古代文字検索のためのフォントからの字形特徴量の抽出および活用可能性の検討

李　康穎† Batjargal Biligsaikhan‡ 前田　亮†‡
†立命館大学情報理工学研究科　〒525-8577 滋賀県草津市野路東 1-1-1
‡立命館大学衣笠総合研究機構　〒603-8577 京都市北区等持院北町 56-1
†‡立命館大学情報理工学部　〒525-8577 滋賀県草津市野路東 1-1-1
E-mail: †gr0319ss@ed.ritsumei.ac.jp ‡biligsaikhan@gmail.com †‡amaeda@is.ritsumei.ac.jp

あらまし　近年，オフライン文字認識の手法として，大量の学習データに基づき，深層学習により画像の抽象化を行う手法が提案されており，手書き文字や印刷文字を認識するタスクにおいて，すでに高い精度が得られている．しかし，現在使用されている文字の形と大きな違いがある古代文字では，同じ文字に対応する字形のバリエーションが多く存在するため，全てのバリエーションを同じ種類として学習データを作成するのは問題がある．本研究では，古典籍ディジタルアーカイブに多く含まれる蔵書印に記されている個々の単一古代文字の検索ベース認識手法として，古代文字フォントの字形から複数特徴量を抽出して認識を行う手法を提案し，検索結果から，複数の人文系データベースとの連携検索の活用可能性を検討する．

キーワード　古典籍ディジタルアーカイブ，情報検索，古代文字認識

1. まえがき

現在，漢字を使用している国の多くで使用されている書体は楷書と呼ばれ，その字形は古代に遡ると篆書や隷書などの書体が存在している．専門家ではない現代人にとって，楷書と異なる篆書体を読むのは困難である．また，時代の変化によって，漢字の形が変化したり，複数のバリエーションの書体が生まれることがある．また，漢字から派生した文字や，表記する言語が漢字と異なるもの，一見漢字にみえる文字も存在する．漢字を使う国の中で，例えばベトナムの「チューノム」は漢字本来の字形を利用した文字体系であり，日本の「平仮名」や，中国の一部の地方で女性のみが使う「女書」は，漢字の変形，崩れなどの扱いを通じて生まれた文字と言われている．使用されることが稀な文字データを大量に集め，それぞれの専門家に文字の種類を判別してもらい，ラベルをつけることは困難である．このため，ラベル付き字形画像が一枚だけであってもユーザのクエリ文字と照合でき，随時文字の種類を更新できる検索ベース古代文字認識システムの構築が期待される．

本研究では，一枚のラベル付き字形画像から複数の特徴量を抽出し，検索により古代文字を認識する手法を提案する．既存の多くの篆書体フォントは美観の追求に主眼に置かれており，「造字」が頻繁に行なわれているため，これらと歴史上の篆文字形には極めて大きな相違が存在する．そのため，使用するフォントに対して選別を行い，古代の著作を参照し作成したフォントを実験データに用いる．立命館大学白川静記念東洋文字文化研究所が公開している『白川フォント』[1]には篆文 2,590 字のフォントデータが含まれている．これらのデータは漢字学者の研究に由来し，厳密な校正を経ており，「造字」で作成された「篆書体」に見られる現代文字の字形は一切含まれず，信頼できる古代文字字形データと考えられる．一方，国文学研究資料館の『蔵書印データベース』[2]には，篆書体で彫られた蔵書印の画像データが含まれており，本研究では，『白川フォント』からの字形情報の複数特徴量の抽出に基づき，類似字形を検索する手法を提案し，その検索結果と『蔵書印データベース』など複数の人文系データベースとの連携検索の活用可能性を検討する．

2. 関連研究

近年，画像検索や画像認識の研究が盛んに行われている．古代文字の文字認識システムの構築は，人文系の研究において重要な役割を果たす．人文系データベースから取得できる情報に応じて，文字認識手法，画像検索の手法をそれぞれ考える必要がある

