類似文字検索機能をそなえた電子くずし字辞典の開発

山田　奨治 *1, 和泉　勇治 *2, 加藤　寧 *2, 柴山　守 *3

*1 国際日本文化研究センター・研究部
*2 東北大学大学院・情報科学研究科
*3 大阪市立大学・学術情報総合センター

古文書の翻刻作業の効率をたかめるためには，標準的なくずし字辞典を電子化し，検索の利便性を向上させることが有効であろう．また，デジタル化された文字画像を使って，ある文字に類似した文字を一覧的に検索することが可能になる．類似文字検索を実現するさいに鍵となるのは，文字の特徴量と文字間の類似度の設定方法である．われわれは，オフライン日本語手書き文字認識技術で使用されている文字特徴量と文字間類似度にストローク情報を加味することで，類似文字検索機能をもった電子くずし字辞典を開発した．

Development of a digital dictionary of historical characters with search function of similar characters

YAMADA Shoji*1, WAIZUMI Yuji*2,
KATO Nei*2, and SHIBAYAMA Mamoru*3

*1 International Research Center for Japanese Studies
*2 Graduate School of Information Sciences, Tohoku University
*3 Media Center, Osaka City University

Developing an computerized dictionary of historical characters would be effective to improve the reading speed of historical documents. Using the digitized character images, we can also browse characters similar to an example. A key issue for implementing the search function of the similar characters is how to define the character feature and the similarity between two characters. We developed a computerized historical character dictionary by using some character features and similarities, which are used in Japanese off-line hand-written OCR technology, and using stroke information in addition to them.

1　はじめに

古文書の翻刻（文書を解読して記述内容を活字にすること）は，歴史研究を前進させるために必要不可欠な基礎的な作業である．国内には翻刻されていない古文書が膨大な数あり，それらの多くは未だ手つかずのまま文書館などに眠っている．古文書の翻刻作業は人手に頼らざるをえないにもかかわらず，翻刻しなければならない古文書数に比して翻刻作業にたずさわっている人間の数は非常に少ないのが現状である．また，古文書の翻刻作業は高度に専門的なもので，一人前の作業者になるまでに長期にわたる訓練を経なければならず，作業者の育成と作業効率の向上が，歴史研究を進めるうえでの困難な問題になっている．

われわれは，古文書の翻刻作業をトータルに支援するシステムの開発をめざした「古文書翻刻支援システム開発プロジェクト（HCR プロジェクト)」を進めている [1][2][3]．HCR プロジェクトの主な目標は，(1) 古文書文字認識システムの研究 [4]，(2) 古文書文字認識システム研究のための古文書文字データベースの作成，(3) 古文書からの文字切り出しの研究 [5][6]，(4) 古文書に関する知識を用いた翻刻支援の研究 [7]，(5) 電子化古文書文字辞典の研究においている．本報告は，これらのうちの (5) 電子化古文書文字辞典の研究の中間的な成果に関するものである．

