Vol.2017-CH-115 No.3
2017/8/4

説文解字書影対比をテンプレートマッチで行う際のパラメータ自動設定について

鈴木 俊哉
suzuki toshiya
〒739-8511 広島県東広島市鏡山 1-4-2　広島大学　総合科学研究科
mpsuzuki@hiroshima-u.ac.jp

概要: 清代に翻刻された説文解字にはいくつかの版本系列があるが、底本が特定されていない系列があることや、現存する宋本が必ずしも誤り最少の資料と言えないことがあり、清刊本の版本比較はまだ残っている課題と言える。 説文解字のうち、特に現存する宋本にもっとも近い小字本の一群は文字感覚の狭さや不均一な字配り、印刷品質の問題のため、OCR 的な画像分解は難しい。そこで、画像分解のコストを下げるため、ある版本に対して手作業で作成したレイアウト分析データを、同系列の他の版本で共用するためのパラメータ自動設定の方法を考える。

1. はじめに

『説文解字』は秦代の小篆に基づいて漢の許慎が編んだ最初の部首引き字書である[1][2]。宋初に徐鉉が校訂した、いわゆる大徐本が現在完本として残る最古の資料だが、南宋の李燾が大徐本を韻書排列に組み替えた『説文解字五音韻譜』が非常に広く通行し、明代には「許慎の説文は韻書排列ではないが、大徐が校訂して韻書排列にした」という誤解が広まった。この誤解を解いたのが明末清初の汲古閣による説文解字の翻刻である。現代の説文学はこの汲古閣本を発端とする一連の流れに連なるとも言えるだろう。

汲古閣本が刊行されてから 84 年後、段玉裁の『汲古閣説文訂』により、通行の汲古閣本は宋本を忠実に翻刻したものではなく、主に小徐本によって改められていることが指摘された。これ以降、変更の少ない宋刊小字本を翻刻しようという動きや、蔵書家が持つ宋刊本を通行の資料と比較する説文の校勘研究が盛んとなった。

この動きは、最終的に宋刊元修本とされる岩崎本が續古逸叢書によって影印出版されることで収束したが、この影印出版にも加筆の可能性があり[1][3]、宋本の状況を正確に伝えているかには疑問があることが知られる(図 1)。

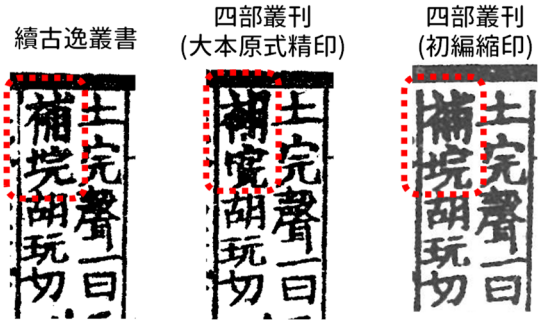

図 1: 岩崎本影印出版の加筆と思われる差異

現在もっとも広く流布し参照される陳昌治本(いわゆる一篆一行本)においても、版本ごとの差異[5]や、影印出版の際の変更が疑われる部分があることが知られる(図 2)。

また、清代の校勘研究で参照された宋本には、岩崎本と若干の異なりがあるものも報告されている。『説文解字詁林』などに収集される古典的な研究は、岩崎本の影印出版以前なので、それらが宋本と言った場合に岩崎本と対照させて良いのかは必ずしも明らかではない。

さらに、近年、修正が少なく permissive な条件で公開されているデジタル画像を用いて、加筆問題や再利用に制限がある影印出版に基づいた研究を検証・更新するなどの需要もある。

