Mem. Fac. Sci. Eng. Shimane Univ.
Series A
35, pp. 91-98 (2001)

前置処理系によるOCR認字率の向上法

曹　　宇*・佐藤　匡正**
*島根大学大学院総合理工学研究科　数理・情報システム学専攻
**島根大学総合理工学部　数理・情報システム学科

An OCR Character Recognition Improvement Method by Preprocessor Approach

Yu CAO and Tadamasa SATOU
**Interdisciplinary Division for Research of Science and Engineering, Graduate School of Shimane University*
***Department of Mathematics and Computer Science, Faculty of Science and Engineering Shimane University*

Abstract

It is convenient to apply OCR to computerizing printed materials as books. Some errors in recognition, however, would be often discovered. They may cause to take not less time for corrections. It is effective to reduce such errors in recognition with covering more range of types used in Nippon-go language. To do so it is tried to develop recognizing ruby types for off-the-shelf OCR's by the preprocessor approach for ease of the development. The preprocessor provides a function by classifying ruby types from regular ones, through checking the area range of a type. This presents the preprocessor approach and its effectiveness.

要　　約

書物など印刷物を電子化するには，OCR文字認識が簡便である．すべての文書に対して正当な認識が行われることが理想であるが，誤認字が生じるため修正には手間がかかる．これまでの経験では，ルビ付きの文書の誤認字が顕著であった．そこで，ルビ文字含む文書の認字率の改善を図ることを企て，ルビOCR機能の実現を前置処理方式によって試みた．本方式は実現の容易さを狙った簡便な方式である．OCR本体はそのままで前置処理系において文字寸法による地文とルビの自動的な識別方法を実現する．本論文では本方式の実現方法と有効性について述べる．

1. 序　　論

OCRによる文章の電子化にあたっては誤認字が問題となる．OCRによる誤認字の問題は，誤認字が正字と類似している字形なので気付き難いとか，ルビに対応していないことなどが挙げられる[1]．これは，日本語文の特徴として，一やーなど仮名(平/片)，漢字，英字，

記号，ルビなど多様な文字種が用いられていることがあげられる．このなかで特に，ルビを含む場合は認字率が低下し[2]，その修正の手間が問題となる．

この解決策として，最も簡単な方法は，ルビに対応しているソフトウェア製品を入手することである．しかし，利用者の立場からすれば，安易な製品購入は戒めるべきである．ルビ機能を求めて製品を購入すれば現行使用製品の廃棄や類似機能の製品の重複購入につながり，金銭上はもちろん，ファイル量等の計算機資源の無駄を招くことになる．別の方法としては，既製品の内部に手を入れて改造を施すことが考えられる．しかし，この方法は著作権の問題や内部構成の知識の獲得方法などが問題となる．

機能改造の方法は，利用者の立場から言えば効果が高く，実現の容易なことが基本要件である．このためには，既製品内部には手を加えずに行えるものが望ましい．そこで，ルビ認字に前置処理方式による機能改造が考えられる．この方法では，一連の処理過程(フェーズ)から成るプログラム処理は，「より以前の結果が前提として行われる」という特性を利用するものである．つまり，既存製品の本質機能を活用するために，入力を製品外部で加工する機能を前置処理として設ける方法である．本方法による開発では，機能の重複開発の問題は残るが，本体には手を加えずに前置処理系だけの開発に限定され，機能改良や追加が容易に行える．

この特徴を生かし，本前置処理方式を既存 OCR にルビ認識機能追加の改造に適用する．ルビを含む文を前処理によってルビ部分と地文部分に分けて，別箇に OCR の入力としてそれぞれに認字する．前置処理では，基本的にルビと地文を自動的に分ける機能を実現する．この場合，技術的にはルビと地文の自動識別方法，およびルビ文の OCR 認字率が問題となる．

ルビは地文と文字寸法が異なる．ルビの文字寸法は対応する地文の 1/2 また 1/2 より小さく，ルビ文字の有効面積は地文の 1/4 また 1/4 より小さい[3]ことが知られている．この結果を利用して，ルビと地文は自動的に分離できる．また，文字寸法は小さなくとも認字率は変わらないことも知られている[3]．これらを利用すれば，前置処理方式で OCR ルビ機能が実現できる．