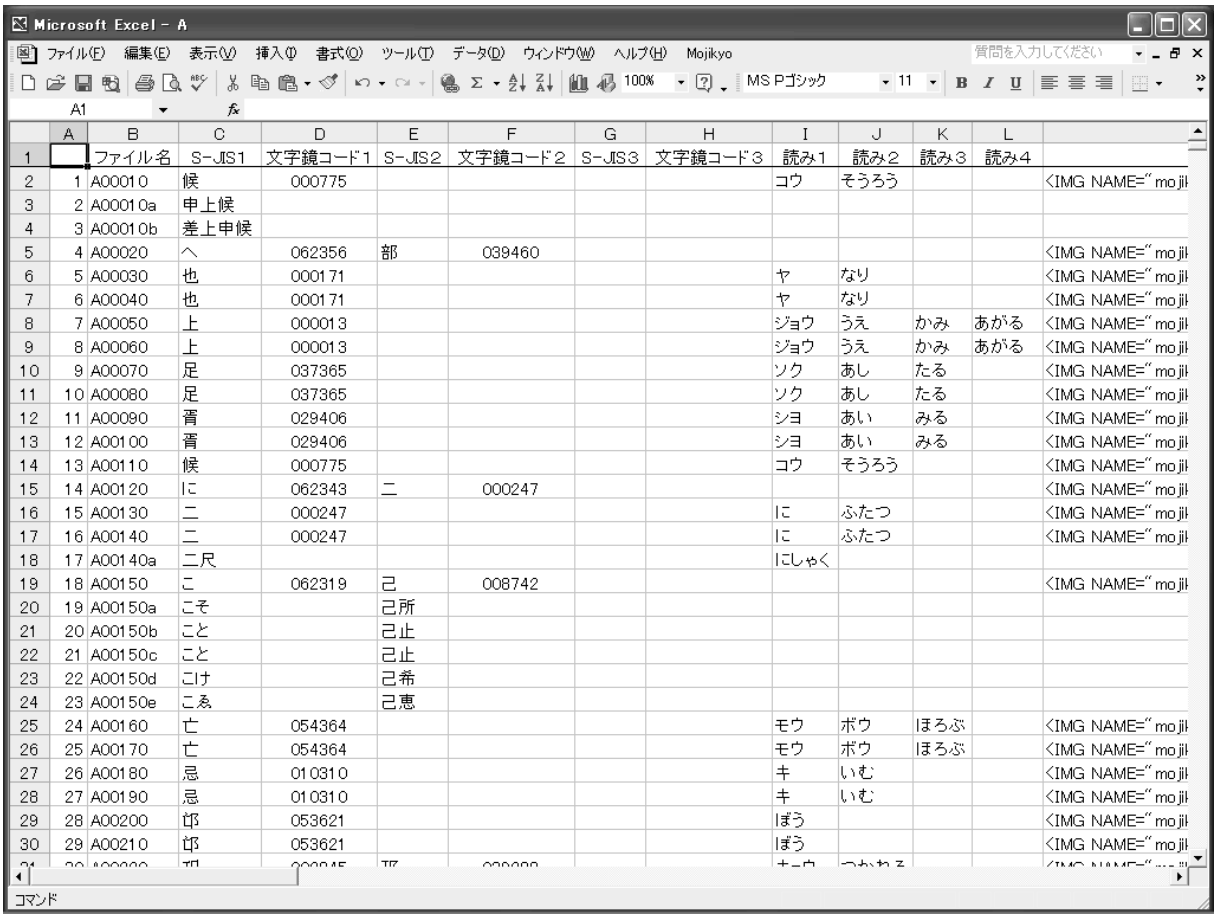
現在のところ，古文書の文字辞典類のなかで電子化されたものはない．古文書の翻刻の際に使用される標準的な辞書を電子化し，検索の利便性をたかめることができたならば，翻刻作業の大幅な向上が見込まれる．辞書の電子化を考えるならば，現在もっともよく使われている辞書を対象にするすることが理想である．専門的な翻刻者がよく使用している辞書のひとつに，東京堂出版『毛筆版くずし字解読辞典』[8]（以後『くずし字辞典』）がある．この辞書は，文字の第 1 ストロークの方向を 5 種類（縦点，横点，斜棒，縦棒，横棒）に分類して検索キーにするという，ほかの辞書にはない特徴を備えている．つまり，不明な文字を調べる際に，第 1 画の方向から探索することができるのである．しかしながら，この辞書を実際に使ってみると，求める文字にたどりつくにはそれなりの時間がかかり，検索漏れがおこる可能性もたかいことがわかる．辞書を電子化して検索の方法を工夫すれば，知りたい文字にたどりつくまでの時間を短縮し，検索漏れをすくなくすることができるだろう．

また，辞書を電子化することによって，紙の辞書では到底できない検索方法を実現することができる．それは，ある文字に類似した文字を一覧的に検索することである．類似文字の検索を実現する際に鍵となるのは，文字の特徴量と文字間の類似度の設定方法である．くずし字の特徴量と類似度は，オフラインの日本語手書き文字認識技術で使用されている手法を応用することによって求めることができる．

以上のようなアイデアのもとに,『くずし字辞典』を電子化し，類似文字検索機能を開発して，電子くずし字辞典を実装した．

2　辞書の電子化

電子化の対象にしたのは,『くずし字辞典』のなかの付録以後を除く章に掲載されている 23,703 文字である．ここには漢字・かな文字のほかに，かなの複合文字「より」のように複数の文字からなる例も 1 文字として含まれている．



