オフライン草書体文字認識システムにおける文字抽出の自動化

西田　茂生・谷　宗一郎

Automatization of the character extraction in an off-line fully-cursive-style character recognition system

Shigeki NISHIDA , Soichiro TANI

An off-line cursive style body character has two main features. One is that the character is connected by one line. Another is that a background is complicated by the quality of paper. Moreover it is that the shades of character differ and also existence of blur, lacking, etc. makes character recognition environment severe. Although the character recognition engine has up to this time been developed, the extraction part of the character which is needed in the stage before character recognizing has been performed manually. This paper reports automatization of this character extraction part. Two-step processing is performed in automatization of character extraction. The first phase decomposes a character image into a string using the difference of pixel power. The second phase is extraction of one character and this is decomposed using variance of pixel power. The character extraction experiment was carried out using the built character extract program by making applicable to an experiment the character image written to the Japanese writing paper with the brush. As a result, the extraction strike rate of 92.8 percent was obtained. Moreover, as a result of carrying out an extraction experiment using the old literature in which a 31-syllable Japanese poem was written, it succeeded in extracting one character at a time.

1．緒言

文字認識の分野において，これまでに様々なタイプの認識システムが提案され実用化されている．しかしながら，和紙に墨書された文字を認識するものは未だ研究段階である．本研究では，過去の先人たちが残してきた貴重な文献を国文学や歴史学に携わっていない者にでも簡単に読むことのできる，オフラインでの草書体文字認識システムを提案してきた．[1] オフライン草書体文字を認識するための要件は，不定形文字の認識と続き文字列の分離である．古文献中の文字は様々な字体を含み，個々の書き手の癖が多く現れるため，書き手によって認識率が影響されない手法が必要となる．そこで，先ず第1段階として，不定形文字の認識のために，認識エンジンには学習機能を持ち，データベースの文字数を増加させることによって認識率向上が見込まれる隠れマルコフモデルを用いた．[2) 3)]

オフラインの草書体文字の認識で問題となる点は，一つは字形が不定形であり文字間を1本の線でつなげて書かれていることであり，一つは文献が古く保存状態が悪くなるにつれ文字のかすれや欠損がおこることである．これまでの認識システムでは，半紙に墨書されたオフライン文字画像を4値化することにより，文字のかすれやつづき部に対応させたが，1文字を抽出するためには手動で行う文字の抽出作業が必要であり，そのために認識時間の高速化を妨げていた．[4)]

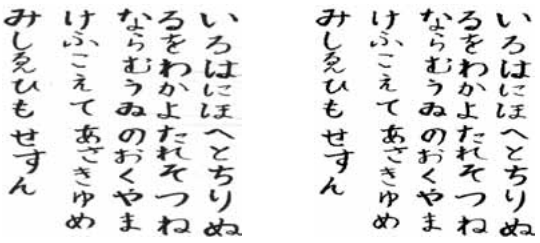
本論文では，和紙に墨書された仮名文字の文書画像から，1文字ずつを自動で抽出する手法を提案する．文字の抽出実験および認識実験対象としては，17名が半紙に墨書した“いろはうた”を用い，文字抽出成功率と認識率を求めることにより，本手法の妥当性の評価を行った．また，文字抽出に関しては和歌の書かれた古文書による文字抽出実験を行い，古典文献の自動認識システムの可能性評価も行った．

2．文字の抽出法

文字の抽出手法について以下に説明する．なお，文字画像は1画素あたり8ビットのグレースケール値で表現し，最も黒い部分を0，白い部分を255とした．また，文字画像中の各文字は手書きの墨書であるため，任意の文字幅と高さをもつ．