本稿では，前置処理方式による OCR 機能改造方法の実現性を確かめ，その有効性についての評価を試みる．

2. 前 置 処 理

ルビに対応していない OCR ソフトウェアでは，ルビを地文の字と合体させて認字することが多いため認字率が低下すると考えられる．そこで，ルビ部分と地文を分けて認識させれば，OCR は両方とも通常の文と判断して認字することになる．双方の認字率は通常文の程度が維持されると期待できる．

OCR ソフトウェアの処理方式は，一般に図 1 に示すように，入力部から始まり出力部までの連続した処理部（フェーズ）から成る[4,5]．ソフトウェア本体に，ルビ対応機能を設けるには，行域の識別を行っている本体の前処理において，ルビを認識するための機能を新た

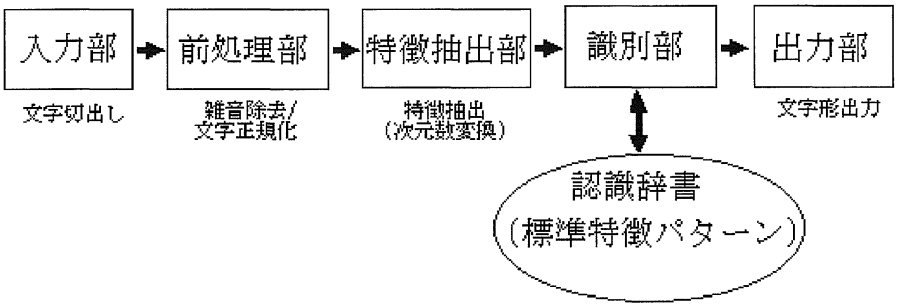

図1　OCR の認字方式

に設ける必要がある．しかし，本機能を外部に前置処理系として設ける方式が考えられる．本体に対する入力を前置処理系で加工してしまい，これをあらためて入力とする．こうすると，外部に設けた前置処理と，本体の前処理は一部処理の冗長構成となるが，前置処理が優先される．前置処理系は処理構成上の冗長性はあるが，本体の認識に対しての影響は与えない．

前置処理系は，システム構成方式として適用性が良好である．既存の OCR ソフトウェアを機能部品としてシステムの中核に据えて，特定の適用に対する機能を前置処理系として実現する．つまり，前置処理系によって前処理部機能の増大が図れる．

本システムでは，スキャナからの出力ビットマップ・ファイルを入力とする．これに基づいてルビと地文との別を判断する前処理を設ける．この処理ではファイルのビット・データを読んで，文字域を確立させ，この大きさによってルビか文字かを自動的に判断する．そして出力として 2 種類のビットマップ・ファイルを作成する．

システムとして見た場合，前置処理は，文字域確定処理およびルビと地文の識別処理をもち，OCR 前処理にも文字域確定処理をもつ．この処理は，本来はシステムとしては，開発上も実行上も単一であるべきで，冗長である．しかし，前置として独立させることができる．

前置処理から 2 種類のファイルを，それぞれ OCR ソフトウェアに入力させることによって，出力として認字した結果である地文認字ファイルおよびルビ認字ファイルが得られる．