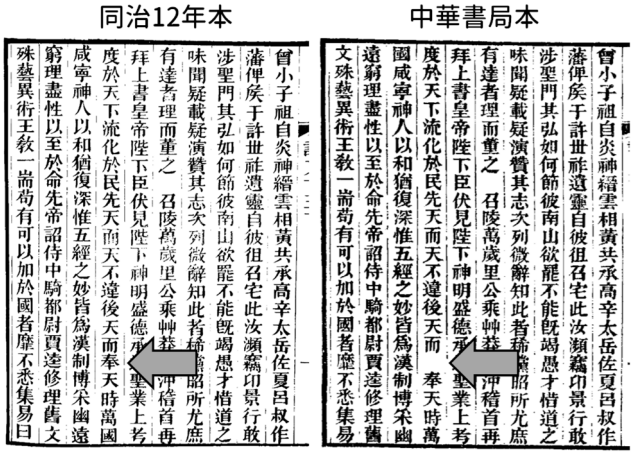

図 2: 陳昌治本の影印出版における変更が疑われる例

以上から、現在でも説文の版本比較には以下のような需要があると言えるだろう。

① 岩崎本以外の宋本について明らかにするため[3]。
② 現存する宋本、清本、および過去の説文研究でひかれた資料の参照関係を明らかにするため。
③ 現行の影印本の加筆箇所を明らかにするため[5]。
④ 影印本に基づく先行研究を原本の画像で再確認するため。

2. 説文版本対照への情報処理技術の応用

文献の版本対照に情報処理技術を応用する場合、すぐに考えられるのは (符号化文字列による)デジタルテキスト化して比較するという方向性であろう。説文の校訂には長い積み重ねがあるが、版本の参照関係を追跡するためには不鮮明な箇所や誤字の出現状況なども重要な情報であり、校訂したテキストを用いれば良いとは限らない。版本比較を目的とするならば、画像データベースのような形式が望ましいだろう。

説文解字のレイアウトは見出し字小篆、小篆と同じ文字サイズでの説解、割注が入り混じっている上、現存する宋本

ⓒ 2017 Information Processing Society of Japan

においては、罫線や前後の文字に接触してしまっている状況も多い。そのため、市販の OCR ソフトなどでは完全な画像分解が困難で、機械的な支援を受けるとしても、現時点では手動補正は必要と思われる[1]。

画像分解に手動補正が必要であれば、その補正結果を複数の版本の画像分解で共用する省力化も期待される。本稿では、レイアウト分析情報の共有のための、書影対応づけの自動化について検討する。

2.1 版面から見た大徐本説文解字の系列とその比較

大徐本説文解字には様々な版本があるが、多くは既存資料の翻刻に基づいたもののため、「巻 x の葉 y の行 z にはどのような字が見えるか」といったレベルまで踏み込んで共通化することができる。共用できる版本が多いグループとしては、①汲古閣本系(1 ページ 7 行、いわゆる大字本形式。汲古閣未改本、汲古閣通行本、朱筠本、江戸時代の官本など)、②宋刊小字本系(1 ページ 10 行。岩崎本、海源閣本、平津館本、藤花榭本、日照丁氏本など)、③一篆一行本系(1 ページ 10 行。陳昌治本とその翻刻・影印本など)の 3 つがある。

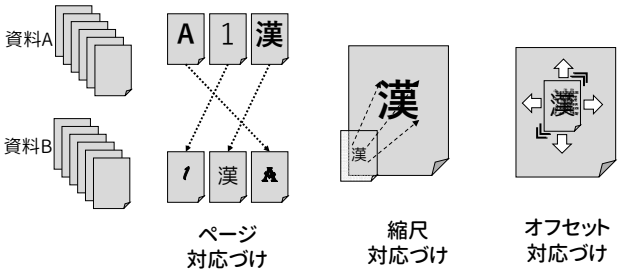

図 3: 対応付け操作の種別

本稿では、もっとも基礎的なレイアウト対応付けとして、同一系列の 2 つの版本で、別々の条件で撮影またはスキャンされた画像データ群に対し、対応するページを重ね合わせた比較ができる状況を目指す。機械支援が期待されるパラメータは、「同じページを指している画像ファイルを探す」(ページ対応)、「版面のサイズの対応づけ」(縮尺対応)、「画像ファイルの中の印字領域へのオフセットの対応づけ」(オフセット対応)の 3 つである。実際には、斜行・回転によるずれや、カメラ撮影の歪曲収差など、より高度な補正を加味すべき状況もあるが、特に影印本では撮影環境の情報も不明であり、数学的に定式化できるか、あるいは細分化してアドホックなパラメータ設定を行うべきかは、はっきりしない。今回はもっとも簡単な拡縮と平行移動だけを考える。