2.1　背景除去

認識対象として，和紙に複数の文字が墨書された画像を入力画像とする．この入力画像に対して，画像の中に存在する濃度の低い画素を背景，つまり画像中に存在する紙の柄やしわなどを文字ではなく背景と見なし，これを除去する．背景であるという判定は，一定の濃度以下の画素を白すなわち画素の値を255に置き換えることで行う．この際，残った画素に対して読みとりやすくするために濃度を強調する操作を行う．これは各画素の濃度の値に対して一定の係数をかけ，その値が255を超えるものについては，255に置き換えるという方法で行う．また，画像内において一定矩形内に収まり，周りを白の画素に囲まれた白でない画素の集合もノイズと見なして除去する．これによりシミなどの文字でない成分を除去することができる．このように処理して得られた画像を加工画像とする．Fig.1に入力画像と加工画像の一例を示す．Fig.1は半紙に筆で墨書された“いろはうた”である．図(a)では一部に紙のしわ後が見られるが，処理後の(b)では除去させていることがわかる．



(a)Original image　(b)Background eliminated image
Fig.1　Image samples for background elimination

2.2　文字列の抽出

Fig.1のように認識対象は文字の大きさが不均一であり，かつ行間も不均一である．この画像から一行一行を切り出す文字列の抽出を行う．加工画像において，その左端から画像の画素の列方向について濃度の分散を求める．求まった分散の値について差分を求め，差分結果の各値と加工画像の列要素を対応付ける．差分結果が一定値以下となる列を空列，すなわち文字の書かれていない列と見なす．これは，文字の書かれている画素の列と文字の書かれていない列で，分散が大きく変わることを利用している．文字の書かれている画素の列は，その文字の存在のために画素の濃淡が大きく移り変わり，分散の値が大きくなることによって差分値も大きくなる．一方，文字の書かれていない画素の列はほぼ白一色であるため，分散の値は小さくなるため差分値も小さい．この特性より，それらの差分が文字の書かれていない列の区間で小さく，文字の書かれている区間で大きくなる．

この操作により，空列でない列の区間を文字の列，空列と空列でない列が切り替わる列を列区切り位置とし，それによって加工画像を区切る．区切られた各画像を部分画像とする．

2.3　1文字の抽出

文字の列が抽出できれば，次に文字列から1文字ずつを抽出する操作を行う．部分画像において，その上端から画像の行方向について濃度の平均値を求める．求まった平均値の列に対して，次式に示す三角関数を用いた最小二乗法を適用する．なお，最小二乗法の次数は部分画像の幅と上下の空白部を除いた高さから推定した部分画像に含まれる文字数から計算した値を用いる．

$$f(x)=a_n \sin x+a_{n-1}\cos x+a_{n-2}\sin 2x+a_{n-3}\cos 3x +\cdots+a_0 \qquad \text{eq.1}$$

ここで，aは係数，nは次数を示す．

このようにして得られた係数を用いて，平均値に対する近似曲線を得る．この近似曲線の各極大の点は，文字の区切りの基準となる．得られた近似曲線と，部分画像の対応をFig.2に示す．図中の黒線は最小二乗法で得られた曲線，灰線は近似曲線である．なお，図の右側の文字列は，文字列抽出により得られた部分画像を示す．

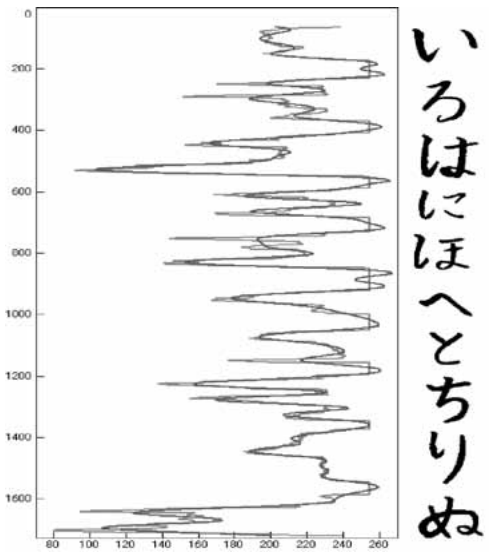


Fig.2 An average power of one line of an extracted character string image and an approximated curve by the method of least square using a character string shown in right-side hand
The black curve expresses the average power and the grey curve expresses the approximated curve.

