漢字 OCR システムでの認識率の向上方法と考察
-携帯機器上の問題点と解決方法-

Improvement of the Recognition Rate in the Japanese Kanji OCR System
-Problems and Solutions of OCR Systems on Mobile Devices -

ベロフ　アレクサンドル[*1]・立田ルミ[*2]
Alexander Belov、 Lumi Tatsuta

Email: k02345@dokkyo.ac.jp

近年、無線ネットワークの急激な進歩の影響により、携帯機器の革命が起きている。次世代型携帯電話に続き、スマートホンやタブレットパソコンなどが市場で広く出回り、普及している。その流れではソフトウェアも大きく変化する傾向がある。アップル社をはじめ、Google 社、Microsoft 社など、各大手メーカーがインターネット上の開発センターを公開し、全世界のソフトウェア開発者が互いに新しいアイデアを交換し、語り合う場が実現した。様々なクラウドシステムやオンラインサービスの利用により、各クライアント機器上の計算量が大幅に減ったので、従来、高度周波数 CPU が必要とされる情報処理システムが、携帯機器上でも実現されるようになった。その様なシステムの一つである日本語漢字 OCR システムを用いて、本研究では文字認識率の向上方法について各実験を行い、分析をし、従来のシステムと大きく異なっている機能の作成を試みた。また、携帯機器上での OCR システムの他の問題点とその解決方法について考慮した。以上の点について本稿で報告する。

In recent years, due to rapid advances in wireless networks, is happening revolution of mobile devices. After the next-generation mobile phones became popular and circulate widely in the market such devices as smart phones and tablet computers. And because of this revolution, the current software also tends to change. Some major companies, such as Apple, Google and Microsoft opened their Development Centers on the Internet to software developers for exchange new ideas with each other developers around the world. The use of a variety of cloud systems and similar online services drastically reduced the amount of computation on each client device. Information processing systems that previously required high-frequency CPU now began to be implemented on mobile devices. In addition, we have tried to create functionally different system in comparison with traditional OCR. We also consider how to solve some other problems of the OCR system and on mobile devices. This paper reports on the above points.

*1: 獨協大学経済学部
*2: 獨協大学経済学部

1. はじめに

現在、携帯電話、スマートホン、タブレットパソコンなどが普及し、市場のハードウェアが急激的に携帯化している（図 1 参照）。また、図 2 で示すように、フォレスターリサーチ社[1]の研究結果では、パソコンの市場で携帯機器の利用がさらに広がると予測されている。

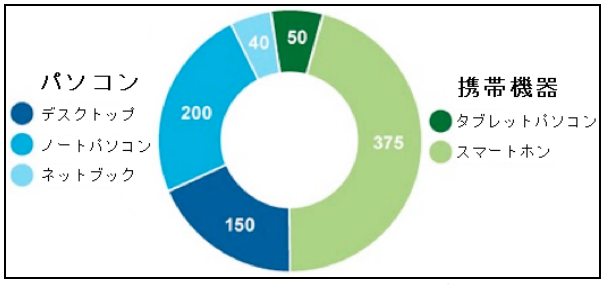

図 1 パソコンと携帯機器の 2011 年度販売台数（単位：100 万台）

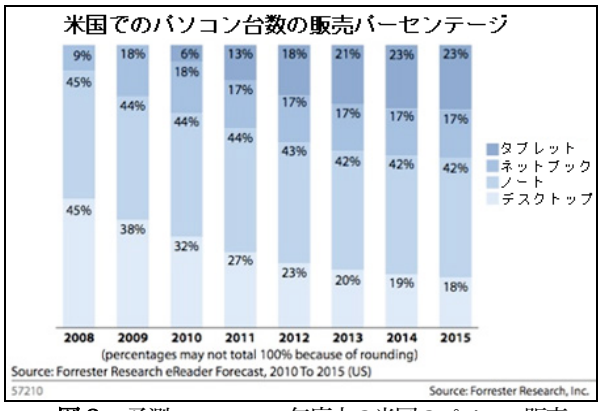

図 2 予測：2008-2015 年度中の米国のパソコン販売のシェア

そのため、ソフトウェアも大きく変化している。クライアントの CPU への負担をかけないで、クライアントから入力データを受け取り、ネットワーク上のサーバでより多くの計算をさせて、結果だけをクライアントへ戻す方法は近年のソフトウェア開発の流れである。その例として各クラウドシステムや多くのオンラインサービスなどがあげられる。「マイクロソフトクラウドオフィス」[2]、「Google 翻訳ツール」[3]、「Yandex スペルチェッカー」[4]など、数多くのオンライン処理ソフトが既に存在している。特に、以前から膨大の計算量でよく知られる情報処理ソフトはオンライン化されてきている。