2.1　画像検索システム

蔵書印を自動的に解析するシステムとして，富士通による蔵書印検索技術のプロジェクト[3]がある．中国の蔵書印画像を対象として構築されたデータベースを用い，蔵書印の印文全体をマッチングなどの技術で認識する実験が進められている．青池ら[4]は，蔵書印のデータも含まれる資料画像の挿絵領域を自動的に抽出し，図案の辺縁特徴の特徴量を利用した画像検索システムを提案した．

2.2　文字認識

近年では，英文字を対象とした文字認識の研究はす

でに高い精度が得られており，文字認識の手法としては，画像に含まれる文章や手書き文字を認識し，テキストおよびその内容に関連する情報をユーザにフィードバックする手法が提案されている．例えば，Luo らによる研究[5]では，文字の構造特徴を抽出する手法を用い，中国の手書き漢字を認識する実験を行った．Deepa らは，Zernike モーメントと対角線の特徴を利用し，タミルの手書き文字を認識する手法[6]を提案した．近年，オフライン文字認識の手法として，畳み込みニューラルネットワーク（CNN: Convolutional Neural Network）の応用が注目されている．人の神経回路網を数式的なモデルで表現し，「畳み込み層」により画像からのエッジ等の特徴を抽出し，画像の抽象化を行い，分類を行う．大量の学習データに基づき，手書き文字を高い精度で認識することができる．CNN を用いた手書き漢字の認識システムとして，スタンフォード大学の学生により『ETL 文字データベース』[7]の漢字データを 99.64%の精度で認識できたとの報告がある[8]．

2.3 古代文字の文字認識

篆文と同じ形式の象形文字には，甲骨文や金文などの書体が含まれる．古代文字認識の研究では，碑刻，木簡，史料などに書かれた古代文字の領域を計算し，画像に含まれる文字をテキストとして抽出することを目指している．古代文字を対象として文字認識を行う研究もあり，深層学習で甲骨文字を 114 種 820 枚から 184 種 2000 枚に文字種を増やして実験を行い，94.4%の認識率を達成した研究もある[9]．他の認識手法として，古代文字データベースを構築し，候補テンプレートを検索することにより文字認識を行う研究がある[10]．そして，Clanuwat ら[11]は，深層学習を用いた日本語くずし字認識手法を提案した．

本研究は，様々な蔵書印画像データと背景情報を公開している国文学研究資料館『蔵書印データベース』[2]，九州大学附属図書館『九州大学蔵書印データベース』[12]の 2 つのデータベースを検索支援の対象として実験を行う．

3. 提案手法

中心線特徴およびコーナー特徴は，字形データを用いたテンプレートマッチングの研究によく使われる特徴量であり，本研究では中心線特徴，コーナー特徴を含む複数の特徴量を抽出し，特徴量のデータベースを構築することで，字形データの拡張性を考慮した検索ベース古代文字認識手法を提案する．

3.2 データの取得および前処理

表 1 に「白川フォント」における篆文の例を示す．

表 1 「白川フォント」における篆文の例

現代文字	人	文	科	学
白川フォント				

より良い字形の特徴量を抽出するため，「白川フォント」から抽出されたデータの前処理を行う．前処理は，画像のサイズの規格化，画像の二値化，字形の細線化を含む．画像のサイズは全て224×224ピクセルに統一し，画素の色特徴と空間分布特徴を利用し，k-means による画像の二値化を行う．図 1 の結果例に示すように，細線化については，Zhang ら[13]が提案した細線化手法で中心線特徴の抽出を行う．

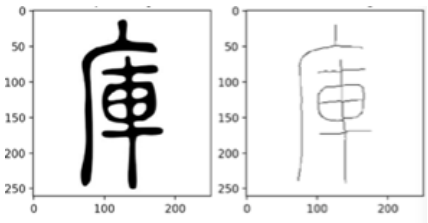


図 1 細線化による結果例

このアルゴリズムでは，以下の二つのステップで 2 値画像に対するピクセル処理を行う：

P9	P2	P3
P8	P1	P4
P7	P6	P5

図 2 細線化フィルタ

図 2 に示すように，処理は一つのピクセル P1 に注目し，周りのピクセルには P2～P9 の番号をつける．前景ピクセル（黒）は 1 で表され，背景ピクセル（白）は 0 で表される．