3. 実現性と評価

前置処理方式による実現性及びシステム機能，つまり OCR 機能についての有効性を評価する．

3.1 実現性

ここでは，前置処理方式の実現性を確かめる．なお，試料は日本語文庫文試料 1 の10

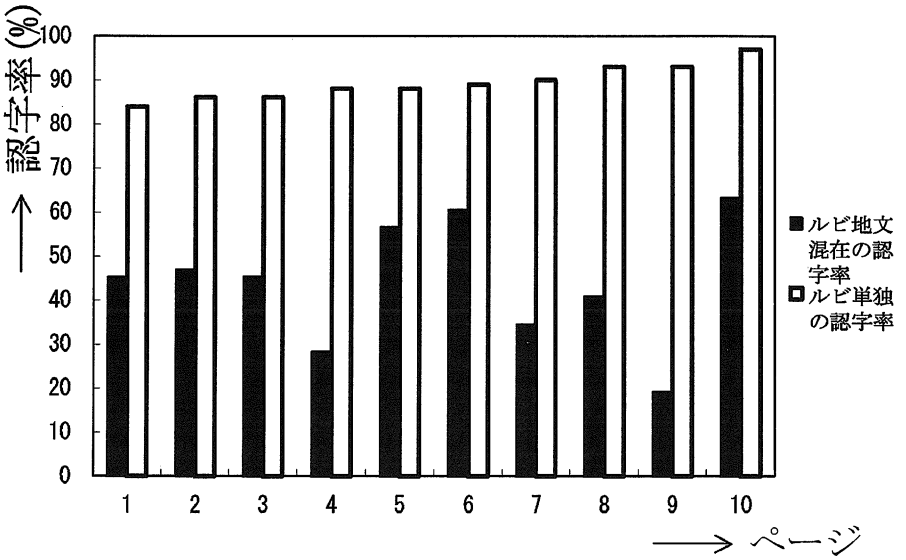

図２　ルビ文字認字率の比較（試料１）

ページ分ルビ文字含む5455字と試料２の30ページ分ルビ文字含む13916字とした．

①　ルビ文字の認字率

試料１と試料２におけるルビ文字と地文混在する場合とルビ単独存在する場合にルビ文字のページごとの認字率を図２，図３に示す．両図ともルビ文字単独存在する場合の認字率によって昇順に整列して示す．

試料１の場合は，ルビ地文混在する場合のルビ文字認字率は19％から63％までの範囲にあり，ルビ文字単独の認字率は83％から97％までの範囲にある．

試料２の場合は，ルビ地文混在する場合のルビ文字認字率は０％から71％までの範囲にあり，ルビ文字単独の認字率は85％から100％までの範囲にある．

②　ルビ付地文の認字率

試料１と試料２におけるルビ文字と地文混在する場合と地文単独存在する場合に地文文字のページごとの認字率を図４，図５に示す．両図とも地文単独存在する場合の認字率によって昇順に整列して示す．

