オフライン草書体文字認識システムの開発

西田　茂生　　辻野　学

Development of an off-line cursive style body characters recognition system

Shigeki NISHIDA and Manabu TUJINO

There are two significant procedures to recognize off-line cursive style characters. One is recognition of the infinite form characters as separate of the character connected. Another is recognition of each separated character. In this paper, we propose an algorithm of off-line cursive style body characters using hidden Markov model (HMM). A characters databank was constructed using researcher's characters. To begin with the each character image is reduced to 50×50 pixels, it makes to four grey scale values. Next it has expanded as it is compared with the length and breadth of the character in the image. Finally the HMM parameter of each character was calculated. In this experiment, four sets of characters were used. The result of this, the recognition rate has become 43.5%. However, there was so little individual difference. It is thought that it becomes a system that can deal with everyone, and it is possible to correspond to various documents.

1．緒　　言

文字認識技術において，楷書体による定型文字やオンライン手書き文字などの認識技術はすでに実用化されているが，オフライン手書き文字については実験段階にあるものが多く，さらに日本文学の古典に多く用いられるオフライン草書体文字の認識技術は開発されていない．その理由として草書体文字は字形が不定形であり，文字間を１本の線で書くという特徴をもつこと，古い文献では文字の擦れや欠損が多くこれらが草書体文字の認識を困難にしている．オフライン草書体文字認識システム技術が開発されれば専門的な知識や訓練の必要なく草書体で書かれた文献を読解することができ，古典の研究に役立つと考えられる．本研究ではオフライン草書体文字認識システムの開発を最終目標とする．草書体文字を認識する場合は２つの重要なプロセスが必要となる．１つは文字間の分離であり，もう１つは不定形文字の認識である．そこで本論文ではその第１段階としてオフライン不定形文字認識手法の提案を行う．不定形文字のオフライン認識手法として，システムの発展性を考え学習機能をもち，データベースを増やすほど認識率が上がると考えられる隠れマルコフモデル（以下HMM）を用いた．また不定形文字としてデータベースには筆者の手書き文字，認識実験用には４名の被験者の手書き文字を用い実験を用いた．

