4G-04

古活字版カタカナ活字データベースの構築
― 活字画像の切り出しと分類手法の検討 ―§

杉山 正治*, 村上 明子**
大谷大学*, 関西外国語大学**

1. はじめに

日本の古文書には古活字版と呼ばれる印刷方法が存在する．これは，安土桃山末期から江戸初期にかけて広く作られた活字印刷本の総称（日本歴史大事典）である．古活字版の技法や伝搬の様相を明らかにする研究では，古文書の紙面に印字された古活字を切り出して分類し，照合していく．しかし，その作業量は膨大であり，人力では限界がある．これに対し，活字の切り出しと分類を自動化できれば，古活字研究のみならず，出版文化史研究の進展に寄与するものと考えられる．

古文書研究の支援には以下２つのアプローチがある．

ひとつは人海戦術型（代表例：みんなで翻刻 [1]）である．長所として，人が読むため認識精度が高いこと，更に入力支援システムが使えることなどが挙げられる．短所として，多大な労力と膨大な作業時間が必要な上，未登録の古文書の場合は即時利用できない問題もある．

もうひとつは自動認識型（代表例：くずし字認識アプリ「みを」[2]）である．長所として，機械学習ＡＩで自動的に解読するため，即時利用できることが挙げられる．短所として，認識精度は学習データ依存である上，学習データの準備には人海戦術が必要である．

これらはいずれも，古文書の記載内容が読めさえすれば良いという単一の方針しかなく，文字領域を囲む矩形とそのテキスト表示しか出力されない．すなわち，古活字研究者が望んでいるような，活字（字の形そのもの）の比較や分類には殆ど利用できない状況である．

著者らはこれまでに，古活字版の漢字活字およびカタカナ活字の一部サンプルのみのデータベースを構築してきた [3][4]．しかし，手作業での分類作業は煩雑を極め，未だデータベースを完成させるには至っていない．

本研究では国立国会図書館オンラインで公開されている和玉篇巻下（慶長年間）[5] を対象とし，カタカナ活字を切り出して分類できる画像処理システムを提案する．本システムは機械学習 AI ではなく旧来の画像処理のみで実装しているが，高い認識精度で活字領域の矩形を抽出して行列位置を特定した上，カタカナ活字を類似度順に並べて表示することができる．以下では本システムの概要と実行例を示し，有効性を述べる．

2. 古活字の種類と寸法

表１に本研究で取り扱う古活字の種類と寸法を示す．これらの寸法は，著者らのこれまでの古活字研究 [3] で確認された活字の現物の大きさであり，活字の字の形を包含できる最大外形寸法（定数扱い）に設定できる．なお，⑥と⑦はエラー判定用に用意したサイズである．

表 1: 古活字の種類と寸法（[3] より設定）

活字の種類	幅 [mm]	高さ [mm]
① 割注用カタカナ活字	5.1	4.8
② 割注用漢字活字	10.3	8.7
③ 割注用二字活字	10.3	12.9
④ 本文用漢字活字	20.7	21.1
⑤ 魚尾活字	26.0	15.1
⑥ 割注用カタカナ二字活字	5.1	9.6
⑦ 活字領域サイズ（最大値）	26.0	21.1

図 1: 画像サンプル [5]

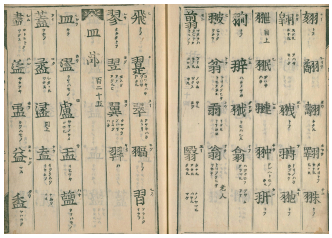

図 2: トリミング画像

3. 印字書式の概要

図１に本研究で取り扱う和玉篇巻下の画像サンプルを示す [5]．この画像には古文書の紙面と共にメジャーが並べられている．このスタイルの概要を以下に示す．

(1) 左右のページは別々の紙および版下
(2) 活字の行列位置は整然と揃っていない
(3) 左右のページ毎に，漢字６行５列で構成
(4) 漢字毎に，カタカナ５行で構成
(5) 漢字の有無，カタカナの有無など，多様な配置
(6) カタカナ領域に割注漢字や二字活字が見られる
(7) 魚尾を含むタイトル行にカタカナは見られない