試料１の場合は，ルビ地文混在する場合の地文認字率は11％から80％までの範囲にあり，地文単独の認字率は95％から100％までの範囲にある．

試料２の場合は，ルビ地文混在する場合の地文認字率は4.4％から88％までの範囲にあり，地文単独の認字率は88％から100％までの範囲にある．

---

*) 試料１　荒　俣宏：「帝都物語８」角川書店
　試料２　壺井　栄：「柿の木のある家」偕成社文庫

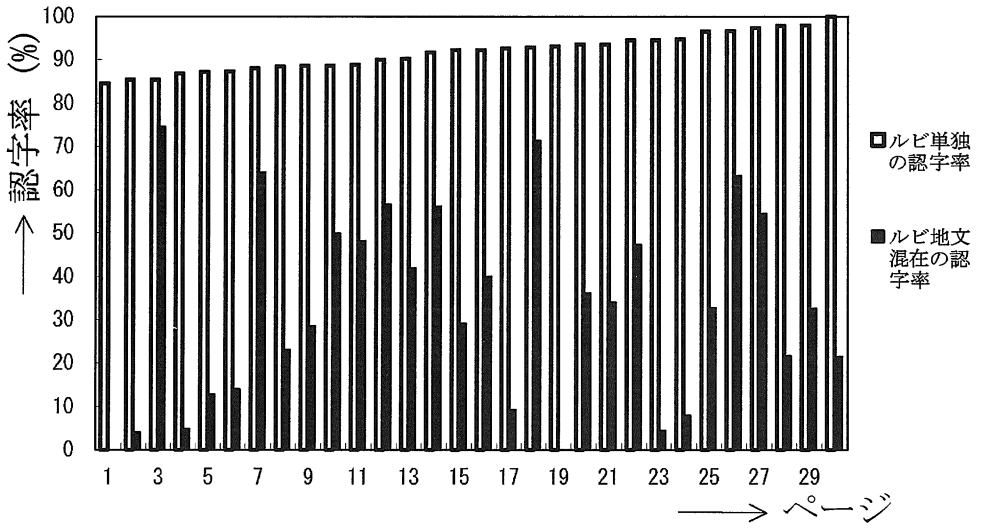

図3　ルビ文字認字率の比較（試料2）

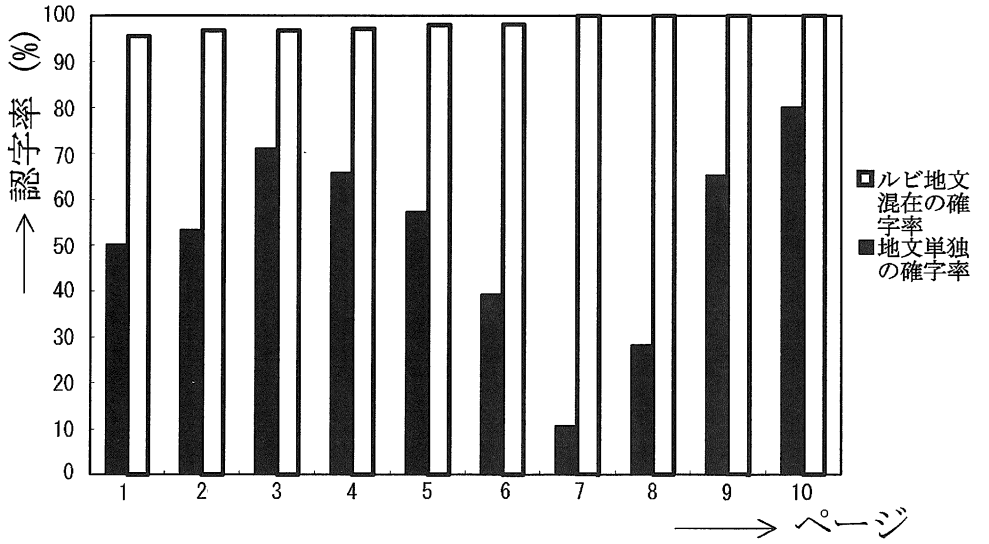

図4　地文認字率の比較（試料1）

以上の観察からルビも地文も本来のOCRの認字率を確保できている．よって，本方式の実現性は有効であると言える．

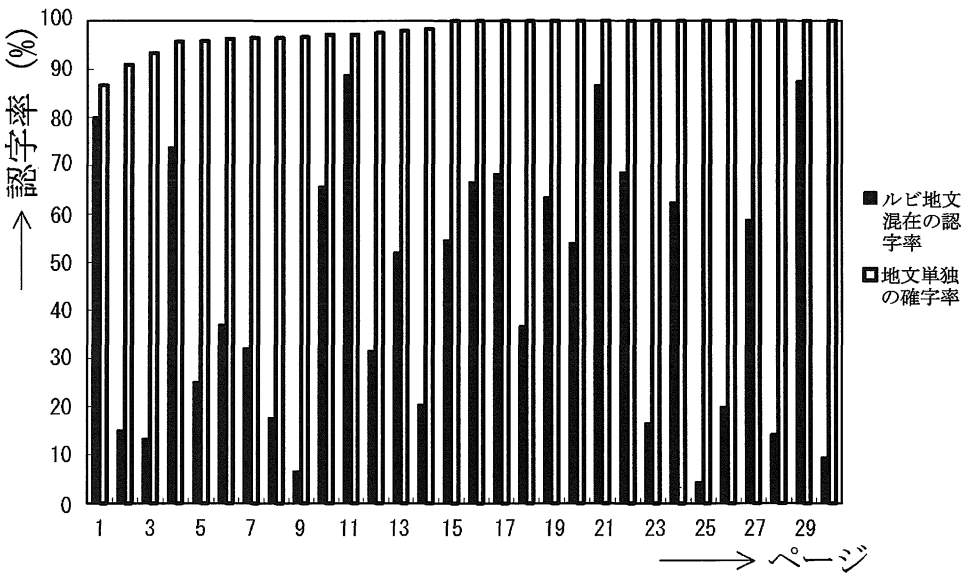

図5 地文認字率の比較（試料2）

3.2 OCR機能の評価

方式の有効性について評価する．

1） 評価の方法

前置処理方式に対する評価の基準は，認字率である．認字率が高ければ高いほど性能が良いものとする．認字率はOCR処理によって，正当に認識された文字数の対象とする全文字数に対する割合であり，次のように定義する．

認字率＝(正当に認識された文字数/対象とする全文字数)×100%

2） 評価

本評価は，両試料を対象にし，通常OCRによる文字認識（ルビと地文混在）と前置処理方式を用い，ルビ文字と地文を分離し（ルビ単独また地文単独），それぞれOCRによる文字認識の認字率を比較する．また，認字の誤り要因について述べる．