2．認識アルゴリズム

2.1　HMMの概要

ＨMMとは確率的な状態遷移と記号出力を備えたオートマトンである．以下にHMMの概要を述べる．

まずHMMは以下５項目により定義される．

⑴　$Q=\{q_1, \cdots, q_N\}$：状態の有限集合

⑵　$\Sigma=\{o_1, \cdots, o_M\}$：出力記号の有限集合

⑶　$A=\{a_{ij}\}$：状態遷移確率分布

a_{ij}は状態q_iからq_jへの遷移確率であり$\Sigma a_{ij}=1$

⑷　B=$\{bi(\mathrm{ot})\}$：記号出力確率分布

$b_i(o_t)$は状態q_iで記号o_tを出力する確率であり，$\Sigma b_i(o_t)=1$

⑸　$\pi=\{\pi_i\}$；初期状態確率分布

π_iは状態q_iが初期状態である確率$P(X_1=q_i)$

４つの状態からなるHMMの例をFig.1に示す[1)]．

Fig.1は状態q_1～q_4があり，その状態から記号o_1～o_4を出力することを表している．円の中にあるπ_1～π_4は添

字番号の状態からはじまる初期状態確率である．

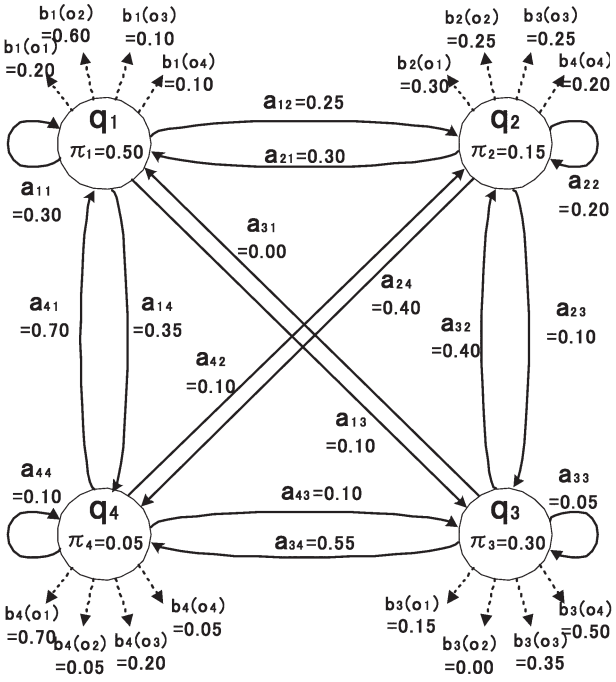


Fig.1　4つの状態にからなるHMM

矢印は状態の遷移を表しており a_{11}～a_{44} はその際の状態遷移確率である．そして状態から記号 o_1～o_4 が出力される記号出力確率が $b_1(o_1)$～$b_4(o_4)$ となる．例えば状態が1→1→2→4→3（q の添え字を表す）と遷移するとき，記号が3→2→4→2→1（o の添え字を表す）と出力されるとすると生成確率Pは次式で求めることができる．

$$\begin{aligned}P&=\pi_1\times b_1(o_3)\times a_{11}\times b_1(o_2)\times a_{12}\times b_2(o_4)\times a_{24}\times b_4(o_2)\times a_{43}\times b_3(o_1)\\&=0.50\times0.10\times0.30\times0.60\times0.25\times0.20\times0.40\times0.05\times0.10\times0.15\\&=1.35\times10^{-7}\end{aligned}\tag{1}$$

したがって，生成確率は 1.35×10^{-7} となる．また記号出力するとき状態遷移が数種類ある場合は，各生成確率の和がその記号出力の生成確率となる．

2.2　HMMの学習アルゴリズム

データベースを作成する際与えられた記号系列からモデルのパラメータを決定しなければならない．しかしHMMでは記号系列を生成した状態遷移系列は非観測であるため，直接最尤推定を行うことができない．このためEMアルゴリズムを用いた繰り返しアルゴリズムによりパラメータを推定する．

最初にモデル M が $O_1^T=O_1\cdots O_T$ を生成して，時刻 t で状態 q_i に到達する確率 $\alpha_t(i)$ を導入する．

$$\alpha_t(i)=P(o_1^t,X_t=q_i\mid M)\tag{2}$$

$\alpha(\cdot)$ を前向き確率と呼び次のように再帰的に計算できる．

$$\alpha_{t+1}(j)=\left[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right]b_j(o_{t+1})\tag{3}$$

次に前向き確率とは双対な考えである後向き確率を導入する．後向き確率 $\beta_t(i)$ は状態 q_i から始まる状態遷移によって $o_{t+1}^T=o_{t+1}\cdots o_T$ が生成される確率である．

$$\beta_t(i)=P(o_{t+1}^T,X_t=q_i\mid M)\tag{4}$$

後向き確率も前向き確率と同様に再帰的に計算できる．

$$\beta_t(j)=\sum_{j=1}^{N}a_{ij}b_j(o_{t+1})\beta_{t+1}(j)\tag{5}$$

そして与えられた記号系列 $O_1^T=O_1\cdots O_T$ に対し状態 q_i から状態 q_j への遷移が時刻 t で生じた確率 $\gamma_t(i,j)$ を考える．

$$\begin{aligned}\gamma_t(i,j)&=P(X_t=q_i,X_{t+1}=q_j\mid o_1^T,M)\\&=\frac{P(X_t=q_i,X_{t+1}=q_j,o_1^T\mid M)}{P(o_1^T\mid M)}\\&=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\alpha_T(i)}\end{aligned}\tag{6}$$

