文字認識のための漢字データベースの解析
An analysis of Kanji Character database for character recognition

北村浩治　二階堂真理恵　中島由美　安田道夫
Koji Kitamura Marie Nikaido Yumi Nakashima Michio Yasuda

要旨

手書き文字認識システムの開発や評価には，対象となる文字に対する手書き文字データベースが必要不可欠である．日本ではひらがな，カタカナ，漢字，数字，アルファベット等の多種多様な文字が日常的に使用されている．

海外に目を向けると，NIST Special Database 19 のように大量のデータを提供しているデータベースも存在するが，そこで扱われるデータの範囲は，国や地域によって異なるため，数字やアルファベットの文字認識システムの利用に限定される．

したがって，日本語の文字を対象とする認識システムを開発し，評価するには，これらの多種多様な文字列を扱った日本語のデータベースが必要である．

公開されており，良く知られた日本語手書き文字データベースとしては，ELT データベースと JEITA-HP データベースとが挙げられる．

実際，これらの 2 つの手書き文字データベース以外にも，文字認識システムを開発している企業では，各社独自で作成した作成しているデータベースがあると思われるが，一般には公開されてない．そのため，大学等の独自にデータベースを持たないところで，日本語のデータを大量に扱うためには，ETL もしくは JEITA-HP のデータベースを利用するしかないのが実状である．

ところが，どちらのデータベースも作成されてから随分と経つが，よく利用される ETL と比較すると，JEITA-HP はあまり利用されていないようである．本稿では，今後の利用に向けて，JEITA-JP データベースの解析を行った．

1　はじめに

文字認識を行うためにコンピュータへデータを取り込むには，スキャナやカメラなどによって書かれた文字を画像として取り込む方法と，ペンデバイスやタッチパネルなどを使って筆跡，筆順の位置情報，ベクトル情報をリアルリアルタイムに取り込む方法とに大別することができる．

前者は，事前に書かれている文字を光学装置を用いて文字認識を行うことから，光学文字認

識 (Optical Character Recognition : OCR) または「オフライン文字認識」と呼ばる．後者はリアルタイム情報を扱うことから「オンライン文字認識」と呼ばれる．

デジタルカメラやタッチパネル機能をもつスマートフォンやタブレット PC が広く普及する現在，文字認識の機能の利用は，従来の郵便区分機のような大規模な業務向けだけでなく，日常生活においても身近なものとなっている．

また近年，ウェアラブルコンピュータまたはウェアラブル端末と呼ばれる身につけて持ち歩くデバイスのひとつとして，google グラスのようなアイウェア型デバイスが登場してきた．アイウェア型デバイスは，今はまだあまり普及していないが，このような装置を用いて，我々の視野に入った画像情報から文字情報を自動で読み取るようなことも近い将来，実現するかもしれない．

文字認識の手法は，現在に至るまで様々なアルゴリズムが提案されている．このところ，ビッグデータが注目を浴びるようになって，機械学習 (machine learning) がもてはやされており，機械学習のひとつであるサポートベクターマシン (Support Vector Machine:SVM) などの手法を使って文字認識 [2] も行われている．筆者らは，文字認識の手法として，相関法の改良，類似度法による参照パターンの選択法，疑似三次元特徴抽出法などを提案してきた．[1]

文字認識システムの開発において，ニューラルネットワークや SVM 法では，教師データによって多数のパラメータを最適化し，相関法による文字認識手法では，個別文字を代表する標準パターンをデータから作成する．教師データからのパラメータの決定，標準パターンの作成には，サンプルデータが必要であり，このデータの質や量によって，文字認識システムの性質は大きく変わる．また，特にオフライン文字認識では，評価のための文字データが大量に必要であり，大量のデータでなければ十分な評価をすることできない．大量のサンプルデータを海外のデータベースを利用して，認識システムを作成，評価することもできるが，その評価が行えるカテゴリは数字やアルファベットに限られる．