ステップ 1：①2 <= N(p1) <=6　②S(P1) = 1　③P2 * P4 * P6 = 0　④P4 * P6 * P8 = 0 の四つの条件があり，N(P1)は注目される画像の周りの前景ピクセルの数，S(P1)は P2，P3，P4，P5，P6，P7，P8，P9 と並べて順番に見ていったとき，0 の次が 1 となっている場所の個数である．以上の全ての条件を満たすピクセルに消去マークを付ける．

ステップ 2：ステップ 1 と同じように，①2 <= N(P1) <=6　②S(P1) = 1　③　P2 * P4 * P8 = 0　④P2 * P6 * P8 = 0 の四つの条件があり，条件をすべて満たすピクセルに消去マークを付ける．

以上の 2 つのステップを，どちらのステップでも変更点が無くなるまで繰り返す．

3.3 字形における幾何特徴の抽出

構造により漢字の字形を分類することも一つの手法として考えられる．図 3 に示すように CHISE project[1]では，ISO/IEC 10646-1:2000 の IDS 形式に基づく漢字の構造情報データベースを開発している．

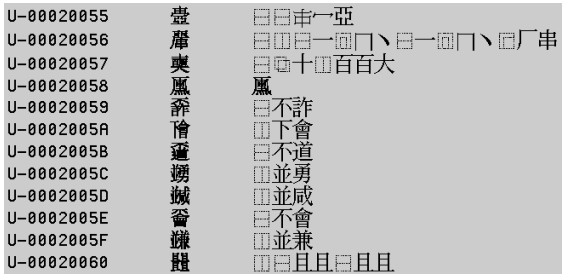


図 3　漢字の構造情報データベース

また,『新漢英字典』(ハルペン, 1990)は SKIP（System of kanji indexing bypatterns:字型式検字法）と 名付けられた新しい検字法を採用している.図 4 に示すように, SKIP は部首の知識がなくても検索できるように, 字形の特徴から 1（左右）型，2（上下）型，3（囲み）型，4（全体）型の 4 つのパターンに分け，各構成要素の画数によってさらに下位分類することによって 3 桁の SKIP 番号を付け，検索する方法である．

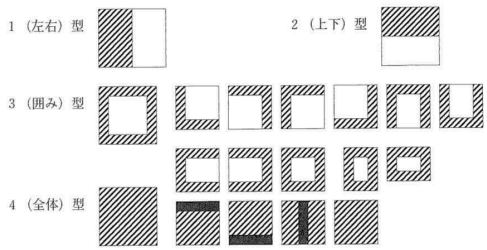


図 4　SKIP (System of kanji indexing by patterns:字型式検字法)

以上で述べた手法は全て字形の構造に基づいた発想である．単一文字に対する構造の分布の分析も重要なポイントであることがわかる．本研究では，図 5 に示すように，カーネル密度特徴による文字を構成する黒いピクセルの密度を計算する．画像の横軸と縦軸におけるカーネル密度推定量を数式 1 に示す．

$$\hat{f}_{bandwidth}(x) = \frac{1}{n*bandwidth}\sum_{i=1}^{n} K\left(\frac{x-x_i}{bandwidth}\right) \quad (1)$$

ここで，$\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$, …$\mathbf{x_n}$は，二値化された画像の印文を示す黒画素の x 座標の集合を表す$\mathbf{y_1}$, $\mathbf{y_2}$, $\mathbf{y_3}$, …$\mathbf{y_n}$は，それらの x 座標に対応する y 座標の集合を表す．bandwidth はバンド幅と呼ばれる平滑化のためのパラメータであり，K は数式 2 に示した Gaussian Kernel を用いる．画像の縦軸におけるカーネル密度推定量も数式 1，2 を用いて計算する．

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right) \quad (2)$$

カーネル密度推定は，画素の分布特性を分析するために使用できる．計算結果を特徴量データベースに保存する．

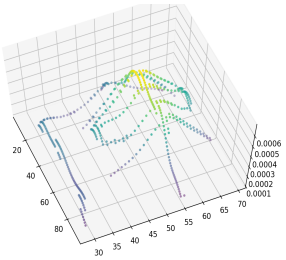


図 5　字形に対するカーネル密度推定の計算

細線化された字形データに対しては，ストロークの分析ができるようになる．Harris コーナーとは，画像の輪郭などの幾何特徴量を分析するため， Harris ら[14]が提案した方法である．