	A	B	C	D	E	F	G	H	I	J	K	L	
1		ファイル名	S-JIS1	文字鏡コード1	S-JIS2	文字鏡コード2	S-JIS3	文字鏡コード3	読み1	読み2	読み3	読み4	
2	1	A00010	候	000775					コウ	そうろう			<IMG NAME="moji
3	2	A00010a	申上候										
4	3	A00010b	差上申候										
5	4	A00020	へ	062356	部	039460							<IMG NAME="moji
6	5	A00030	也	000171					ヤ	なり			<IMG NAME="moji
7	6	A00040	也	000171					ヤ	なり			<IMG NAME="moji
8	7	A00050	上	000013					ジョウ	うえ	かみ	あがる	<IMG NAME="moji
9	8	A00060	上	000013					ジョウ	うえ	かみ	あがる	<IMG NAME="moji
10	9	A00070	足	037365					ソク	あし	たる		<IMG NAME="moji
11	10	A00080	足	037365					ソク	あし	たる		<IMG NAME="moji
12	11	A00090	胥	029406					シヨ	あい	みる		<IMG NAME="moji
13	12	A00100	胥	029406					シヨ	あい	みる		<IMG NAME="moji
14	13	A00110	候	000775					コウ	そうろう			<IMG NAME="moji
15	14	A00120	に	062343	二	000247							<IMG NAME="moji
16	15	A00130	二	000247					に	ふたつ			<IMG NAME="moji
17	16	A00140	二	000247					に	ふたつ			<IMG NAME="moji
18	17	A00140a	二尺						にしゃく				
19	18	A00150	こ	062319	己	008742							<IMG NAME="moji
20	19	A00150a	こそ		己所								
21	20	A00150b	こと		己止								
22	21	A00150c	こと		己止								
23	22	A00150d	こけ		己希								
24	23	A00150e	こゑ		己恵								
25	24	A00160	亡	054364					モウ	ボウ	ほろぶ		<IMG NAME="moji
26	25	A00170	亡	054364					モウ	ボウ	ほろぶ		<IMG NAME="moji
27	26	A00180	忌	010310					キ	いむ			<IMG NAME="moji
28	27	A00190	忌	010310					キ	いむ			<IMG NAME="moji
29	28	A00200	邙	053621					ぼう				<IMG NAME="moji
30	29	A00210	邙	053621					ぼう				<IMG NAME="moji

図 1: 電子辞書の文字情報

電子化の手順は，つぎの通りである．まず，毛筆で書写された文字の画像を 400dpi の 2 値画像でスキャナ取り込みした．同時に，文字画像に対応する文字のフォントを Windows 内蔵のフォントで割り当て，内蔵フォントにない文字については今昔文字鏡フォントを使用した．また，複数文字からなる例を除くすべての文字について，今昔文字鏡の文字コードを付与し，読みの情報を最大で 9 種類付与した．作成した文字情報の一部を図 1 に示した．

3　類似文字の検索手法

文字の特徴量の算出方法として，われわれは改良型方向線素特徴量 [9] を採用した．この特徴量は，日本語手書き文字認識の研究用データベースとして定評のある ETL9B を対象にした実験で，たかい認識性能を示している．

改良型方向線素特徴量の算出方法の概要を以下に示す．はじめに前処理として，スムージング，輪郭線抽出，正規化をおこなう．スムージングは，文字の局所形状の変化をなめらかにしてノイズを軽減するためのものである．文字の形は，輪郭線抽出によって取り出す．輪郭線抽出ではなく細線化をおこなうと，文字がつぶれていた場合に文字の形状情報が失われてしまう．その点で，細線化よりも輪郭線抽出のほうが，毛筆の特徴抽出においても優れている．正規化は，津雲による非線形正規化 [10] を採用している．津雲の正規化法は，ストローク

間隔の逆数を正規化関数にするもので，ストローク間の間隔をある程度均一化できるという特徴がある．

文字特徴量は，以下の手順で算出する．

1. 輪郭線の線素化
2. 方向線素特徴量の算出
3. 外側加重による方向線素特徴量の補正
4. 方向線素ベクトルの算出

輪郭線の線素化は，輪郭線上の黒画素を方向づける作業である．輪郭線に対して3×3 のマスクを用いて，線素の方向を縦，横，+45 度，-45 度のいずれかに分類する．ただし図2のような場合は，たとえば(a)では縦と+45 度の 2 方向に線素があると判断する．