このように，文字認識システムを実現するためには，入力デバイス，認識手法，文字データベースが必要であるが，整備された日本語の文字 (ひらがな，カタカナ，漢字を含む) の手書き文字データベースは，公開されているものはあまり多くなく，不足しているとも言えなくはない．

文字認識システムを開発している各企業では，独自のデータベースを作成していると思われるが，これらのデータベースは各社の企業秘密であり，一般に公開されていない．

ETL 手書き文字データベースは最も良く利用されているデータベースである．漢字に限ってみると，ETL8 では 160(文字/字種)，ETL9 では 200(文字/字種) が格納されている．

今回，解析を行った JEITA-HP データベースは，“DATASET_A” で 480(文字/字種)，“DATASET_B” で 100(文字/字種) が格納されている．

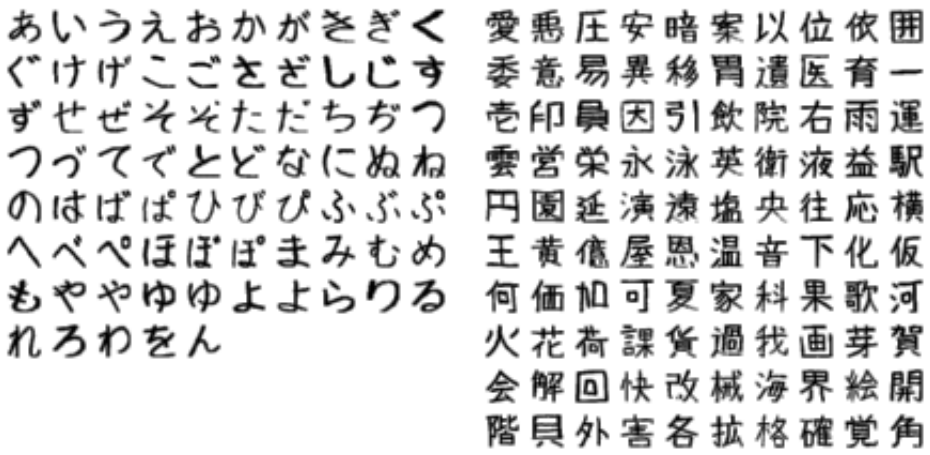

図 1 ETL8 の手書き文字データの例

2 ETL 手書き文字データベース

日本語手書き文字データベースで良く利用されているものは， ETL データベースである．ETL データベースは現在の産業技術総合研究所 (National Institute of Advanced Industrial Science and Technology) の前身である電子技術総合研究所 (Electro-Technical Laboratory:ETL) から提供されている．データ収集は，1973 年から 1984 年にかけて収集されてものであり，あまり新しくはない．

ETL データベースは，ETL1 から ETL9 までで構成されており，このうち，手書き文字として良く利用されているデータベースは ETL6，ETL8，ETL9 である．

ETL6 は，対象が数字，英大文字 26 字種，カタカナ 46 字種，特殊文字 32 字種で 1383 文字ずつ格納されてる．

ETL8 は，対象が手書き教育漢字 881 字種，平仮名 75 字種で 160 文字ずつ格納されている．ETL8 に格納されている手書き文字データの例を 図 1 に示す．

ETL9 は，対象が JIS 第一水準漢字 2965 字種，平仮名 71 字種で 200 文字ずつ格納されている．ETL9 に格納されている手書き文字データの例を 図 2 に示す．

ETL8, ETL9 のデータサイズは，128×127 (pixel) で，濃度レベルは 16 階調である．例で示したデータの作成では，まず，前処理としてこのグレイレベルの階調データを適当な閾値で白黒に 2 値化している．