また $\gamma_t(i)$ を以下のように定義する．

$$\gamma_t(i)=\sum_{j=1}^{N}\gamma_t(i,j)\tag{7}$$

$\gamma_t(i)$ は，時刻 t に状態 i に滞在した確率である．

$\gamma_t(i,j)$ および $\gamma_t(i)$ を用いてパラメータの再推定を以下のように行うことができる．

$$\bar{\pi}_i=\gamma_1(i)\tag{8}$$

$$\begin{aligned}\bar{a}_{ij}&=\frac{\text{状態}i\text{から状態}j\text{へ遷移する回数の期待値}}{\text{状態}i\text{から遷移する回数の期待値}}\\&=\frac{\sum_{t:o_t=k}^{T-1}\gamma_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}\end{aligned}\tag{9}$$

$$\begin{aligned}\bar{b}_i(k)&=\frac{\text{状態}i\text{に滞在し記号}k\text{を出力する回数の期待値}}{\text{状態}i\text{に滞在する回数の期待値}}\\&=\frac{\sum_{t:o_t=k}^{T}\gamma_t(i)}{\sum_{t=1}^{T}\gamma_t(i)}\end{aligned}\tag{10}$$

上の再推定式を用いてパラメータを求める方法を本研究の学習アルゴリズムとして使用する．

2.3　HMMの文字認識への応用

今回の実験では1画素を1つの状態とし背景と同色（白）の場合，記号出力は無しとした．また入力画像を2値化ではなく，背景を除いて4値化し文字の濃淡にも

対応できるよう考慮した．入力画像をMとして4値化に用いた式(11)を以下に示す．

$$\left.\begin{array}{l}(medianM)/4\times3-0.1\le o_1<(medianM)/4\times4-0.1\\(medianM)/4\times2-0.1\le o_2<(medianM)/4\times3-0.1\\(medianM)/4-0.1\le o_3<(medianM)/4\times2-0.1\\0\le o_4<(medianM)/4-0.1\end{array}\right\}\quad(11)$$

ここで$median$は中央値を意味する．画像の大部分が背景であることから，中央値と背景の値は近似していると考え，誤差を0.1に設定し背景を除去した．残った値で4値化を行う．その時の色調をFig.2に示す．

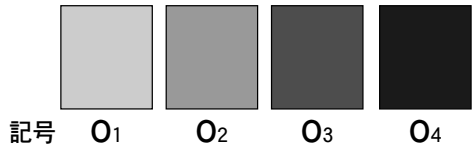


Fig.2　4値化の色調

これらの記号o_1～o_4を出力する場合の状態q_i，初期状態確率π_i，状態遷移確率a_{ij}，記号出力確率$b_i(o_t)$をFig.3に示すように設定した．

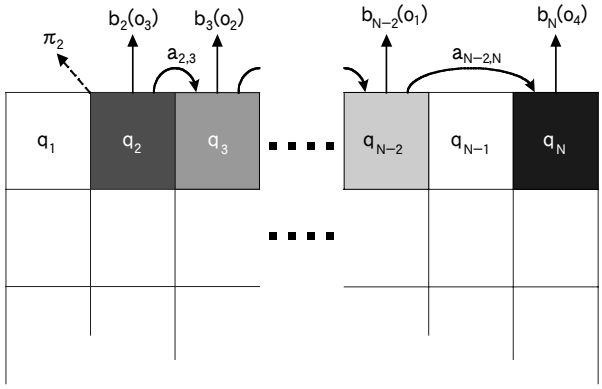


Fig.3　パラメータの設定方法

まずHMMパラメータを1行ごとに定義し，計算を行った．画素の左端を状態q_1，右端を状態q_Nとし左から右にむかって状態が遷移すると定める．その時，最初の記号を状態iで発見した場合に初期状態確率π_iに1を代入し，その時の記号出力確率$b_i(o_t)$に1を代入する．次に記号が出力されている状態jへ向かい状態遷移確率a_{ij}に1を代入する．後は初期状態確率を除いて同様の手順で記号出力確率，状態遷移確率に1を代入していく．これで1枚の画像各行にHMMを定義することができる．この方法でデータベースを作成する場合，1文字につきL枚画像を取り込んだとする．そしてデータベースのπ_i, a_{ij}, $b_i(o_t)$を$(\pi_i)_D$, $(a_{ij})_D$, $\{bi(o_t)\}_D$，取り込んだK枚目の画像のπ_i, a_{ij}, $b_i(o_t)$を$(\pi_i)K$, $(a_{ij})K$, $\{b_i(o_t)\}K$と表して書くと，式(12)により各パラメータを求めることができる．