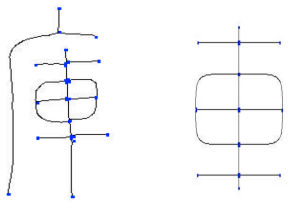


図 6 字形に対するコーナー情報の検出

Harris コーナーの検出により，ストロークの交差点情報が簡単に取得できる．図 6 に示すように，類似している漢字に対してストロークの情報により区別できる．本研究では，検出されたコーナーの座標情報を特徴量としてデータベースに保存する．

3.4 学習済みニューラルネットワークモデルから特徴量の抽出

[1] http://www.chise.org/ids/index.ja.html

古代文字の字形には多くのバリエーションがあり，同じ文字でも字形が変わることで，輪郭，中心線特徴，コーナー特徴による特徴の抽出結果にも大きな影響がある．畳み込みニューラルネットワークの構造により，画像を抽象化することができる．

データが足りない状況における深層学習では，近年，転移学習がよく使われる手法である．転移学習では，足りない訓練データのある属性と似たような特徴がある充実させたデータを収集し，訓練データとしてモデルを訓練する．学習済みモデルの前の層のパラメータをフリーズさせ，認識したい少量のデータで再訓練する．本研究では，再訓練を行わず，学習済みモデルを特徴抽出器として使う．特徴抽出器の訓練データには，現代漢字の手書きデータセット「CASIA Online and Offline Chinese Handwriting Databases」[15]を用いる．ニューラルネットワークには，VGG16[16]を使用する．

3.4.1 現代漢字の手書きデータを用いた特徴量抽出器の作成

現在使われている漢字は，形の上で古代の書体と大きな共通の特徴がある．ニューラルネットワークを用いることで字型の抽象化方法を学習できることが期待される．本研究では，字形の特徴を字形の種類を区切りやすい空間に圧縮し，その空間で入力された画像を表示できる特徴量を抽出する．図 8 に示すように，畳み込みニューラルネットワークの訓練データは現代文字の単一画像であり，訓練済みモデルを特徴量検出器として使う．

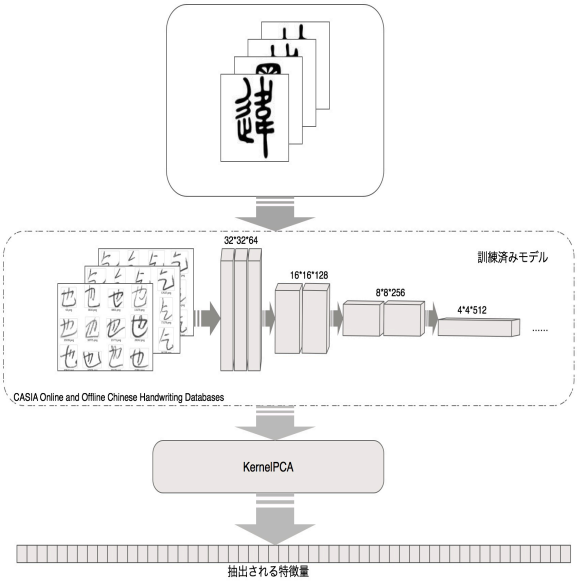


図 7　特徴量抽出器の作成

抽出された中間層データは高次元ベクトルであるため，本研究では KernelPCA[17]で特徴量の次元削減を行う．正定値カーネルは，数式 3 に示したガウス RBF（radial basis function）カーネルを用いる．

$$exp(-\frac{1}{2\sigma^2})\|x-y\|^2 \quad (3)$$

3.4.2 特徴量抽出ための中間層の選択

Kavukcuoglu [18]らの研究によると，入力層に近い中間層の出力は抽象的な特徴量であることがわかる．入力層に近い中間層の情報を用いることで，同一の種類に属する画像の共通の特徴を抽出することができると思われる．本研究では，プーリング層の出力に注目する．特徴抽出器の各プーリング層の出力特徴の部分データの可視化画像を図 8 に示す．