次に個別文字それぞれの縦方向，横方向それぞれにヒストグラムをとって切り出し領域を決定する．

切り出し領域の縦と横の比率は，個々の文字によって異なるが，縦と横の比率があまり大き

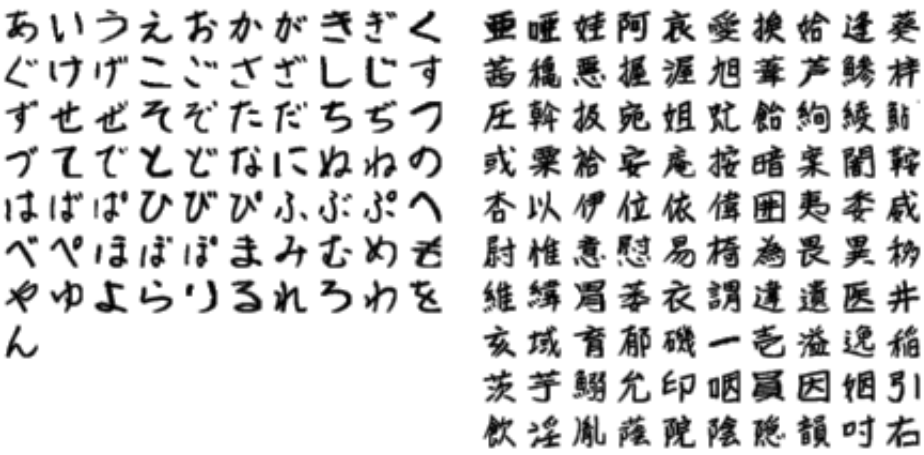

図 2 ETL9 の手書き文字データの例

くなりすぎると，あとで切り出した領域を正方形に正規化すると，線幅が広がりすぎたり，字形が潰れたりして，元の形状と変わりすぎてしまう可能性があるため，切り出し領域は，最初から正方となるように調整した．

図示した文字データは，この切り出したデータを 16×16 の文字サイズになるようにしてから，濃度を 256 階調で揃えたものである．

ETL データベースのデータの格納順は，各カテゴリのデータが連続に格納してあるので，カテゴリ毎のデータとしての扱いが易しい．

また ETL の漢字，ひらがなのデータには，いわゆるゴマ塩ノイズを含むデータは存在するが，ノイズだらけの文字データや，文字として人が判別することができないような汚い文字データはほとんど存在しない．

筆者らは，連続したデータを先頭を 1 番とするシリアル番号を個別データに割り当て偶数データと奇数データのグループに分け，一方を学習データや標準パターン作成のためのデータとして利用し，もう一方を未知パターンとして認識システムの評価に用いてきた．

3 JEITA-HP 手書き文字データベース

JEITA-HP 手書き漢字データベースは，日本ヒューレットパッカード株式会社 (日本 HP) で文字パターンのデータを収集し，電子情報技術産業協会 (Japan Electronics and Information Technology Industries Association) を通して公開，提供されていたデータベースである．

JEITA-HP データベースには “DATASET_A”，“DATASET_B” の 2 種類がありそれぞれのディレクトリに収録されている．

各ディレクトリ内には，「writer-(n).img」のファイル名でデータが格納されており，n は，

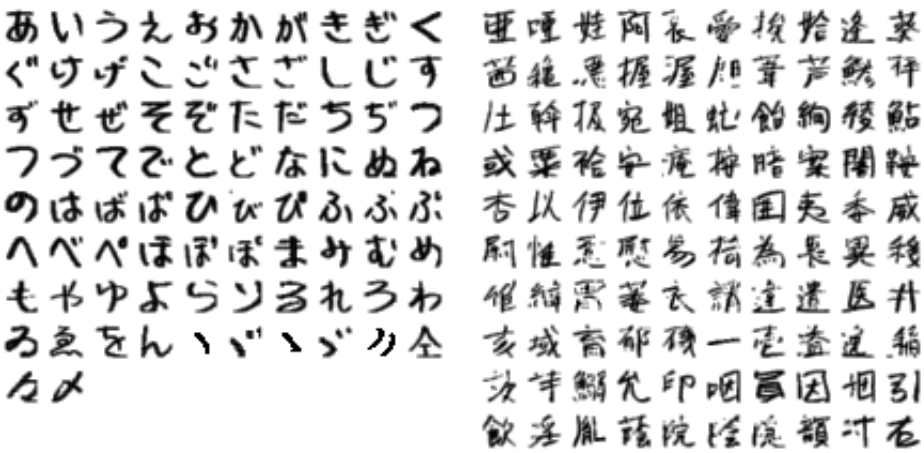

図 3 JEITA-HP の手書き文字データの例 (“DATASET_A”)