2.2 対応付けパラメータの依存関係

書影画像の重ね合わせに必要なパラメータ 3 つ(ページ対応、縮尺対応、オフセット対応)をテンプレートマッチによって自動設定することを考えた場合、最初に処理すべきものは縮尺対応と思われる。なぜなら、本稿で前提としている説文解字の場合、最も少ない②宋刊小字本系でも 500 ページ、最も多い③一篆一行本系では 1200 ページ以上あるため、縮尺が不明なままページ対応をとろうとすれば、「500～1200 種類の図形の識別を領域分割せずに行う」という文字認識の課題とほぼ同じになってしまうからである。拡縮・回転に影響されない特徴量によって物体検出を行う研究は少なくないが、それらは数百種類の物体を識別するものではないこと、1 ページの書影を 1 つの画像と考えた場合 500 種以上の物体を区別可能なほど説文の各ページに図形的差異があるとは考えにくいこと、などを考え合わせると、縮尺を最初に解決する手順の方が容易と思われる。そこで、本稿では、まず縮尺対応のパラメータをテンプレートマッチで検出し、それをもとにページ対応を自動検出するという手順を想定した。

3. 評価実験

本稿での評価実験について整理する。

3.1 テストデータ

本稿での評価実験は、②宋刊小字本のグループで行った。

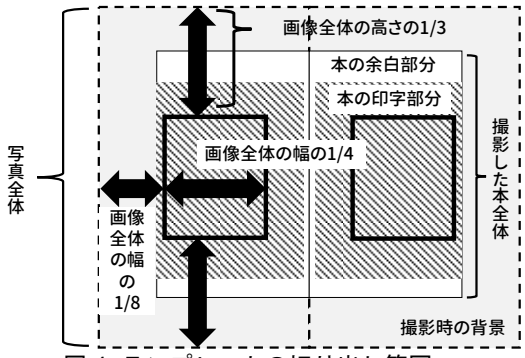

図 4: テンプレートの切り出し範囲

- テンプレート
  テンプレートは、国立公文書館所蔵の五松書屋版平津館本説文解字[2]を 400dpi でオーバーヘッドスキャナでデジタル化したデータを、一度 2 値化した上で 64dpi にスケールしたものを用いた。撮影画像から図 4 のような範囲を切り出した。画素数では 198×238 ピクセル、文字数では 6.5 行×11.5 文字程度のサイズとなる。テンプレートのピクセル数は全頁で共通だが、そこに含まれる文字の状況により JPEG 画像のデータ量は 10kB(巻末の、罫線のみで文字がない領域)から 42kB まで分布する。
  この領域は、主に国立公文書館および京都大学人文科学研究所東方学デジタル圖書館の画像データから、特段のパラメータ調整をせずに、背景や余白を含まない、印字部分の一部だけが取り出せるよう設定したものである。

[1] 坂内[6]・守岡[7]は『説文解字繫傳』を市販 OCR の PDF 出力を解析してデジタルテキスト化する試みを行った。その対象となった祁寯藻本は 1 ページ 7 行の大字本様式であり、字配りは十分周期的であるため、その領域分割はある程度の精度を保証できたように思われる。ただしこの場合でも見出し小篆が誤って割注の文字サイズの部分図形に分解されてしまうこともあったようである。

[2] 平津館叢書は孫星衍存命中に刊行された嘉慶年間本と、卒後に刊行された光緒年間本がある。国立公文書館に所蔵されるのは嘉慶年間本であり、世界書局の影印は光緒年間本を底本とする。出版経緯や内容の変更に関しては下見氏の論文に詳しい[8]。また、この他にも平津館本説文解字は多く翻刻されている[9]。