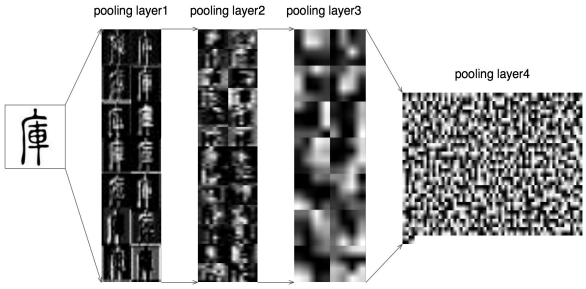


図 8　中間層出力の可視化例

図 9 に示すように，蔵書印データベースにおける出現頻度の高い文字として，「庫」と「印」二種類のそれぞれ 20 枚の単一文字の印文画像をランダムに選択し，これを例として中間層特徴の選択について説明する．

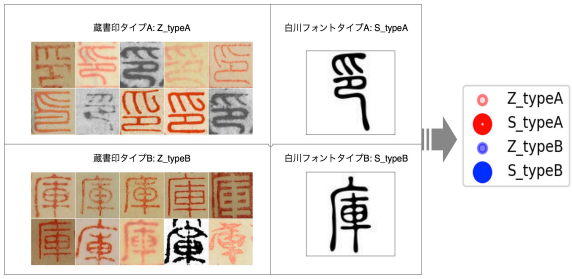
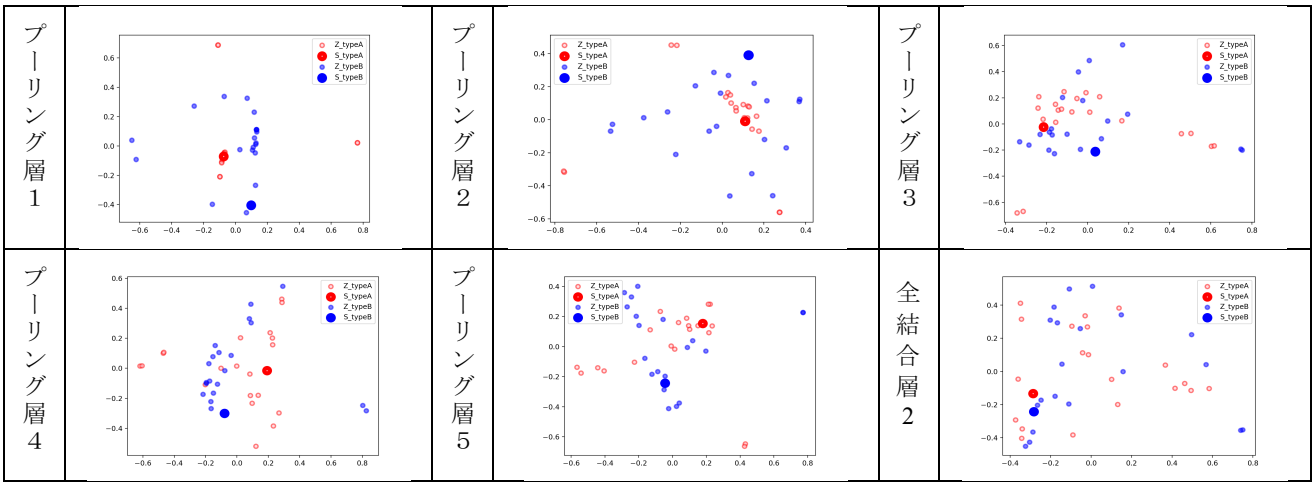


図 9　例として使用したデータとそれぞれの表示マーク

KernelPCA で 2 次元に次元削減した散布図を表 2 に示す．図の縦軸と横軸は 2 次元におけるデータの座標を示す．結果より四番目のプーリング層（プーリング 4 層）の特徴空間で，二種類のデータ分布が規則的になっている．異なる種類のデータが分離され，同類のデータが集まっている．このような分布は類似度に基づいた検索に有利である．そのため，本研究では，プーリング 4 層の特徴量と，分類情報を含む全結合層 2