計算量の大きいシステムの一つは、日本語 OCR（Optical Character Recognition）ソフトである。本稿では携帯用日本語 OCR システム（以下では「本システムと呼ぶ」）の開発とその認識率の向上について述べる。また、携帯機器上の問題点とその解決方法を証明する。

2. 本システムの開発

2.1 本システムのニーズ

現在、大学や図書館などではまだデジタル化されていないテキストが数多く存在している。そして、例えば図書や歴史的な書類処理施設、保管所などでは保管管理によりペーパ上の書類を持ち出すことができない場合が少なくない。それらの施設を利用する研究者、歴史家や記者などは、資料をデジタル化するシステムである本システムのようなシステムがあれば、研究室などで資料を解読することが可能である。モバイル端末からこのようなシステムを閲覧できれば、移動中にでも仕事が可能となる。その他にも多くの携帯 OCR システムのニーズに関する例があると考えられる。

2.2 計算量と認識率のバランス

通常、OCR ソフトでは各文字のデフォルトのマトリクスが使用されている[1,2]。そのマトリクスの各ピクセルのデータを、テキスト上の漢字の領域と思われる画像の一部のピクセルデータと比較している[3]。このマトリクスのサイズが計算量と認識率に大きく影響を与える。確かに、マトリクスのサイズが 8×8 ピクセルの場合、計算量が少なく、認識処理が早く行えるが、認識の正確度が落ちる。他方、元のマトリクスが 32×32 ピクセルの場合、漢字の細かい部分でも正確に認識ができるが計算量が圧倒的に増える。それぞれのマトリクスサイズの例を図 3 参照に示す。

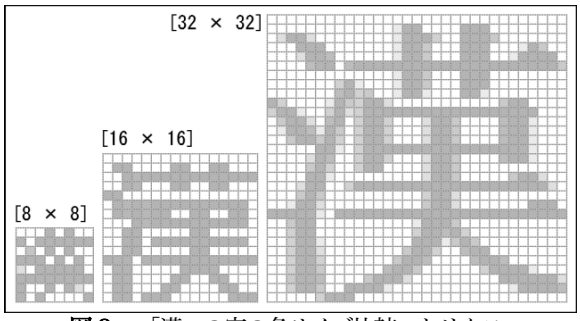

図 3 「漢」の字の各サイズ比較マトリクス

比較マトリクスのサイズを大きくすることにより認識の正確度を向上させられると思われるが、印刷の質によりペーパ上の漢字の位置が数ピクセルにずれることもある。それにより、誤った認識が増える。その場合、逆により小さい比較マトリクスを設けなければいけない。したがって、効率のよい文字認識システムの開発にあたって、認識の正確度と計算量のバランスの調節が必要と思われる。

2.3 本システムの構造

2.2 で提示した問題を考慮し、本システムでは図 3 に示したように、8×8、16×16、32×32 の三つの種類の比較マトリクスを設けた。それぞれのサイズ上の比較結果として、各漢字に対するランキングリストを作成する。

各種ランキングリストのトップに一番多く現れる漢字が最終的な結果として出力される。iPhone 用単漢字ランキングリストの例を図 4 に示す。

3 種類の比較マトリクスを設置することで、図 7 の通り、従来の研究と比較して [1] 認識の正確度があがった。

しかし、通常の OCR と比べて、各漢字の比較のために約 3 倍の計算量がかかった。現在の携帯機器のデータ処理能力を考えると、場合によっては、各文字の認識には数秒の時間がかかる。幸い、無線ネットワークの機能の進歩により、携帯機器上でのサーバとの接続やデータ交換を素早く行えるようになってきているので、処理速度の問題は解決できる。

図 4 iPhone 用漢字ランキングリストの例

日本語オンライン OCR システムの一つである「WeOCR Project」[5]と同様、本システムもインターネット上に認識サーバを備えている(図 5 参照)。よって、クライアントの機器の OS の種類や CPU の処理能力にもかかわらず、短時間の認識処理に成功した [4]。

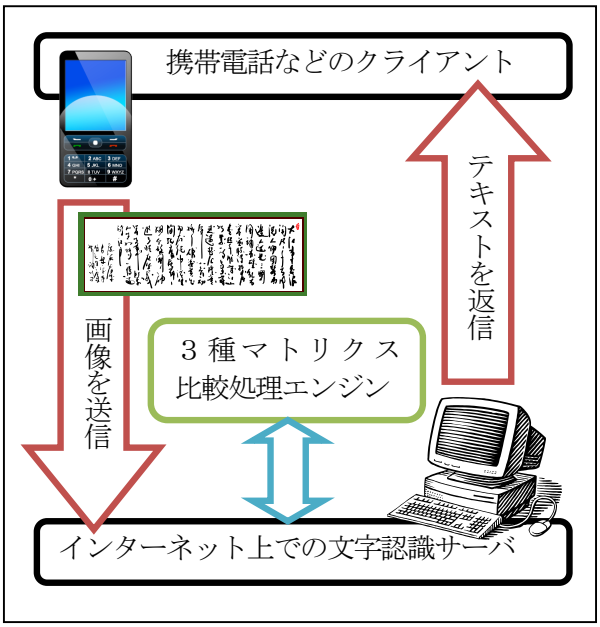

図 5 本システムの構造

2.4 本システムの認識処理

本システムは図 6 に提示した通り、以下の順に認識処理を行っている。

1）文字認識サーバは画像を受け取る。サーバ側は WEB プロトコールを使用しているため、クライアント側の機器の種類に関わらず、JPEG 形式の写真を受け取ることができる。