ⓒ 2017 Information Processing Society of Japan

- 母集団データ
  テンプレートと対照させるデータは以下を用いた。
  - 世界書局影印平津館本[12](以下、世界書局本)
    光緒年間の呉縣朱記榮重刻平津館叢書に含まれる説文解字を世界書局が影印出版したものである(A5サイズ影印本の1ページあたり、原本の1ページが影印されている)。同一の底本に直接に由来する複数の資料を対比するテストとして選定した。複合機スキャナにより600dpiモノクロ2値でスキャンし、200dpiにスケールした。縮尺テスト範囲は11～13%。テストしたページ総数は480ページ。
  - 四部叢刊影印岩崎本[13](以下、四部叢刊本)
    商務印書館の大本原式精印四部叢刊正編での岩崎本の縮印本(B5サイズ影印本1ページあたり2葉4ページが縮印されている)である。宋本翻刻本と宋本影印本を対比するテストとして選定した。複合機スキャナにより600dpiモノクロ2値でスキャンし、200dpiにスケールした。縮尺テスト範囲は70～80%。テストしたページ総数は505ページ。
  - 藤花榭本[14]
    早稲田大学図書館デジタルアーカイブが原本を撮影したカラー画像を公開しているもの。底本の関係がはっきりしない複数の翻刻本を対比するテストとして選定した。元画像データは140dpi程度と思われる。モノクロ2値化して使用した。縮尺テスト範囲は40～45%。テストしたページ総数は506ページ。
  - 日照丁氏重校本[15](以下、日照丁氏本)
    京都大学人文科学研究所の東方學デジタル圖書館が原本を撮影したカラー画像を公開しているもの。藤花榭本と同様に底本の関係がはっきりしない複数の翻刻本を対比するテストとして選定した。元画像データは200dpiである。モノクロ2値化し100dpiにスケールして使用した。縮尺テスト範囲は70～80%。テストしたページ総数は500ページ。

3.2 評価手法

本稿でのテンプレートマッチは、実装はOpenCV[10]のPythonバインディングのmatchTemplate()を用い、差異の評価アルゴリズムにはOpenCV組み込みの評価関数のうち最も単純な差分相関(SQDIFF)を用いた。テストする縮尺率へスケールした書影画像(全体)と、テンプレートの両方に$\sigma = 1$のガウスぼかしをかけ、マッチさせる処理を繰り返す(テンプレートは拡縮しない)。以下では、差異については理論的な最大値[3]で正規化した割合で記した。

3.2.1 縮尺率の自動検出

縮尺率の自動検出は以下のようにテストした。まず対象データに対しテスト範囲の縮尺率を1%単位で変化させた画像群を作り、巻号・葉数で対応づけたテンプレートを使ってマッチを行う。

あるページに対し、複数の縮尺率で作成した画像にテンプレートマッチを行い、スコアが最良となるものから、そのページの最適な縮尺率を得たとする。ある資料が同一の環境で撮影され、縮尺率だけが異なるものであれば、一つの縮尺率に収束すると期待できる。

3.2.2 ページ対応の自動検出可能性の評価

テンプレート画像のうち、JPEG圧縮後のデータ量が大きいものは、画像に細かい図形が多く、本来対応しないページにマッチさせた場合にスコアが悪くなると期待できる。そこで、まず最もページ対応の自動検出が有利になるであろう、JPEG圧縮後のデータ量が大きいものからテンプレートを10個とり[4]、対象データの全てとマッチさせた。

得られたスコアのうち、実際に対応づくページに対してテンプレートマッチを行ったものだけが特異的にスコアが良い(差異が少ない)という結果が得られれば、ページ対応の自動判別が有望と言える。

4. 実験結果

4.1 縮尺率

まず、縮尺率の自動検出のための実験結果は図5のようになった。各ページで、最適と推定された縮尺率(左縦軸)と、その縮尺率でのテンプレートマッチのスコア(横軸)をプロットし、各誤差以下のページ数を集めた累積率(右縦軸)もプロットした。