の出力を抽出した．

表 2　特徴量の散布図



3.5 ランキング計算結果による文字の推定

図 10 に示すように，本研究では，各特徴量におけるランキングの重み付け計算により出力ランキングを計算する．

特徴量抽出器により取得した特徴量の距離計算によるランキング結果を用いた枝刈り演算を行うことで，正解データの平均順位により，上位複数枚の結果画像が残され，次の計算に用いられる，w_i は検索結果を改善するために設定する重みであり，初期設定は 1 である．

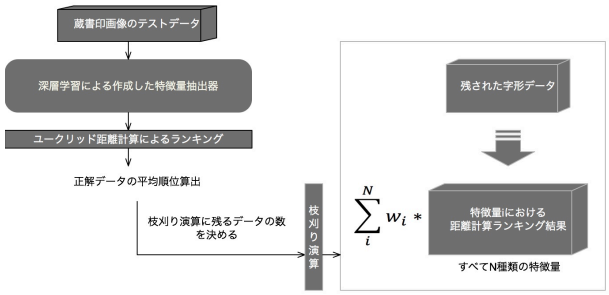


図 10 ランキング計算

4. 実験

本実験で用いたテスト用の単一文字の種類は 10 種類であり，データベースから抽出したラベル付き画像の文字の出現頻度の計算により決められる．

手書き漢字の訓練データは，いずれも蔵書印の背景特徴を持っていないため，本研究では単一の文字の蔵書印画像を：①二値化処理だけ行う②二値化した後に中心線の特徴を抽出する，の二つの組に分けて前処理を行う．文字「庫」のデータを図 11 に示す．

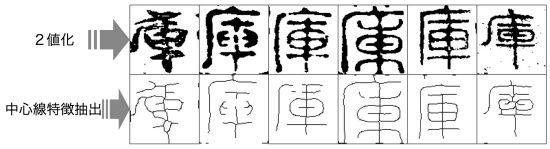


図 11　実験用の単一文字データの一部

20 枚のテスト画像の検索結果の可視化例を示す．プーリング 4 層（pooling4）と全結合層 2（fc2）から抽出した特徴量だけを使った検索結果を図 12 と図 13 に示す．コサイン類似度の計算によりランキング順位を計算する．

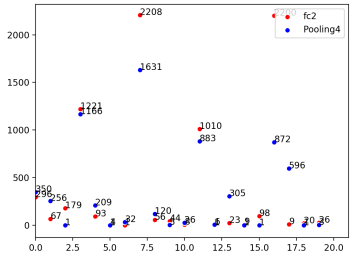


図 12　二値化処理だけ行われたテストデータのランキング結果

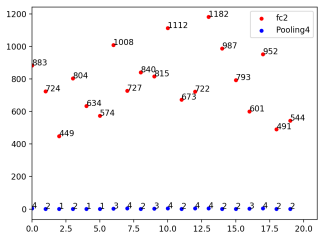


図 12　テスト画像の中心線特徴における検索結果

図 12，13 に示す赤い点と青い点は全結合層 2 とプーリング 4 層の特徴量における正解データのランキング順位を表す．図の縦軸は，横軸の画像番号に対応するランキング結果を示す．結果より，本研究の文字推定では，入力画像の二値画像の全結合層 2 の特徴量と中心線特徴図のプーリング 4 層の特徴量を用い，他の特徴量とともに 3.5 節で述べたランキング計算を行う．

テスト画像は文字の種類ごとに 20 枚を用いる．評価は平均逆順位（Mean Reciprocal Rank，数式 4）を用いて行う．Q は検索された画像の総数，i は画像番号，rank はランキングの順位である．

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank_i}} \quad (4)$$

実験結果の MRR のスコアを表 3 に示す．

表 3　検索結果の MMR スコア

文字	MRR スコア	文字	MRR スコア
印	0.0066	書	0.0212
文	0.2588	庫	0.4375
蔵	0.0188	之	0.2563
木	0.1112	氏	0.1880
図	0.1427	山	0.4833