$$\left.\begin{array}{l}(\pi_i)_D=\{(\pi_i)_1+(\pi_i)_2+\cdots+(\pi_i)_K+\cdots+(\pi_i)_{L-1}+(\pi_i)_L\}/L\\(a_{ij})_D=\{(a_{ij})_1+(a_{ij})_2+\cdots(a_{ij})_K+\cdots+(a_{ij})_{L-1}+(a_{ij})_L\}/L\\\{b_i(o_t)\}_D=[\{b_i(o_t)\}_1+\{b_i(o_t)\}_2+\cdots\{b_i(o_t)\}_K+\cdots+\{b_i(o_t)\}_{L-1}+\{b_i(o_t)\}_L]/L\end{array}\right\}\quad(12)$$

本来であれば式(12)を用いるが，今回のHMMの状態・記号出力設定の場合，このまま後に示す式(15)を計算すると全ての値が少数であり，似ている文字ほどデータベースに含まれるパラメータ数が多いことから，式(15)の計算結果は小さくなる．しかし，似ている文字ほどパラメータ値は大きいので，パラメータ数が同じであると式(15)の計算値は大きくなる．このように計算値の大小と認識したい文字が一致しづらいため式(13)を用いる．

$$\left.\begin{array}{l}(\pi_i)_D=\{(\pi_i)_1+(\pi_i)_2+\cdots+(\pi_i)_K+\cdots+(\pi_i)_{L-1}+(\pi_i)_L\}\\(a_{ij})_D=\{(a_{ij})_1+(a_{ij})_2+\cdots(a_{ij})_K+\cdots+(a_{ij})_{L-1}+(a_{ij})_L\}\\\{b_i(o_t)\}_D=[\{b_i(o_t)\}_1+\{b_i(o_t)\}_2+\cdots\{b_i(o_t)\}_K+\cdots+\{b_i(o_t)\}_{L-1}+\{b_i(o_t)\}_L]\end{array}\right\}\quad(13)$$

式(13)は，画像を取り込んだ枚数で割らず，HMMパラメータを全て整数で扱うようにした．パラメータ数が多い場合でも式(15)の計算結果が大きくなるので，計算値と認識したい文字の不一致を大幅に改善することができた．この式(13)を用い平仮名46文字のデータベースを作成する．ここまでの作業が終了したら認識実験を行う．まず実験画像をFig.3の方法でパラメータ化し作成した平仮名46字のデータベースとひとつひとつ重ね合わせる．その時用いた式で入力画像をMとし，π_i, a_{ij}, $b_i(o_t)$を(π_i)M, $(a_{ij})M$, $\{b_i(o_t)\}$Mと表してものを式(14)に示す．

$$\left.\begin{array}{l}\pi_i=(\pi_i)_M\times(\pi_i)_D\\a_{ij}=(a_{ij})_M\times(a_{ij})_D\\b_i(o_t)=\{b_i(o_t)\}_M\times\{b_i(o_t)\}_D\end{array}\right\}\quad(14)$$

これにより，実験画像に対する平仮名46文字の各HMMパラメータを求めることができる．

後は生成パラメータを式(15)によって計算する．

$$P=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{4}\pi_i a_{ij}b_i(o_t)\quad(15)$$

ただしπ_i, a_{ij}, $b_i(o_t)$の中に0が含まれている場合，乗算であるため他の値に関係なく結果が0になるので，0を飛ばして計算している．またHMMパラメータは各行ごとに計算されるので実験画像の行数の数だけ計算される．