得られた近似曲線を用い，各部分画像の上端の行から下端の行までの範囲内で次の⑴から⑶の操作を繰り返すことによって，1文字の抽出を行う．なお“操作行”とは，操作を行うべき対象の行のことを指す．

⑴文字の高さの推定

多くの文字は正方形に近い四角形に収まることから，列の幅を文字の高さであると仮定する．Fig.3に文字の高さの推定方法の概略図を示す．列の中での若干の文字の大きさの変化に対応するために，操作行付近の一定の行からなる領域について，その左右に存在する空白を除いた幅を求める．その値と列の幅の重み付き平均を求め，その値を文字の高さとする．図中の細実線は列の幅，点線は操作行付近の幅，太実線は求められた文字の幅および高さを示す．

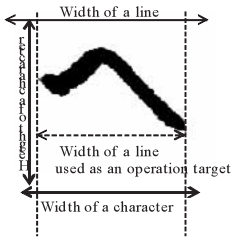


Fig.3 Schematic diagram of evaluation of character height

⑵仮の区切り位置の推定

操作行から推定した高さ分だけの領域を，文字の範囲であると仮定し，領域下端を仮の区切り位置と推定する．

⑶文字領域の決定と操作行の移動

仮の区切り位置に対応する近似曲線の点において，その点付近にある極大の点のうち，隣り合う極大と極小の差が少ないものを除外した点前後にある直近の極大の点を以下に示す条件で選択し，文字の区切りとして文字領域を記録する．ただし，その区切り位置の行の画素の濃度が一定以下であるならば，極狭い範囲において一定以上の濃度をもつ行へと，区切り位置を修正する．

・極大点選択の条件

より近い方の画素が白と見なせるならばその極大の点を選択し，さもなければより近くにある極大点を選択する．

区切り位置を次の文字の開始位置へと設定する．ただしこのとき，区切り位置の下方向に余白であると見なせる領域があるならば，その領域下端を次の文字の開始位置とする．区切り位置の修正と操作行の移動方法の概略図をFig.4に示す．図中の点線部分は補正前の文字領域，実線部分は補正後の文字領域を示し，補正終了後次の文字区切りへ移動する．なお，余白とは，その行の濃度の平均値と分散値が一定範囲にあるものである．こうして得られた区切り位置と，文字の列の区切り位置から文字が収まっている範囲である文字領域を記録する．

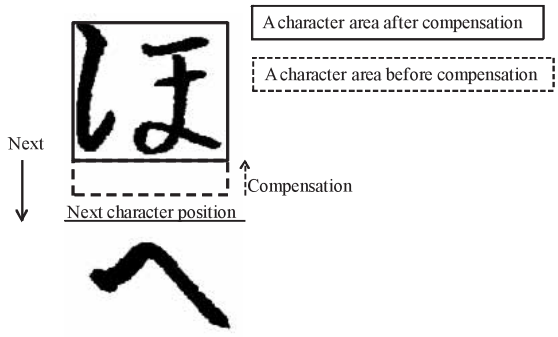


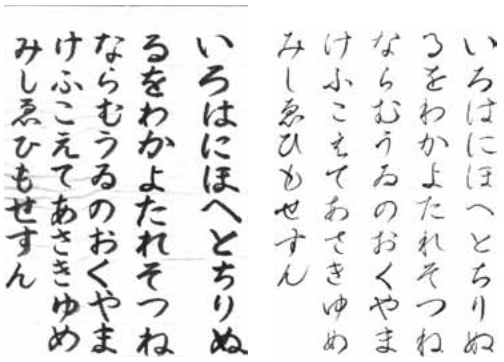
Fig.4 Schematic diagram of compensation method of character area and a method of movement to next operation target

最後に記録された文字領域を用いて，入力画像から文字画像を抽出してファイルとして出力する．このとき，文字領域の上下左右の各辺から，最も近い白でない画素が存在する行または列を調べ，それぞれ差をとることにより文字の幅と高さを得る．この幅と高さを比較して，小さい方にこの2つの値が同じ値になるように空白領域を追加する．空白領域の追加は文字が出力結果の画像の中心にくるように，上下または左右に均等に追加する．