世界書局本に対する縮尺率の自動検出は、かなり良く収束する結果を得た。当初テスト範囲を11, 12, 13%で試験したが、11%や13%が最適となったページが得られず、11.5%や12.5%での試験も追加したが、大半が12%に集中している(480ページ中428ページが12%である)。また、差異も7%未満の範囲に収まっている。世界書局本に対する重ね合わせの結果の例を図6に示す(全体が世界書局本の書影で、中央の矩形領域でテンプレート画像を半透明で重ねている)。

この場合は、機械的にパラメータ設定をすることも充分有用と思われる。テンプレートの底本である五松書屋の平津館本とは重刊本の関係にあり、また底本は清刊本であって孤本ではないことから、被せ彫りのような精密な翻刻を行った可能性も考えられるだろう。

次に、日照丁氏本も、世界書局本ほどではないが、縮尺率はかなり良く収束する。73～74%の範囲であることはほぼ明らかで、差異も7%までにほぼ収まっている(図5の範囲外になるが、7%を越えるものは5ページある)。

日照丁氏本は、平津館本の翻刻ではなく、汲古閣旧蔵宋刊小字本をもとにしたとされるため、直接の参照関係がない平津館本と高い精度で一致するという結果から、葉徳輝の日照丁氏重校本は平津館本を加筆したもので実際には宋本に拠らないとした観察[3][16]にもある程度の妥当性があったと言えるだろう。

[3] テンプレートの全ピクセルが白のものと黒のものを合成して差異を取った結果、matchTemplate()の差分相関が返す差異の値は0～9192714240の範囲であった。

[4] 次に具体的なページとテンプレートのJPEGデータのサイズを列挙する。巻02下葉06右(42.1kB)、巻04上葉07右(42.0kB)、巻04上葉09右(42.2kB)、巻04下葉06右(42.1kB)、巻05上葉01左(42.3kB)、巻05上葉02右(42.4kB)、巻05上葉03左(42.1kB)、巻06上葉01左(42.1kB)、巻08葉05右(42.0kB)、巻13上葉07左(42.1kB)。