2）画像を各文字の領域に分ける。まず、画像の上下と左右空白を切り取り、バックグラウンド色の直線で貫通できるポイントを探す。それにより、全体的のテキスト各行に分ける。基本的に日本の漢字が四角い領域に当てはまるため、行の高さと比べながら、各文字の領域を取得する。日本のテキストでは「弓」と「引」のような字が存在しているため、各文字の切り方が重要な役割を果たしている [1,5]。

3）各漢字の領域のピクセルデータを各比較マトリックスのサイズに合わせて圧縮し、メモリスタンプを作る。その時、スタンプのサイズが小さいほど漢字の元の形が崩れるため、画数の多い漢字の場合、大きいマトリックスの使用により、認識率があがる。しかし、画数の少ない漢字の場合、画像データでの上下左右への数ピクセルずれを避けるため、数ピクセルの列を合体させる小さいスタンプの使用により認識率が上がる。

4）各文字のピクセルデータを三種類のマトリクスと比較し、各ランキングリストを作成する。比較マトリクスの表面データだけではなく、各ピクセルの「深さ」も計算される。本研究では 1 バイト、0 から 255 までのピクセルの「深さ」を備えている。人間の目でその「深さ」のバイトを確認しやすくす

るため、図３では灰色を備え、その明るさでデータレベルを示した。

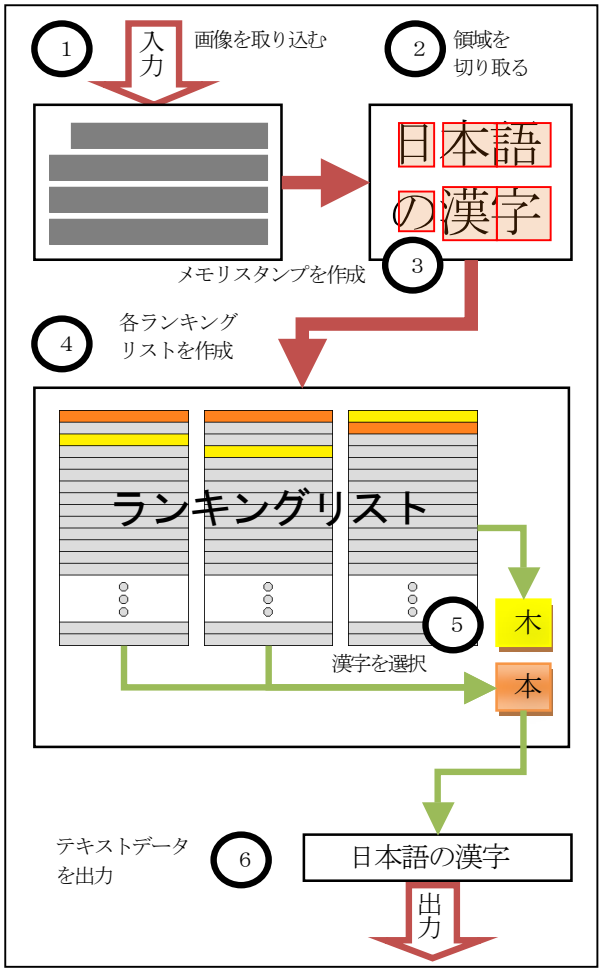

図６　本システムの認識処理

５）各ランキングリストを分析し、トップに一番多く現れた漢字を出力テキストへ挿入する。

６）結果テキストをクライアントへ送る。WEB プロトコールではユニコードテキストデータもサポートされているため、結果としてプレーンテキストを送ることができる。よって、クライアント機器の種類とそのオペレーティングシステムに関係がなく、文字化けしないでテキスト表示が可能になっている。

3.　実験とその結果

本システムを用いて認識処理実験を行った。iOS、Android、Windows Phone の各種 OS 上で動作している携帯電話を使用し、本システムの同テキストの認識処理にかかる時間を計った。その時間は、100 文字あたりで約 35 秒であり、これらの OS 間で大きい差は見られなかった。