なお，処理時間短縮のため，本研究では図２のように予め余白をトリミングした画像を用意して解析する．

4. 開発環境

本システムは HTML/CSS/Javascript で構築している．モダン Javascript が動作するブラウザ（Edge や Chrome 等）が必要である．OS には殆ど依存しないが，キーボードとマウスの利用を前提として設計している．

5. しきい値・換算値

画像の２値化しきい値を初期値 128 とする．また，サンプル画像 16 枚のメジャーを読み取ってピクセル数の平均を求め，整数の有効数字３桁の換算値として 50[mm] を 856[ピクセル] とする．この換算値を用いて，表１の寸法をピクセル数に変換できる．なお，画像毎のしきい値と換算値は解析前に手動で変更可能である．

§Segmentation and Classification method for developing a Database with the Japanese Katakana Characters Printed by the Old Movable Types
*Seiji Sugiyama: Otani University
**Akiko Murakami: Kansai Gaidai University

6. 探索アルゴリズム

6.1 枠情報取得

以下の手順により全ての枠情報を配列 **P** で管理する．

(1) グレースケール変換後，２値化
(2) ラベリング
(3) ラベル毎の外形包含枠を取得
(4) 枠同士の重なりを１つの枠に統合
(5) 表１⑦の活字領域サイズを超えた枠を除外
(6) 枠に表１①〜⑥の活字の種類（フラグ）を設定

6.2 漢字行列位置取得

以下の手順により漢字行列位置を配列 **C** で管理する．

(1) 左右ページで処理を分割
(2) 配列 **P** から漢字枠と魚尾枠のみ抽出
(3) 黒塗りした枠のみでヒストグラムの帯を取得
(4) 縦横の帯の交差領域を漢字位置と推定
(5) 交差領域内の枠を統合して漢字枠を確定
(6) 漢字枠が交差領域外ならタイトル行と仮定

6.3 カナ行列位置取得

6.3.1 ５行のカナ領域

以下の手順によりカナ行列位置を配列 **R** で管理する．

(1) 漢字毎に，５行のカナ領域を設定
(2) 上端位置は漢字からのカナ幅・高さで設定
(3) 下端位置は次のカナ領域直前に設定
(4) 配列 **P** からカナ枠のみ抽出
(5) １行目に重なるカナ枠を抽出
(6) カナ枠を包含するよう領域位置を左右シフト

6.3.2 木枠ゴミ対策

以下の手順により境界線付近のゴミ枠を除外する．

(1) 行毎に，カナ領域１行目の全カナ枠を抽出
(2) 黒塗りした枠のみでヒストグラムの帯を取得
(3) 帯の右側にカナ幅以上離れたゴミ枠を除外

6.3.3 上下左右分離カナ枠の統合

以下の手順により分離されているカナ枠を統合する．

(1) 左右に並ぶ枠を統合
(2) 枠の上下隙間リスト作成
(3) 近い順で上下に並ぶ枠を統合
(4) 極小枠を除外
(5) １行目のカナ枠を確定

6.3.4 クラスタ処理

以下の手順によりカナ行列位置を確定する．

(1) 行毎に，カナ領域１行目の全カナ枠抽出
(2) 枠の上下隙間リスト作成
(3) 上下近接カナ枠同士でクラスタ作成
(4) クラスタ中心位置の最寄りの漢字を特定
(5) 管理する漢字カナ領域の修正
(6) カナ行列位置を確定
(7) カナ領域の高さを最小化