ⓒ 2017 Information Processing Society of Japan

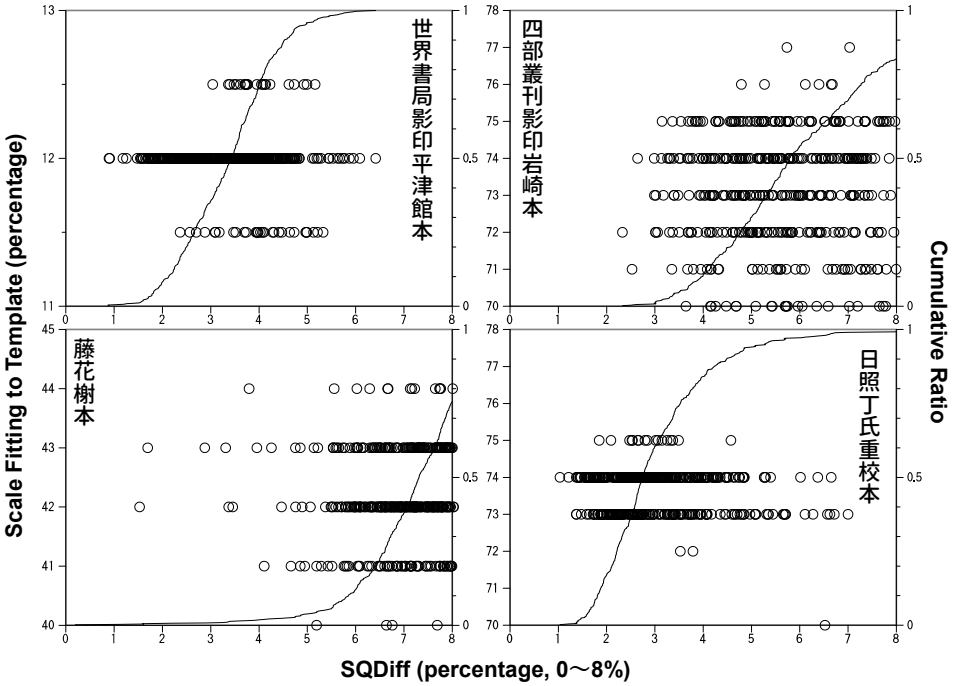

図 5: テンプレートマッチによる縮尺率の自動検出試験結果

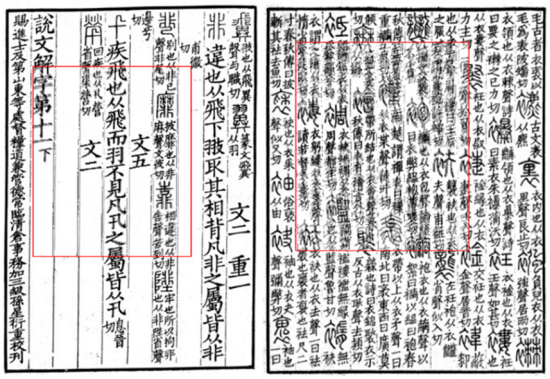

図 6: 世界書局本に対するマッチ例
(左:差異 1.3%、右:差異 6.4%)

図 7: 日照丁氏本に対するマッチ例
(差異 6.0%)

一方、四部叢刊本と藤花榭本はやや複雑である。四部叢刊本の場合、差異 7%までを念頭におけば縮尺率は 72～74%の範囲となるが、世界書局本や、日照丁氏本のような縮尺率が特定の値に集中する傾向が見られない。また、この範囲でカバーできる累積ページ割合は 70%程度であり、これを越えるものが 154 ページ残っている。テンプレートマッチの結果の例を図 8 に示すが、世界書局本に比べて影印自体に大きな歪みが残っていることがわかり、異なる撮影条件の影印を貼り合わせて作られた可能性が疑われる。

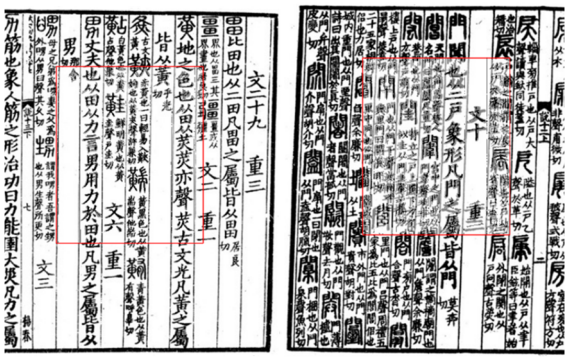

図 8: 四部叢刊本に対するマッチ例
(左:差異 2.6%、右:差異 11.3%)

図 9: 藤花榭本に対するマッチ例(差異 4.9%)

ⓒ 2017 Information Processing Society of Japan

藤花榭本は縮尺率に関しては四部叢刊本より分散が少なく、41～43%に収まっているが、差異については四部叢刊本より多い。7%を越えるページが全506ページのうち314ページと半分以上になっている。撮影条件が一定の藤花榭本においてここまで差異が多いことは、画像処理に起因する問題ではなく、版面自体に由来する問題と思われる。テンプレートマッチの結果の例を図 9 に示すが、多くのページで縦横比が合わないずれが目立つ結果である。

4.2 ページ対応

次に、縮尺率の判定結果をもとにしたページ対応の自動判別の結果を図 10 に示す。縦軸にテストしたテンプレート番号、横軸に正規化した差異をとってプロットした。

世界書局本(縮尺テスト範囲 11～13%)は、縮尺率においても良い収束を示したのと同様に、ページ対応においても、対応づくページとの差異は、対応しないページとの差異より明らかに少ない。差異 6～7%に閾値を置き、「ページ対応が正しいもの」「誤っている可能性が高いもの」を区分することができるようにも思われる。ただし、最もページ識別に有利なテンプレートであっても、誤ったページ対応から最大 15%程度の差異しか得られないことには注意が必要である。