3．文字抽出実験

3.1　文字抽出プログラムの評価

2章で提案した，文字抽出プログラムの検証実験をおこなった．文字画像として，同一文字列，ここでは“いろはうた”を使用し，これらを異なる17名が半紙に毛筆で墨書したものを用いた． Fig.5に文字抽出実験に使用した文字画像の例を示す．



(a) Object A　　(b) Object B

Fig.5 Image samples for a character extraction experiment

図のように画像の一部に文字以外の紙の汚れやしわなどが存在する他，文字の濃さが個人で異なり，線の太さや濃淡が書き手によって異なる様子がわかる．

文字抽出実験の結果をFig.6，抽出に成功した文字の内訳をFig.7に示す． Fig.6では横軸に被験者番号，縦軸に抽出成功率をとり，被験者ごとのいろは48文字に対する抽

出成功率を示す．また，Fig.7では横軸に抽出対象文字すなわち，いろは48文字，縦軸に抽出成功率をとり，各文字での抽出成功率を示す．

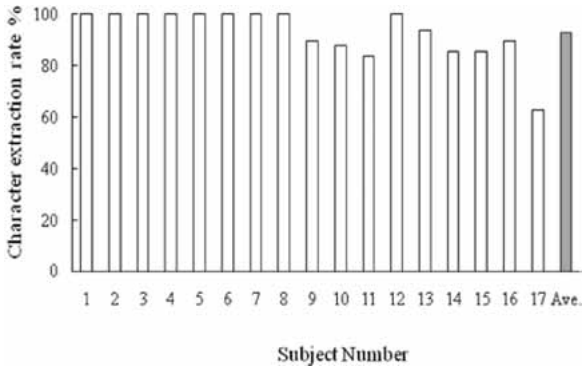


Fig.6 Result of character extraction experiment
The number of the characters for an extraction experiment is forty eight of “いろはうた”.
"Ave." expresses the average character extraction rate for all the subjects.

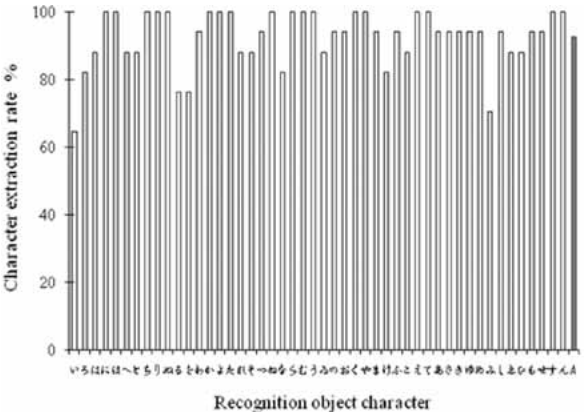


Fig.7 Character extraction rate for each character
“A” expresses the average character extraction rate for all characters.

Fig.6より，書き手により62.5%の成功率はあるもののほとんどの被験者では背景を含む文字でも抽出できており，全文字のうち抽出に成功した文字の割合は92.8%であった．このことより本文字抽出方法が書き手や筆記環境によらず高い文字抽出率を得ることのできる手法であることが検証された．

さらに文字抽出プログラムにより出力された文字画像を観察する．抽出に成功した文字画像の一部をFig.8，抽出に失敗した文字画像の一部をFig.9に示す．

Fig.8をみると，文字に汚れやしわによる影などの背景が含まれている場合や，文字の濃淡および滲みやかすりが含まれている場合でも抽出処理に影響がないことがわかる．すなわち本手法が背景を含む画像からの文字の抽出に対して有効な方法であることを示している．