3　データベース

3.1　データベースの作成

データベースの作成手順をFig.4に示し，各手順の処理方法を述べる．

（i）データベース画像を取り込み平仮名各46文字の手書き文字のデータベースを作成する．600dpiのスキャ

ナで文字をパソコンに8ビットのグレースケールで取り込む．1文字につき10個とりこむ．文字は研究者のものを用いた．

(ii) 手作業により一文字ずつ切り出す．

(iii) 取り込んだ画像を50×50ピクセルになるようにサイズを調整する．これは計算機のメモリ容量に制限があるために行っている．「あ」を例にあげFig.5(a)に示す．これらの画像はデータベース作成のために取り込んだ研究者の文字画像である．この時点では画像は8ビットのグレースケールで表されている．

(iv) 式(11)を用いて背景，ノイズを除去し4値化した画像をFig.5(b)に示す．この時点で画像の各画素の光強度は2ビットに減少するが文字の濃淡の情報が残っていることを確認することができる．しかし文字ごとに大きさのばらつきがあるため次の作業を行う．

(v) Fig.5(b)で画像内の文字を縦横比は保持したまま50×50ピクセルに拡大しFig.5(c)に示す．文字の左右の余白は計算により均等になるように調節している．この作業は文字の大小による個人差をなくすためである．

(vi) Fig.5(c)の画像を式(13)によりパラメータ化し「あ」のデータベースが完成する．同様に平仮名46字のHMMパラメータを設定する．このとき1文字につき取り込む文字数を増やせばより緻密なHMMパラメータを設定できる．

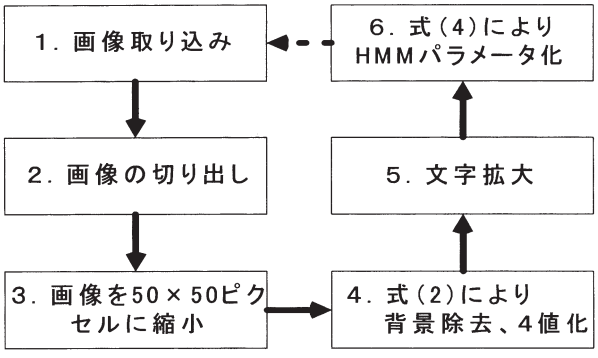
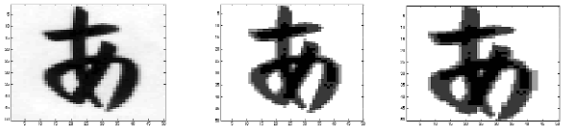