２行目以降も１行目と同様の探索と統合でカナ行列位置を確定する．ただし，クラスタ処理は不要である．

6.4 類似度一覧表取得

以下の手順によりカナ枠の類似度一覧表を作成する．

(1) カナ枠を正方形中央に配置して 1/4 倍リサイズ
(2) 正方形画像の２値情報をベクトルとみなす
(3) 全カナ枠正方形同士でコサイン類似度算出
(4) コサイン類似度の値で降順ソート

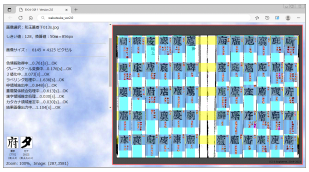

図 3: 活字領域表示の例

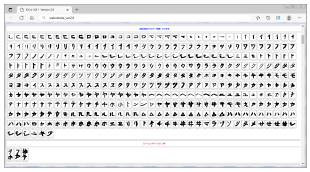

図 4: 活字一覧表示の例

表 2: カタカナ活字抽出結果

	抽出率 [%]	正答率 [%]
検証データ	96.9	99.8
テストデータ	95.5	99.5

抽出率＝正しいカナ枠総数÷取得したカナ枠総数，正答率＝正しいカナ枠総数÷カナ活字総数

7. 実行結果

図 3 に活字領域表示の例を示す．漢字行列位置の帯，漢字毎のカナ領域５行の帯，および活字の枠が表示される．枠をクリックすると画面左下にその活字が並ぶ．

漢字位置は (紙面, 漢字行, 漢字列) の３項で構成される．カナ位置は (紙面, 漢字行, 漢字列, カナ行, カナ列) の５項で構成される．例えば，カナ位置 (左,4,3,2,1) は紙面左側の漢字４行３列目のカナ２行１列目を表す．

図 4 に類似度順のカタカナ活字一覧表示の例を示す．画面下にカナ２文字が繋がった誤検出一覧が表示される．

8. 考察

表 2 に抽出率と正答率を示す．この値は検証データ 16 画像，テストデータ 16 画像の平均である．本研究では文字認識処理を一切行わず，現物の古活字寸法のピクセル換算値とラベリングで得た枠寸法の比較により活字の種類を特定し，行列位置を求めただけであるが，抽出率も正答率も 95[%] を越える高い値を得た．

誤検出の多くは木枠の縦横罫線または点在するゴミであったが，これらは類似度順に並べることで容易に区別できるため，多くの場合，殆ど問題にはならない．

今回の仕様では，漢数字「二」などの一部をカナと誤認する．また，紙面の傾きが大きい場合，１字の複数ラベルが複数行に分離することがある．さらに，紙面内で濃淡が不揃いな場合，しきい値が上手く機能しないこともある．これらについては要検討事項である．

9. おわりに

本研究では，旧来の画像処理のみで，和玉篇の書式に合わせたカタカナ活字切り出しと類似度一覧表示を実現し，活字の比較・分類が容易になった．今後は更なる認識精度向上と複数画像間での整理手法を検討する．

謝辞

本研究は科学研究費補助金 20K00355 の助成を受けたものである．

参考文献

[1] 橋本雄太，加納靖之，一方井祐子，小野英理，“『みんなで翻刻』の運用成果と参加動向の報告”，情報処理学会人文科学とコンピュータシンポジウム 2020, pp 39–46, 2020
[2] カラーヌワット・タリン，北本朝展，“資料調査のための AI くずし字認識スマホアプリ「みを」”，情報処理学会人文科学とコンピュータシンポジウム 2021, pp 302–309, 2021
[3] 村上明子，“古活字版成立に関する総合的研究–仁和寺所蔵古活字を中心にして– ”，科学研究費・研究成果報告書 19520179, 2008
[4] 村上明子，“古活字版伝播に関する研究–仁和寺所蔵古活字のデジタル画像分析を基盤にして– ”，科学研究費・研究成果報告書 25580059, 2016
[5] “和玉篇 巻下（慶長年間）”，国立国会図書館オンライン，http://dl.ndl.go.jp/info:ndljp/pid/2532152，閲覧日 2023.11.18