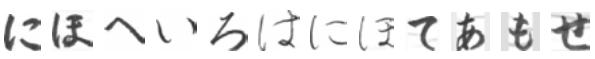


Fig.8 Success examples of character extraction

Fig.9 Failure examples of character extraction

Fig.7およびFig.9の抽出失敗例をみると，「ろ」「る」「を」などの文字の途中で行の濃度が大きく変わるものが主になっている．またFig.9のように文字の列のはじめの文字で失敗しているものも多い．これは，画像上端に含まれる濃い影が，背景として除去し切れていないためであると推測される．また，ある一つの文字の抽出に失敗すると，その付近の文字も失敗である場合が多い．これは抽出の失敗が抽出位置の極の選択の失敗であり，次回抽出位置の選択の失敗であるためであると思われる．この極の選択の失敗の原因としては，現在文字抽出プログラム内で固定値として設定されている各値を，入力画像から推定して設定できるようにすることや，2.4で述べた極大点選択の条件を修正することで改善できると考える．

3.2　古文献の文字抽出実験

次に古文書から文字抽出を行う実験を行った．実験対象文字画像には和歌の書かれた色紙を用いこれをFig.10に示す．図をみるとわかるように，この画像には仮名以外に変態仮名，漢字，落款が混在したものである．抽出さ

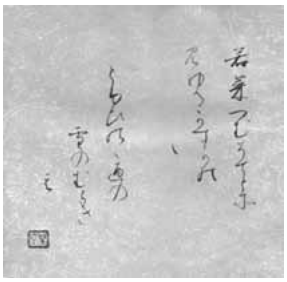


Fig.10 An image sample for a character extraction experiment
A 31-syllable Japanese poem written to Japanese writing paper

Fig.11 Extracted string lines by the character extraction program

Fig.12 Extracted characters of the forth line

れた行をFig.11，第4行の文字をFig.12に示す．

Fig.11，12より和紙に墨書された和歌の文字が1文字ずつ抽出されていることが確認でき，しかも変体仮名などを含んでいても文字抽出が可能であることが確認できた．

4．認識エンジンの改良

4.1.　データベース作成プログラムの改良

本文字認識システムの認識エンジンでは，拡張性の高い隠れマルコフモデルをベースとして構築している．しかし，これまでに構築したデータベース作成プログラムでは，作成の際に文字ごとにプログラムを用意し，さらにデータベースの元となる画像ごとに処理を記述していた．このため，処理が複数あることによるプログラム容量が肥大化していた．また，1つの処理を実行するのに広範囲を確認する必要があることや，ある処理に変更を加えるとすべてのプログラムを確認して変更をしなければならないという可読性・保守性の低さ，データベースの文字を1つ増やすごとに処理を増やさなければならないという拡張性の低さが問題となっていた．このため隠れマルコフモデルの利点が有効に使えていない状況であった．

そこで，まずプログラム容量の肥大化を解消するために，分散していた各画像についての処理をループ処理の中に置き，このループを各文字を網羅するループ処理の中に入れ子構造にして，分散されていた処理を単一のプログラム，単一の処理記述にまとめた．また，文字ごとに作成していたデータベースファイルを単一のファイルにまとめることで，データベースを取り扱いやすく改良した．これらのプログラムの改良によって，現状の認識処理プログラムは以前と比較して，容量を約1/350に抑えることができた．次に，拡張性の低さを解消するために，いわゆるマジックナンバーとなっていた，画素の濃淡の量子化ビット数を定数から変数として置き換え，その値の将来の変更に際して，処理を変更せず数値の変更だけで対応可能な処理記述を行った．また，データベースの元となる画像のディレクトリ構成を，基準となるディレクトリの下にデータベースの各文字に割り振った一意の番号のディレクトリを用意し，その中にそれぞれの元となる画像を置くように変更した．さらに，プログラム中に画像のファイル名をフルパスで記述せずに基準となるディレクトリの指定のみを記述するように変更した．