Fig.4　データベース作成手順



(a)取り込み画像　(b)4値化後画像　(c)拡大後画像

Fig.5 データベース文字画像「あ」の一例

3.2　データベース内文字の認識実験

認識実験手順をFig.6に示し各手順の処理方法を述べる．

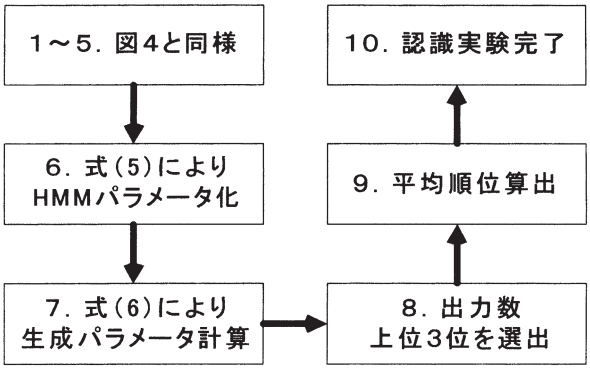


Fig.6　認識実験手順

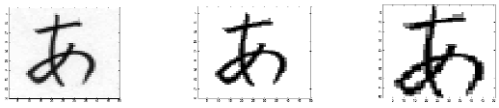
(i) 認識対象画像を8ビットのグレースケールで取り込む．

(ii) 手作業により一文字ずつ切り出す．

(iii) 取り込んだ画像を50×50ピクセルになるようにサイズを調整する．調整後の文字画像をFig.7(a)に示す．

(iv) 式(11)を用いて背景，ノイズを除去し4値化する．処理後の文字画像をFig.7(b)に示す．

(v) Fig.10で画像内の文字を縦横比は保持したまま50×50ピクセルに拡大しFig.7(c)に示す．



(a)取り込み直後　(b)4値化画像　(c)拡大後画像

Fig.7 認識実験に用いた文字「あ」の一例

(vi) (c)をパラメータ化し式(14)によりデータベース46字と重ね合わせ(c)に対する平仮名46文字の各HMMパラメータを求める．

(vii) HMMパラメータより式(15)を用いて46文字の生成パラメータを計算する．

(viii) 各文字につき生成パラメータが0以外になる時の出力数をカウントする．その結果をFig.8に示す．この結果から出力数の1位「あ」，2位「え」，3位「ん」となった．

(ix) この3文字の生成パラメータを計算し，カウント数と式(16)で示す平均順位を算出したものをTable1に示す．

$$\text{平均順位} = \frac{\text{カウント数順位} + \text{生成パラメータ順位}}{2} \quad (16)$$

(x) 表からわかるようにこの認識実験結果から認識対象文字は「あ」を示しており，文字認識が行えたといえる．なお，データベース内の文字による認識実験結果は認識率100％であった．

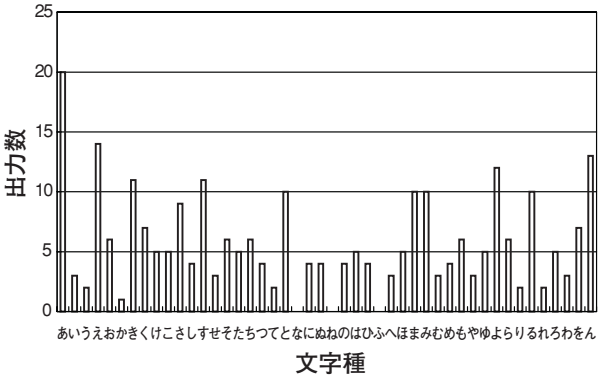


Fig.8　各文字のカウント数（生成パラメータ 0 を除く）

Table 1　平均順位算出

	カウント数	生成パラメータ	平均順位
あ	20	2.07×10^{10}	1.5
え	14	9.36×10^{9}	2.5
ん	13	5.03×10^{10}	2

4．認識実験

4.1　認識実験結果

前章の実験例で示したような実験を，Fig.9 に示す被験者 4 人（データベース作成者と他の被験者 3 名）が手書きした平仮名46 文字を用いて認識実験を行った．

(a)研究者　(b)被験者Ａ　(c)被験者Ｂ　(d)被験者Ｃ

Fig.9 認識実験に使用した文字

この図以外にも濁点や半濁点を付け加えることができる文字があるが，本研究はデータベースにある文字種しか認識できないため，データベース内の文字を選んで認識させる必要がある．認識実験結果を Table2 に示す．表の認識率は式(17)で定義したものを用いる．

認識率(%)=(認識文字数／合計文字数)×100　(17)

認識率は全体で 43.5 % と実際に文字認識システムとして使用するには厳しい数字となってしまった．特にデータベース本人の認識率が予想に反し低く，逆に他の被験者と同じような認識率になった．この事実は本研究で提案する HMM を用いた認識手法が筆記者を選ばない不特定筆記者の文字認識の可能性を示していると考えられる．

Table2　認識実験結果（認識率）

	認識（文字種）	認識不可（文字種）	認識率（%）
研究者	20	26	43.5
被験者 A	19	28	41.3
被験者 B	21	25	45.7
被験者 C	20	26	43.5
平均	20.0	26.3	43.5