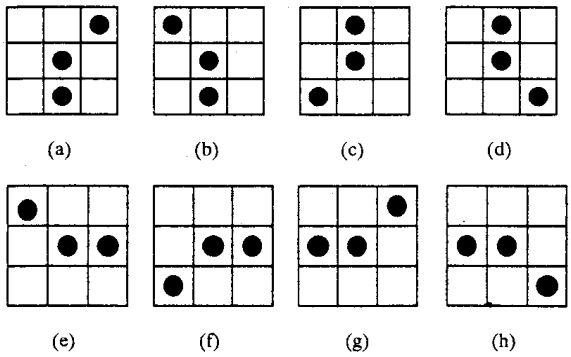


図 2: 二つの方向をもつ線素（文献 [9] より引用）

方向線素特徴量の算出方法は，図 3 に示した．64×64 ドットからなる文字画像領域を 8×8 ドット単位に分割する．隣接する 4 単位をまとめて 16×16 ドットの小領域とし，

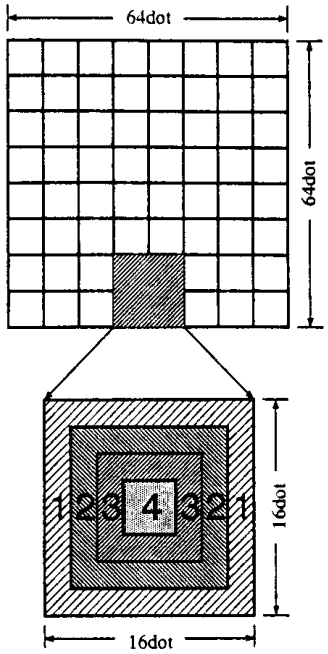


図 3: 方向線素特徴量に対する外側加重（文献 [9] より引用）

縦と横の両方向に，それぞれの半分づつをオーバーラップさせてとっていく．小領域は全部で 7×7 の 49 個えられる．

外側加重による方向線素特徴量の補正は，図 4 に示した．文字画像領域の外側に 16×16 個ドットからなる 32 個の仮想小領域を設けて，他の小領域と同様に方向線素特徴量を求め，それぞれ対応する周辺部の小領域の特徴量に加算する．ただし，文字画像領域の 4 隅の小領域には，4 隅を中心とする 16×16 ドットの仮想小領域と，それに半分づつ重複する隣接の仮想小領域の特徴量を加算する．このような外側加重による方向線素特徴量の補正によって，比較的つぶれのすくない文字周辺部の特徴をより有効に利用することができる．

方向線素ベクトルの算出は，つぎの手順でおこなう．各小領域を図 3 の下図に示すような 4 つの部分に分割し，部分領域にそれぞれ重み 4,3,2,1 を対応させる．各小領域の方向線素特徴量を，つぎの 4 次元のベク

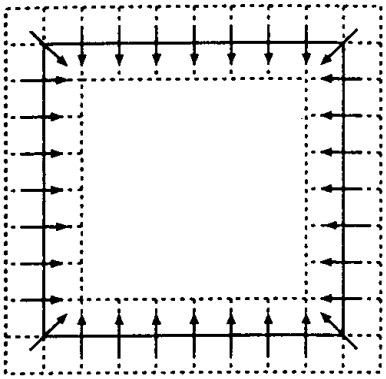


図 4: 方向線素特徴量に対する外側加重（文献 [9] より引用）

トル

$$(x_1, x_2, x_3, x_4)$$

で定義する．ただし，

$$x_i = 4x_{1i} + 3x_{2i} + 2x_{3i} + x_{4i} (i = 1 \sim 4)$$

である．ここで添字 i はそれぞれ，縦，横，+45 度，-45 度の方向線素を意味する．$x_{1i}, x_{2i}, x_{3i}, x_{4i}$ はそれぞれ中心から外側に向けて 4 つの各部分領域での方向線素 i の個数をあらわす．