4.2 文字認識プログラム

これまでに作成した文字認識プログラムでは，上述のデータベース作成プログラムと同様の問題のほか，結果の出力が各文字に割り振った一意の番号であったため，結果がわかりにくいという欠点があった．これらの問題を解消するために，データベース作成プログラムと同様の修正を加えたほか，文字の表を作成することにより出力を各文字に変換するようにした．また，データベース作成プログラムと同一の処理が存在したため，それを関数化してどちらのプログラムからも同一の処理記述を使うようにした．その結果，処理変更の際の書き換える手間を減らすことができた．これらの改良のほかに，さらに処理の効率化による認識実行速度の高速化を図った．これは冗長な繰り返し処理を行わないようにすることや，使用プログラミング言語の特性上，行列演算を高速化できることから，単純なスカラ演算を行列演算に置き換えるなどを行った．その結果，これまでと同じ環境で文字認識プログラムを実行したところ，旧文字認識プログラムでは1文字あたりの処理時間が約3秒であったのが，改良後のプログラムでは，約1.5秒となり，約2倍の高速化が実現できた．

5. 文字認識実験

3章で用いた画像と同一のものを対象として，4章で述べた認識システムを用いて認識実験を行い，出力文字列を正しい文字列と比較する．ただし，比較の際に抽出の失敗などによって1つの文字が2つに，2つの文字が1つに抽出された場合などの出力は，その失敗した文字らを無条件で認識失敗とする．また，その文字の位置に正しく文字が抽出されていた場合に，そこにあるべき文字数だけの文字があったものとして，次の文字の比較を行う．

文字認識実験の結果をFig.13に示す．また，認識に成功した文字の内訳をFig.14に示す．

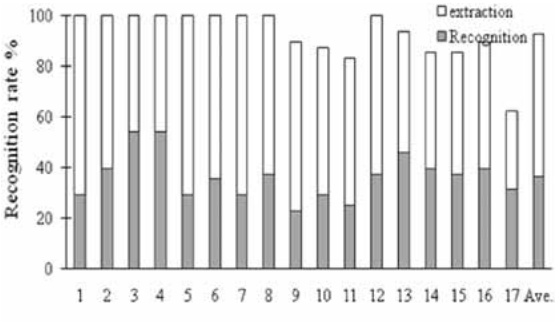


Fig.13 Result of recognition experiment with all subjects

A white rod expresses a character extraction rate and a solid rod expresses a character recognition rate. "Ave." expresses the average character extraction rate and the character recognition rate for all the subjects.

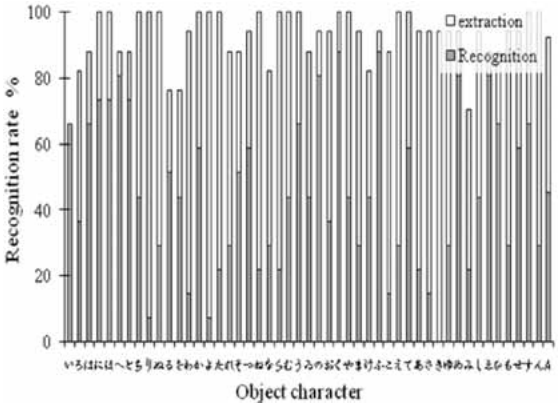


Fig.14 Result of recognition experiment with all characters

A white rod expresses a character extraction rate and a solid rod expresses a character recognition rate. "A" expresses the average character extraction rate and the character recognition rate for all characters.

Fig.13では横軸に被験者番号，縦軸に抽出成功率および認識成功率をとり，文字抽出実験結果に認識結果を重ねたものである．また，Fig.14では横軸に抽出対象文字すなわち，いろは48文字，縦軸に抽出成功率および認識成功率をとり，文字抽出実験結果に認識結果を重ねたものである．ともに白抜き棒が文字抽出成功率，塗りつぶし棒が文字認識率を表わす．

Fig.13,14を見ると，全体の認識率が36.3%と，これまでの認識プログラムで得られている87.7%と比較すると，著しく低い．また，文字抽出率100%の被験者が必ずしも認識率が高いわけではなく，文字抽出率の高い文字種が必ずしも高い認識率ではない．Fig.15に文字種における抽出率と認識率の関係を示す．この図は横軸に抽出率，縦軸に認識率をとり，それらの相関を示している．