また、3 つの比較マトリクスを単独でとその組み合わせを用いて、漢字の認識率の比較実験を行った。

その結果の平均値データを図７に示す。

実験に使用した入力素材として 4 つの新聞の記事の写真を携帯電話のカメラで撮影した。同時に、認識正確度を計るため、同テキストをキーボードから入力し、テキストデータを備えた。それぞれの比較マトリクスのソフトの認識データをテキストデータと比較しながら、漢字の画数と認識正確度の関係の平均データを作成した。そのデータをまとめて、グラフ化した。

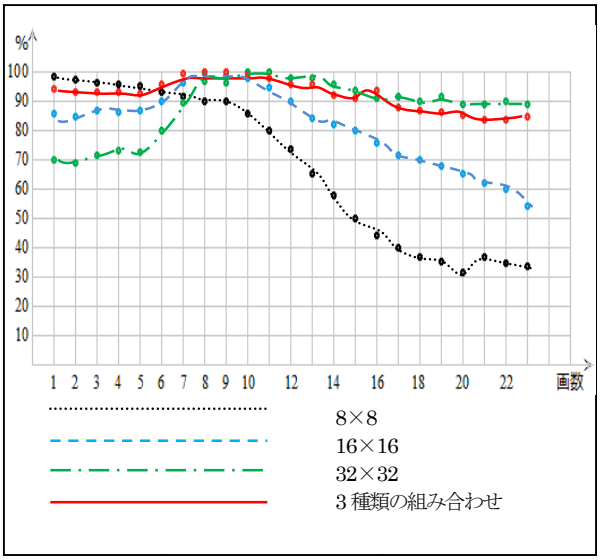

図７　各種マトリクスの認識率比較実験の結果

実験の結果を分析すると、漢字の画数の少ない場合、小さい比較マトリクスを用いた比較処理エンジンの仕様により認識率が上がる。画数の大きい漢字の場合、小さい比較マトリクスを用いた比較処理エンジンの仕様により認識率が下がる。逆に、大きい比較マトリクスを使用しているシステムの場合、反対の結果が見られる。本システムの特徴である、三つの種類の比較マトリクスを備えることにより、漢字の画数に依存しない、全体的に認識率の高い日本語 OCR システムの作成にほぼ成功したと思われる。

4.　今後の課題

今回の実験で使用した比較マトリクスの種類は３つだけである。実験として備えたテキスト数が 1452 文字と少ないため、より正確なデータを得るためにそれぞれの数を増やすことが必要と思われる。

また、将来に期待される本システムを利用するユーザーの増加を考慮し、サーバ側の処理アルゴリズムを同時進行化し、疑似混雑実験を行うことが必要となる。

5.　おわりに

本研究の結果は以下の通りまとめられる。

1)　現在の携帯機器が互換性のないさまざまな OS

とハードウェアを備えているため、互換性のあるインターネットプロトコール上での認識サーバにより、文字認識処理が効率よく行われる。

2) 文字認識率を向上させるため、複数の比較マットリクスとその各リストをバランスよく設けることが効果的であると思われる。

謝辞

実験とデータ処理に際して、協力いただいた学生や友人の皆様に厚く御礼申し上げる。

参考文献

（1）ベロフ・アレクサンドル :"ワンタッチエンジンの構造と解析アルゴリズム"、獨協大学情報センター「情報科学研究」、第 22 号 pp.1-12、(2004)

（2）中野 康明、花野井 歳弘、丸山 稔、宮尾 秀俊、丸山 健一:"複数の文書理解システムを用いた文書理解の高度化(文字とドキュメントの認識・理解)"、 電子情報通信学会技術研究報告、PRMU, パターン認識・メディア理解 103(659), pp.55-60, (2004.2)

（3）熊谷 勝彦、鈴木 真一、上野 浩司 :"OCR の認識率アップ法とそのシステムの簡素化"、全国大会講演論文集 第 42 回平成 3 年前期(2)、pp.104-105、(1991.2)

（4）齋藤 靖二、後藤 英昭、小林 広明 :"シーン中の文字領域検出における周波数特徴の分析と比較"、電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 104(523) 、pp.31-36、(2004.12)

（5）能隅 進一、 福田 亮治、玉利 文和、 鈴木 昌和 :"絞り込み法による数式文字認識とその日本語/数式領域切出しへの応用"、 電子情報通信学会論文誌、D-II、情報・システム、II-パターン処理 J83-D-II(3) 、pp.895-906、(2000.03)

参考 URL

[1] フォレスターリサーチ社
http://www.forrester.com/rb/research
[2] マイクロソフトクラウドオフィス
http://www.microsoft.com/ja-jp/office365/online-software.aspx
[3] Google 翻訳ツール
http://www.google.co.jp/language_tools
[4] Yandex スペルチェッカー
http://api.yandex.ru/speller/
[5] WeOCR Project
http://weocr.ocrgrid.org/

(2011 年 9 月 30 日受付)
(2011 年 12 月 21 日採録)