一方、日照丁氏本(縮尺テスト範囲 73～74%)は、そのような区分は難しい。各テンプレートでのテスト結果において、ページ対応が正しいものと誤るものの間に明白なギャップがあるが、たとえばテンプレート#9 の「ページ対応が正しい場合の差異」は、テンプレート#1 の「ページ対応が誤る場合の差異の最小」とほぼ同じ程度であり、日照丁氏本全体に対して一つの閾値を設定することは難しいと思われる。

四部叢刊本(縮尺テスト範囲 72～76%)は「ページ対応が正しい場合の差異」と「ページ対応が誤る場合の差異の最小値」の間に閾値をおくことがまだ可能な程度のギャップがあるように思われる。ただし、「ページ対応が誤る場合の差異」の分布は、日照丁氏本に比べると四部叢刊本はやや狭い分布となった。

藤花榭本(縮尺テスト範囲 41～43%)では、既に「ページ対応が正しいもの」「ページ対応が誤るもの」の差異の間に大きなギャップは見られない。テンプレート#2, #5, #7, #9 はギャップがあるどうかの判断も難しい状態である。縮尺率の自動検出の結果に対して考察したように、もともとの版面の差異があり、差異をある程度以上に小さくすることができないと思われる。

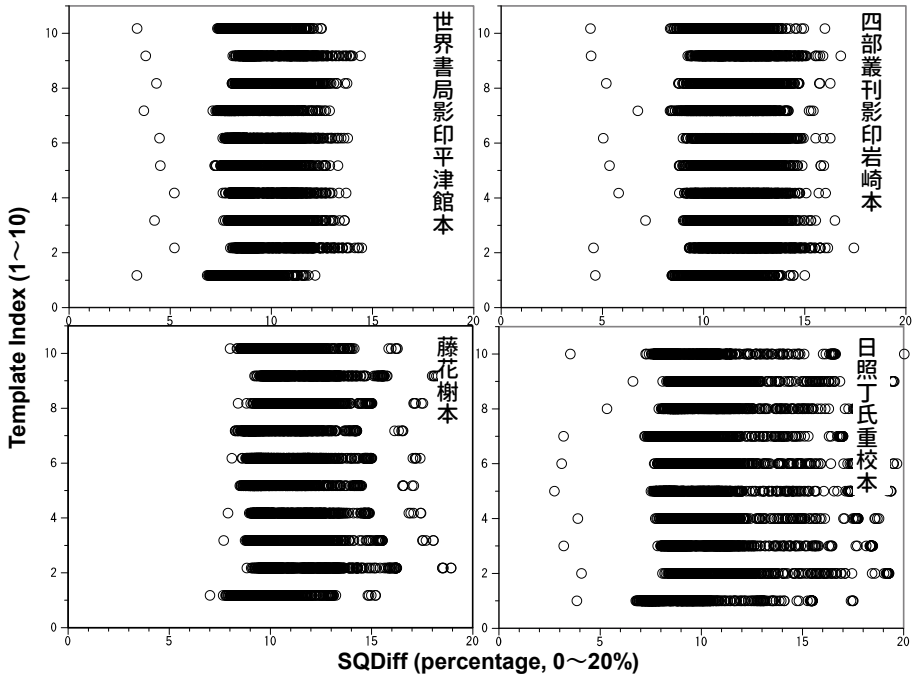

図 10: テンプレートマッチによるページ対応検出試験結果

5. まとめ、今後の課題

本稿では、同系列の説文解字の書影画像を比較するためのパラメータ設定について、テンプレートマッチによる自動設定の可能性を検討した。まず初期段階の検討として縮尺率およびページ対応について、差分相関での自動検出可能性を試験した。

その結果、縮尺率に関しては、底本に明確な参照関係があるもの(今回の場合、五松書屋版平津館本と世界書局影印平津館本)での有効性を確認でき、さらに同系列ではあるが明確な参照関係がないもの(今回の場合、四部叢刊影印岩崎本、日照丁氏重校本)でも有望であることがわかった。