(1) 認字率の比較

試料1と試料2のルビ文字と地文が混在とルビ文字，地文単独存在の場合の認字率は表1，表2，表3に示す．

① ルビ文字認字率＝(正当に認識したルビ/全ルビ数)×100%

② 地文認字率＝(正当に認識したルビ付地文/全ルビ付地文)×100%

③ 統合認字率＝(正当に認識した文字数/ルビと地文の総数)×100%

上の結果から，ルビ文字を地文から分離して別箇に認字する前置処理方式は有効であると言える．

表1 ルビ文字認字率の比較

認字率 / 試料	ルビ地文混在の認字率 (%)	ルビ単独の認字率 (%)
試料1	45	90
試料2	30	91

表2 ルビ付地文の認字率の比較

認字率 / 試料	ルビ地文混在の認字率 (%)	地文単独の認字率 (%)
試料1	52	98
試料2	44	98

表3 統合認字率の比較

認字率 / 試料	ルビ地文混在の認字率 (%)	ルビ単独＋地文単独に認字率 (%)
試料1	87	99
試料2	88	99

(2) 誤り要因

評価の結果により，ルビ単独認字する場合と地文単独認字する場合でも，認字率の誤りが生じる．なぜ誤りが起こるか，またどういう誤りであるかについて分析する．誤り要因として，次に例挙する．

① 読み取り上の問題

ルビ文字と地文の行間域の空線に着目するため，試料の読み取り上の傾きによって，ルビ文字と地文とうまく分離できないことがある．つまり，取り込んだデータの回転角度によって認字率が低下することがある．

② 紙面の物理的な要因

紙面の汚れ，皺また印刷する際にインクによる汚れなどが文字認識に影響を与える．ルビ文字の寸法が小さいため，汚れなどルビ文字の一部と認識される場合がある．例えば，ルビ文字「し」のあたりに黒い点が存在すると，「じ」と誤認識されることがある．また，印刷際にインクの漏れによる汚れで，隣のルビ同士とつなぎ，一つの字として誤認識することがある．例えば，縦書きの「よう」を「真」に，「あざ」を「き」に認識される．

③ 解像度

ルビ文字は寸法が小さいため，地文文字より解像度の影響を受ける．解像度が低い場合は分離されたルビがOCRによる文字認識する際に紙面の汚れとして無視されることがある．

表4 「う」,「じ」,「ぼ」の誤認識

ルビ文字	誤認識のパターン
「う」	スノ，入ノ，よノ，Aノ，つ ネノ
「じ」	‥し，いレ，Mレ，んし
「ぼ」	ば，ぽ

また，ルビ文字の構造による誤認識が起こることがある．例えば，濁点つきの「ぼ」,「ぶ」などのようなルビ文字は濁点を無視され,「ほ」,「ふ」に認識される.「く」を「ノ」と「＼」に,「う」を「つ」に認識される.

④ 縦書き文書に対するOCR認識特性

文字寸法の揃った文字においても，縦書き文は誤認識される場合がある．これは，OCR処理の特性と考える．ここでは，ルビ文字を対象にして典型的な例を表4に示す.

4. 結 論

本論文では，文字寸法によってルビと地文を自動的識別する機能をもつ前置処理系によるルビOCR機能の実現方式の提案，実現方法を示し，合わせて試行システムによる実現性，および有効性を確認した.

本機能の追加により，システムのOCR機能の認字率は2～3倍改善され，また90％以上の認字率が得られるようになった.

参 考 文 献

1) 岸本頼紀，曹 宇，佐藤匡正：「OCR文字認識におけるルビ文字の影響」情報処理学会第59回全国大会講演論文集（1999）
2) 佐藤匡正，岸本頼紀：「利用者からみたOCR誤認特性」島根大学紀要シリーズA，33，pp. 51-62（1999）
3) 曹 宇，佐藤匡正：「文字種別による文字寸法の違い」情報処理学会第61回全国大会講演論文集（2000）
4) 飯島泰蔵：「パターン認識理論」基礎情報工学シリーズ6，森北出版株式会社
5) 舟久保登：「パターン認識」情報・電子入門シリーズ11，共立出版株式会社