受けられなかったが，システム起動後の初回のデータ読み込みにおいて 5〜6 秒待機する必要がある．このことから，システムの実用化に向けて，データの入力方式について再考する必要があると思われる．

また，検索結果の表示部において，類似画像検索や，KWIC の表示をはじめとする基本実装部分は問題なく動作している．一方で，プロトタイプでは行単位で内部処理や表示を行っているため，二つの行をまたいだ語の検索結果が存在している場合，テキスト位置の抽出や，KWIC の表示が正しく動作しなかった．この部分については，全ての行を一次元的に処理するなど，内部的な処理の見直しで改善が可能と思われる．

4.2.2 補助機能

本研究で提案した補助機能のうち，頻出文字の強調表示について実験を行い，基本機能で提供される翻刻文校正作業をサポートできるかどうかを検証した．

検証文献を利用して，翻刻されたテキストに新字と旧字の異なる二種類の表記がなされている場合を想定し，そのような中で正しく頻出文字が抽出され，文字の強調表示が効果を示しているか実験を行った．

検索件数 N を 50 件，概算テキストの拡大幅は前後 2 文字とし，検索対象となる文字は「異國船」とした．この語は文書画像中の 30 箇所にある．各箇所に該当する翻刻文は，表記の違いが生じていることを想定し，「異國船」30 箇所のうち，半分の 15 箇所の翻刻文の表記を「異国船」と新字体による表記に変更した．検索クエリは，文書中にある 30 箇所のうち，初回に現れる部分の画像とした．この結果得られた頻出文字の重みの算出結果を**表 1** に示す．

この結果，頻出文字と判定されたのは「異」と「船」の二種類であり，「国」「國」はそれぞれ頻出語とみなされなかった．それらの文字よりも出現回数の多い「艘」「一」は，対象となる語の前後に共起する割合が高い文字であったが，判別分析法による閾値の算出により，頻出語とはみなされなかった．ここで得られた頻出文字を KWIC 表示画面で強調表示させると，二つの文字がそれぞれ強調され，さらに双方が入っている語では，間にある非強調の「国」「國」の文字を発見しやすくなり，翻刻の表記の違いの確認が容易に行えるようになった．

5. おわりに

本研究では，古文書の翻刻文校正のため，文書画像データを用いた画像検索を用いることで，校正作業の効率化するシステムを提案し，プロトタイピングを通じてその基本的な機能の実装を検証した．

今後は，プロトタイプとして作成したシステムを，実際の運用に耐えうるように機能とインターフェースをより検証し，提案システムとして完成度を高めていく必要がある．また，実際の翻刻文校正作業を通じた評価実験を行い，フィードバックを得ることで機能面の改善を行い，最終的な翻刻文校正支援システムの提案につなげていきたいと考えている．

参考文献

[1] 国立公文書館：国立公文書館 デジタルアーカイブ（オンライン）, 入手先 ⟨http://www.digital.archives.go.jp/⟩, (2013.07.17).
[2] 北村啓子：古文書を分析するための HCI，情報処理学会研究報告，ヒューマンインタフェース研究会報告，Vol.94, No.23，pp.63–70 (1994).
[3] 山田奨治，柴山守：古文書を対象とした文字認識の研究，情報処理，Vol.43，No.9，pp.950–955 (2002).
[4] 相原健朗，林晋：画像化主義に基づく文献資料研究用ツール SMART-GS とその発展，情報処理学会研究報告，デジタルドキュメント，Vol.2011-DD-79，No.5 (2011).
[5] 大浦真，林晋，久木田水生：SMART-GS: 文献研究のためのソフトウェアツール，言語処理学会 第 19 回年次大会 発表論文集，pp.382–385 (2013).
[6] 村川猛彦，福岡整，丁敏，中川優，三宅徹誠，落合俊典：バージョン管理を用いた漢訳仏典翻刻支援システム，情報処理学会論文誌 データベース（TOD），Vol.4，No.2，pp.101–113 (2011).
[7] 小川泰弘，西森寛敏，外山勝彦：Bilingual KWIC – 対訳抽出の可視化による翻訳支援，言語処理学会 第 11 回年次大会 講演論文集，pp.811–814 (2005).
[8] 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).
[9] Terasawa, K. and Tanaka, Y.: Slit Style HOG Features for Document Image Word Spotting, Proc. 10th International Conference on Document Analysis and Recognition, vol.1, pp.116–120 (2009).
[10] 文書画像検索システム（オンライン），入手先 ⟨http://records.c.fun.ac.jp/⟩, (2013.07.30).
[11] 服部雄輝，寺沢憲吾：トランスクリプトマッピングのための書式とその活用，情報処理学会研究報告，vol.2012-DD-84, No.3 (2012).