4.2　考察

次に本認識手法による認識率が低い原因を考察する．まず 4 人の認識結果から各文字の認識数を Fig.10 に示す．

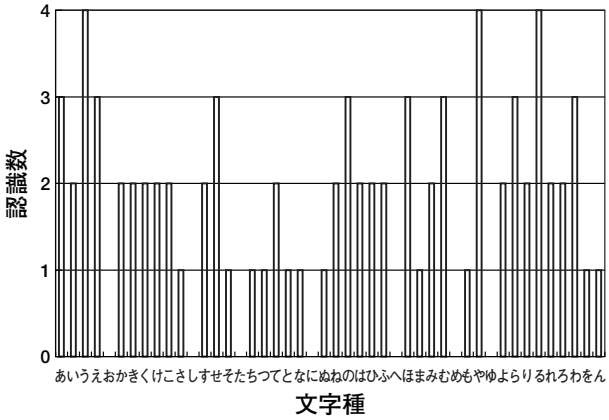
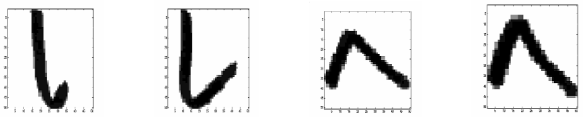


Fig.10　各文字の認識数（被験者 4 名の総数）

図より「う」「や」「る」に関しては全員認識できているが，「お」「し」「た」「に」「へ」「め」「ゆ」は全員認識できなかった．認識できなかった文字について考察する．「し」「へ」に関しては文字の縦横比のばらつきにあると思われる．データベースに用いた「し」「へ」について幾つかを Fig.11 に示す．



(a)し；縦長　(b)し；横長　(c)へ；横長　(d)へ；縦長

Fig.11　縦横比のばらつき

このように文字の形にばらつきが出やすいため，データベースのHMMパラメータも低くなり式(5)により求められる生成パラメータも小さくなる．そのために認識されづらいと考えられる．対策としてはデータベースの拡大などがあると考えられる．

「お」「に」「た」の関しては，形の似ている文字の誤認識がほとんどであった．「お」は「む」と間違えやすく，「に」であれば「け」「ほ」「は」のどれかと間違えていた．また「た」では「む」「ん」と間違えている．
現在のHMM設定では背景を記号出力に含まずに設定していて，文字部だけを評価する手法をとっている．そのため，字形の似ている部分しか評価できずこのような誤認識が起こると考えられる．対策としては字形の違う部分も評価するために背景（白）も記号出力に設定し，画像全体を評価することで解決できると考える．

「め」「ゆ」に関しては上記に示したものの複合的な理由であると考える．対策方法も上記によると思われる．またTable3に上位選出数と非選出数，選出率を示す．

Table3　認識実験結果（選出率）

	上位選出（文字種）	非選出（文字種）	選出率(認識率)（%）
研究者	28	18	60.9（43.5）
被験者A	27	19	58.7（41.3）
被験者B	28	18	60.9（45.7）
被験者C	26	20	56.5（43.5）
平均	27.3	18.8	59.3（43.5）

選出率の算出は式(18)による．

$$選出率(\%)=(上位選出数／合計文字数)\times 100 \quad (18)$$

この表より選出率と認識率の差が15.8％もあることがわかった．選出されても認識文字と一致しなかった文字が多かったことから，新たな判定基準を設け，この差をうめなければならない．

5．結　　言

本研究では，オフライン草書体文字認識システムの構築を行うため，その第1段階として，HMMを用いた認識アルゴリズムを提案した．認識率は全体で43.5%と決して高い数字ではないが，個人差がほとんどなかった．このことから改良を加え認識率を上げれば不特定筆記者に対応できる草書体文字認識システムの可能性が示された．

参考文献

1) T.Artieres,N.Gauthier,et al.,A Hidden Markov Models combination framework for handwriting recognition, IJDAR,No.5 233-243(2002)
2) 北研二，確率的言語モデル，1999，東京大学出版会