一方、ページ対応検出に関しては、定本に明確な参照関係がない場合にはテンプレートマッチで計算される差異に対して一定の閾値を設けて判断することは難しいという結果を得た。

藤花榭本に関しては例外的に、特にページ対応の検出が難しいという結果を得たが、これは他の小字本と縦横比が異なるためと思われる。藤花榭本が底本となった宋本を忠

ⓒ 2017 Information Processing Society of Japan

実に翻刻しているかには疑問が指摘されており[3][17]、さらなる調査が必要である。

今後の課題としては、まず本稿で無視した縦横比の変化や回転歪みを補正するパラメータ自動設定がある。また、書影比較の最終目標は版本比較を機械的に行うことであるので、差異のスコアが大きいとすればどの部分によるのかを拾い出せるようにすべきである。本稿では 1 ページに対し 1 テンプレートした用いなかったが、複数のテンプレートを用いることでページを細分化した上での対比が今後の課題である。

謝辞

本研究は科研費課題番号 26330377, 16K004600 の補助を受けました。大西克也先生、高橋由利子先生、金木利憲先生、永崎研宣先生、王一凡氏、中山陽介氏、川幡太一氏に大変有益な議論と示唆を頂きました。ここに御礼申し上げます。

参考文献

[1] 頼惟勤監修、説文会編:『説文入門』, 大修館書店(1983) p.25-30.

[2] 福田襄之介:『中国字書史の研究』, 明治書院(1979) p.182-187.

[3] 倉田淳之助:「説文展觀餘錄」東方学報(京都) 第 10 冊第 1 分冊 (1939), p.145-154.

[4] 鈴木俊哉:「清刊大徐本説文解字の版本評価の再検討に向けて」, 環境科学研究 11 (2016), p.77-100.

[5] 田泉:「五種陳刻大徐本《説文》文字互異同挙例」, 古籍整理研究学刊 (2003/05), No.3, p.72-73.

[6] 坂内千里:「『説文解字繫傳』データベース構築の試み」, 漢字と情報(2003/10), No.7, p.4-5, https:// repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/57067/1/kanji-and-info-7.pdf

[7] 守岡知彦:「文字画像のマークアップの試み」,『東洋学へのコンピュータ利用』第 14 回研究セミナー(2003/03), http://www.kanji.zinbun.kyoto-u.ac.jp/~tomo/svg/tid-oricom2003.pdf

[8] 下見 隆雄:「「平津館叢書」本『抱朴子』の成立について」, 福岡女子短大紀要 6 (1973-03-31), p.29-37, http://ci.nii.ac.jp/els/contents110001151601.pdf?id=ART0001408658

[9] 周祖謨:『問学集』, 中華書局 (1966-01), 下巻, p.760-800.

[10] OpenCV: http://opencv.org/

[11] 昌平坂学問所旧蔵 平津館本説文解字, 内閣文庫・漢籍・叢書部・平津館叢書、請求記号 371-0043、冊次 19-21

[12] 樸學叢書第 2 集 第 1 冊 平津館校刊説文解字・説文検字・補遺, 世界書局(1960) BN12107517

[13] 大本原式精印四部叢刊正編第 4 巻, 岩崎本説文解字, 台湾商務印書館(1979) BN05526666

[14] 藤花榭本説文解字, 早稲田大学古典籍総合データベース http://www.wul.waseda.ac.jp/kotenseki/html/ho04/ho04_00025/index.html

[15] 日照丁氏重校本説文解字, 京都大学人文科学研究所東方學デジタル圖書館 http://kanji.zinbun.kyoto-u.ac.jp/db-machine/toho/html/A025menu.html

[16] 葉德輝:「(説文解字三十巻)又一部 光緒壬午山東丁氏刻本」,『郎園読書志』, 澹園鉛印 (1928), 第 2 巻,葉 41-45.

[17] 王貴元:「《説文解字》版本考述」, 古籍整理研究学刊 (1999 年 第 6 期), p.41-43, p.34

ⓒ 2017 Information Processing Society of Japan