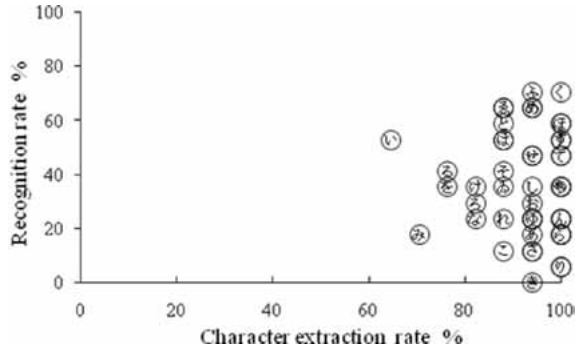


Fig.15 Relation between character extraction rates in a character type and character recognition rates

図からわかることは文字種において抽出率と認識率の間に相関がないことである．この理由としては認識エンジン側に原因があり，使用しているデータベースが毛筆文字を含んでいないことが大きく影響している．すなわち，文字の濃淡に関するデータが少ないことと字形の違いが認識率を低下させていると考えられる．

次にどの文字で認識できないかを考えるため，Fig.14を変更したものをFig.16に示す．

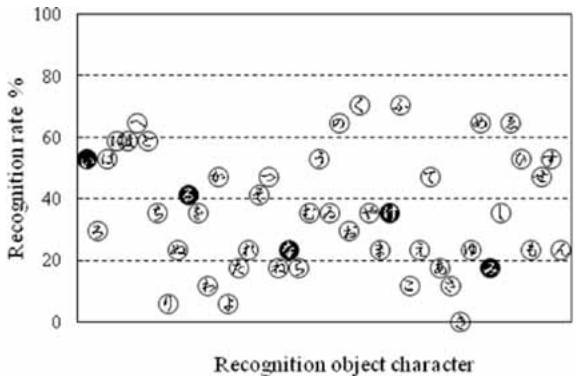


Fig.16 Result of recognition experiment with all characters

この図は文字ごとの認識率をプロットしたもので，上方にあるものほど認識率が高いことを示す．また，塗りつぶしの点は「い」「る」「な」「け」「み」で行頭文字を表す．認識率40%で2群に分けると行頭文字は低認識率群に属するものが多い．また，「わ」「れ」「ね」「お」「あ」や「ろ」「ら」「る」，「さ」「き」などの字形の似たものが低認識率群に属し，「り」「よ」「ま」「こ」「し」「も」などの縦長の文字が低認識率群に属している．これらの原因も文字抽出失敗の原因とは関係なく，データベースにあると考える．

認識率を改善するためには，第一に毛筆で書かれた文字をデータベースに追加することで，改善が見込まれる．なぜなら，毛筆文字をデータベースに追加することで，文字の濃淡や字形のバランスなどが認識要因として追加され，隠れマルコフモデルの特徴である学習機能が有効に発揮されると考えられるからである．その他の方法としては，認識時の画像サイズを大きくすることや，画素濃度のビット数をより多くしたりすることが考えられる．

5．結言

本論文では，複数の毛筆で墨書された仮名画像を認識可能にするために，画像に含まれる文字を1文字ごとに自動抽出する手法を提案し，そのプログラムを作成した．

その結果，改善すべき点があるものの，毛筆の楷書体の抽出において92.8%の妥当な抽出ができる手法を実現することができた．また，実際に和歌のかかれた古草書体の古文献から文字抽出が可能であることも示した．

次に文字抽出プログラムを認識エンジンに組み込むことにより，文書を自動認識するシステムを構築した．その結果は平均36.3%と低い認識率となったが，その原因がデータベースであることを確認した．今後の改良により，画像として取り込まれた古典文献を人の手を介さずに文字認識を行うシステムを構築できる可能性を示した．

参考文献

1）オフライン草書体文字認識システムの開発；西田，辻野奈良工業高等専門学校研究紀要40(2004)39-44
2）Hidden Markov Models combination framework for handwriting recognition; T.Artieres,N.Gauthier,et al., A, IJDAR,No.5 233-243(2002)
3）確率的言語モデル；北研二；東京大学出版会1999
4）隠れマルコフモデルを用いたオフライン手書き文字の認識；近澤，西田；精密工学会　2006，3.18　講演