今回のランキング重みはすべて 1 に設定されているため，重みの調整により検索精度の向上が期待される．

5. まとめ

本研究では，筆者らによる先行研究[19]の提案手法を特徴抽出手法の一部として使い，複数特徴を用いた蔵書印の字形検索を試みた．検索効率を高めるために，最終稿では，特徴量マージ実験，NGT[20]などの検索手法に基づく検索実験を行い，比較評価について詳しく述べる．今後は，文字解説データベース，漢字の構造情報データベースとの連携検索の活用可能性を検討する必要がある．

参　考　文　献

[1] 立命館大学白川静記念東洋文字文化研究所：白川フォント <http://www.ritsumei.ac.jp/acd/re/k-rsc/sio/shirakawa/index.html>（参照 2017-10-15）：

[2] 国文学研究資料館：「蔵書印データベース」，入手先：＜http://base1.nijl.ac.jp/~collectors_seal/＞（参照 2017-12-25）

[3] “Seal Retrieval Technique for Chinese Ancient Document Images， Fujitsu Research & Development Center Co. Ltd”. http://www.fujitsu.com/cn/en/about/resources/news/press-releases/2016/frdc-0330.html, (参照 2018-2-10)

[4] 青池亨，里見航，川島隆徳．資料画像中の挿絵領域の自動抽出及び画像検索システムの実装，人文科学とコンピュータシンポジウム論文集，pp.97-102 (2018)

[5] Luo, Yuchen, Rui Xia, and M.Abdulghafour. Offline Chinese Handwriting Character Recognition through Feature Extraction. Proc. 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV) (2016)

[6] Deepa，Ashlin，and R. Rajeswara Rao. An Efficient Offline Tamil Handwritten Character Recognition System using Zernike Moments and Diagonal-based features. International Journal of Applied Engineering Research 11(4), pp.2607-2610 (2016)

[7] “ETL 文字データベース”. http://etlcdb.db.aist.go.jp/, (参照 2018-2-10).

[8] Tsai， Charlie. Recognizing handwritten Japanese characters using deep convolutional neural networks. Report of Stanford University, pp.1-7 (2016)

[9] 紙徳直生，伊藤大喜，多田晃己，孟林，山崎勝弘．深層学習を用いた甲骨文字認識．第 80 回全国大会講演論文集， pp.513-514 (2018)

[10] 石井康史，藤川佳之，孟林，山崎勝弘．特徴量を用いた甲骨文字の候補テンプレート抽出と認識．第 78 回全国大会講演論文集，pp.211-212. (2016)

[11] Tarin Clanuwat, Alex Lamb, Asanobu Kitamoto. End-to-End Pre-Modern Japanese Character (Kuzushiji) Spotting with Deep Learning，人文科学とコンピュータシンポジウム論文集, pp.15-20 (2018)

[12] 九州大学附属図書館：「九州大学蔵書印データベース」< https://www.lib.kyushu-u.ac.jp/ja/collections/qstamp>（参照 2017-12-25）

[13] Zhang, T Y, Suen, C Y. A fast-parallel algorithm for thinning digital patterns. Comm ACM, Vol.27, No.3, pp.236-239 (1984)

[14] Harris, C. and Stephens, M. A Combined Corner and Edge Detector. In Proceedings of the 4th Alvey Vision Conference. pp.147-151 (1988)

[15] Liu, Cheng-Lin. CASIA online and offline Chinese handwriting databases. Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE (2011)

[16] Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint, arXiv:1409.1556 (2014)

[17] Mika, S., Schölkopf, B., Smola, A. J., Müller, K. R., Scholz, M., and Rätsch, G. Kernel PCA and de-noising in feature spaces. In Advances in neural information processing systems, pp.536-542 (1999)

[18] Kavukcuoglu, Koray, et al. Learning convolutional feature hierarchies for visual recognition. Advances in neural information processing systems, pp.1090-1098 (2010)

[19] 李 康穎，Batjargal Biligsaikhan，前田 亮．古代文字フォント字形の特徴抽出に基づく蔵書印の検索支援．人文科学とコンピュータシンポジウム論文集, pp.123-128 (2018)

[20] Iwasaki, M.: Proximity search using approximate k nearest neighbor graph with a tree structured index. IPSJ Journal 52(2), pp.817-828 (2011)