したがって 1 文字の方向線素特徴量は，49 個の小領域の方向線素特徴量をならべたもので，次元数は 196 となる．各文字間の類似度は，196 次元の方向線素ベクトルのユークリッド距離を使って求めた．

『くずし字辞典』の特徴は，第 1 ストロークの方向によって文字を分類している点にある．すなわち第 1 画を，①下に向かって連続する点で起筆する「縦点」，②右に向かって連続する点で起筆する「横点」，③右上から左下へ斜めに伸ばす棒で起筆する「斜棒」，④上から下へ伸ばす棒で起筆する「縦棒」，⑤左から右へ伸ばす棒で起筆する「横棒」の 5 種類に分けて，その種類ごとに文字が掲載されている．

文字の類似度は，第 1 ストロークがおなじ文字間についてのみ計算した．そうすることによって，第 1 ストローク情報による候補の絞り込みができ，すべての文字間を対象にするよりも検索精度の向上が見込まれるからである．いわば，オフライン文字認識手法に，第 1 ストロークというオンライン的な要素を取り入れた手法といえるだろう．

4 電子古文書文字辞典の実装

以上の作業を経て，電子くずし字辞典 (eKuzushi) を実装した．この電子辞典は，『くずし字辞典』から採録した 23,703 文字の画像とテキスト情報，そしておなじ第 1 ストロークからはじまるすべての文字間の類似度情報を持っている．この電子くずし字辞典は，Windows 環境で稼働し，実装には Microsoft 社の Visual Basic 5.0 を使用した．

現在のところ，検索の入口は文字コードのみとなっている．調べたい文字を ATOK あるいは IME などの日本語入力 FEP を使って入力し，検索ボタンを押すと（図 5），文字コードに該当するすべての文字画像が一覧で表示される（図 6）．文字画像にカーソルを合わせると，その読みが小さなウィンドウに表示される．一度に表示される候補文字数は，最大で 5 文字である．スライドバーを操作すると，右のほうに隠れた他の候補文字をみることができる．

文字画像をダブルクリックすると，類似文字を検索することができる．たとえば図 6 の右から 2 つめの「預」と似た字を調べたいときは，その画像をダブルクリックすると別のウィンドウが開き，そこに類似文字が表示される（図 7）．類似文字の一覧は，検索対象文字とおなじ第 1 ストロークをも

つ文字のなかで，196 次元方向線素ベクトルのユークリッド距離が小さい順に最大で第 10 候補まで表示される．類似文字検索はいずれのウィンドウからも可能で，前に開いたウィンドウに戻って類似文字の検索をすることもできる．

図 5: eKuzushi 検索画面 1（「預」を入力して検索）

図 6: eKuzushi 検索画面 2（「預」の検索結果）

5　利用試験

このようにして実装された電子くずし字辞典 eKuzushi について，簡易な利用試験を

図 7: eKuzushi 検索画面 3（画面 2 の右から 2 つめの文字に類似した文字の検索結果）

おこなって，その有効性を検証した．利用試験は，つぎの方法をとった．

1. 古文書文字データベース HCD2[1] に収録されている証文類の標題行から，翻刻文が同一でない 10 種類を無作為に選択して試験文字列とする

2. 10 種類の試験文字列を 5 種類づつ 2 群に分けて，一方を冊子辞書条件，一方を電子辞書条件とする

3. それぞれの群について辞書を使って 1 文字あたり 1 分（3 文字ならば全体で 3 分）の制限時間内に翻刻する

被験者は 3 名で，古文書読解の能力は初級から中級程度である．

表 1･2 に，利用試験結果を示した．少数の被験者と文字数による試験ではあるが，つぎのような傾向を観察することができた．

- 冊子辞書よりも電子辞書を使用したほうが，1 回の辞書引きに要する所要時間が少ない

- 電子辞書の使用によって，翻刻まちがいによる誤謬を減らす効果はみられなかった

表 1: 利用試験結果（冊子辞書条件：総文字数=38）

被験者	辞書引 1 回あたり所要時間（分）*	不明文字数	誤謬文字数
A	1.6	1	4
B	1.8	0	4
C	1.9	0	4
平均	1.8		

* 制限時間切れの場合を除外した

表 2: 利用試験結果（電子辞書条件：総文字数=35）

被験者	辞書引 1 回あたり所要時間（分）*	不明文字数	誤謬文字数
A	1.0	1	6
B	2.0	1	3
C	0.7	1	9
平均	1.2		

* 制限時間切れの場合を除外した

- 電子辞書の使用によって，不明のまま残る文字数を減らす効果はみられなかった

これらの傾向のうち第 2 点目は，被験者が正しいと判断した文字がまちがっていたことを意味しており，辞書の性能よりもむしろ翻刻者の能力に起因する部分であると考えられる．今後，古文書で頻出する用例を示す機能を電子辞書に追加することで，翻刻者を支援し，翻刻まちがいを減らすことができるだろう．第 3 点目については，冊子辞書が筆順をキーに文字を検索できる構造を持っているのに対して，現状の eKuzushi は文字コードからしか検索できないという機能的な制限に起因する部分があると考えられる．今後，冊子辞書が備えているような，筆順からの検索を可能にするような機能追加が必要であろう．

6 おわりに

以上のように，われわれは日本語手書き文字認識技術で使用されている文字特徴量にストローク情報を加味した類似文字検索機能をもった，電子くずし字辞典を開発した．この電子辞典は，現在のところワープロ的に入力された文字が最初の入口となっているため，わからない文字が何であるかのおおよその検討を利用者がつけなくてはならない．将来的には，①タブレットなどで手書き入力された文字からの検索，②スキャナなどで画像入力された文字からの検索，③より深い階層のストローク情報からの検索機能をもたせるべく，研究を進めているところである．

謝辞

本研究は，日本学術振興会科学研究費補助金・基盤研究 (B)(1) 一般研究「古文書解

読プロセスの知能情報学的解明」（平成11〜13年度，研究代表者：山田奨治），同「古文書OCRの試論的研究」（平成11〜13年度，研究代表者：柴山守），同展開研究「手書き文字ＯＣＲ技術を援用した古文書翻刻支援システムの開発」（平成11〜13年度，研究代表者：山田奨治），同「古文書解読支援システムの開発と電子辞書技術の応用に関する研究」（平成12年〜14年度，研究代表者：柴山守）の支援を得て実施しているものである．本研究のために辞書の電子化を許諾いただいた（株）東京堂出版に感謝申し上げる．

参考文献

[1] HCR プロジェクトのホームページ http://www.nichibun.ac.jp/~shoji/hcr/

[2] 山田奨治, 加藤寧, 川口洋, 原正一郎, 石谷康人, 柴山守, 笠谷和比古, 小島正美, 梅田三千雄, 山本和彦: 古文書翻刻支援システム開発プロジェクト報告（１）ープロジェクト概要ー, 情報処理学会研究報告, Vol.2000, No.8, pp.1-8, 2000.

[3] 山田奨治, 柴山守: 平成11〜13年度日本学術振興会科学研究費補助金研究成果（中間）報告書　古文書翻刻支援システムの研究 (1), 2000.

[4] 和泉勇治, 加藤寧, 根元義章, 山田奨治, 柴山守, 川口洋: ニューラルネットワークを用いた古文書個別文字認識に関する一検討, 情報処理学会研究報告, Vol.2000, No.8, pp.9-15, 2000.

[5] 尾崎浩司, 柴山守, 荒木義彦: 古文書画像のレイアウト認識と標題抽出, 情報処理学会研究報告, Vol.2000, No.67, pp.47-54, 2000.

[6] 原正一郎: 古典 OCR のための文字切り出しについて, 情報処理学会研究報告, Vol.2000, No.67, pp.55-64, 2000.

[7] 山田奨治, 柴山守: n-gram による古文書証文類翻刻支援の検討, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, Vol.2000, No.17, pp.185-192, 2000.

[8] 児玉幸多編: 毛筆版くずし字解読辞典, 東京堂出版, 1999.

[9] 孫寧, 安部正人, 根元義章: 改良型方向線素特徴量および部分空間法を用いた高精度な手書き文字認識システム, 電子情報通信学会論文誌, Vol.J78-D-II, No.6, pp.922-930, 1995.

[10] 津雲淳: 手書き漢字認識における非線形正規化処理, 昭和62年度電子情報通信学会情報・システム部門全国大会, p.68, 1987.