0 からの通番が割り当てられている．

この番号は “DATASET_A” では 492 まであり，“DATASET_B” では 99 まである．ただし，“DATASET_A” には，欠番ファイルが 13 ファイル存在する．そのため，“DATASET_A” のファイル数は 480 である．“DATASET_B” には，欠番ファイルは存在せず，ファイル数は 100 である．

1 つのファイルの中には，筆記者 1 人分のデータが格納されており，格納されているデータの内容は，表 1 の通りである．表 1 の内容で，1 ファイルには，最大 3,306 文字が収録されているが，どのファイルにも，全てのデータが揃って収録されているわけではなく，何らかの理由で収録されなかったデータもあり，それは欠落した状態になっている．

したがって，個々の文字についてのサンプル数を正確に把握するには，それぞれのファイルを開いてアクセスし，内容を確かめる必要がある．手書き文字データひとつのフォーマットを表 2 のとおりである．

表 1 JEITA-HP ファイルの内容

開始番号	終了番号	収録内容
0	2964	手書き漢字
2965	3046	平仮名等
3047	3128	平仮名等
3129	3138	数字
3139	3148	数字
3149	3305	英字，カタカナ，記号等

表 2 JEITA-HP データフォーマット

項番	サイズ	内容
1	2(byte)	データ番号，short 型 (ビッグエンディアン)
2	2(byte)	カテゴリ番号，short 型 (ビッグエンディアン)
3	2(byte)	シフト JIS コード
4	512(byte)	バイナリイメージデータ (サイズ：64×64)

4 JEITA-HP データの解析 (ヒストグラム)

JEITA-HP データベースに収録されているがどのようなデータであるかを確かめるための簡単な手法として，ヒストグラムによるデータ解析を行った．JEITA-HP データベースの手書き文字データは，バイナリの画像データであるため，画像データは，0(白) と 1(黒) とに解釈できる．画像のサイズは、64×64 の大きさなので，黒の部分を加算してヒストグラムを作成すると，まっ白な状態では 0 となり，すべてが黒で埋めつくされた状態では，4096 となる．

データベースの中のひとつひとつの手書き文字について，ヒストグラムをとり，ヒストグラムの値をよこ軸，ヒストグラムの値の出現回数をたて軸にとってグラフに表したものが，図 3 および図 4 である．

対象とするデータは，データベースの全データで，“DATASET_A” は $1,58,0779$ 件あり，“DATASET_B” は，$330,595$ 件あった．

“DATASET_A” では，最小値のヒストグラムは 0 で，データベースの中に 2 データあり，ヒストグラムの最大値は 2667 で 1 データあった．ヒストグラム値の最頻値は，610 で，2877 回あった．これに次ぐ頻度の高いヒストグラムは，614，641，597 で，それぞれ 2826 回，2822 回，2803 回 だった．

“DATASET_B” では，最小値のヒストグラムは 4 で，データベースの中に 1 件あり，最大値のヒストグラムは，1895 で，1 データあった．“DATASET_B” でのヒストグラムの最頻値は，614 で，672 回あった．これに次ぐ頻度のヒストグラムは，636，584，607 で，それぞれ 657 回，650 回，649 回であった．

どちらのヒストグラムのグラフも正規分布の形状をしており，最頻値もグラフのピーク周辺に分布していることがわかる．

5 JEITA-JP データの解析 (ノイズ／不読文字)

手書き文字画像をデータとして取り込むときに，幾らかのノイズがデータと一緒に取り込まれてしまうことがある．考えられる理由としては，用紙に最初からついていたシミ，筆記者が残した，消し跡や枠の外からのはみ出し，スキャナについた埃，アナログからデジタルに変換するさいのノイズなどが考えられる．

データベースを作成するときに発生した読み取りエラーの検知や，データベース編纂者によ

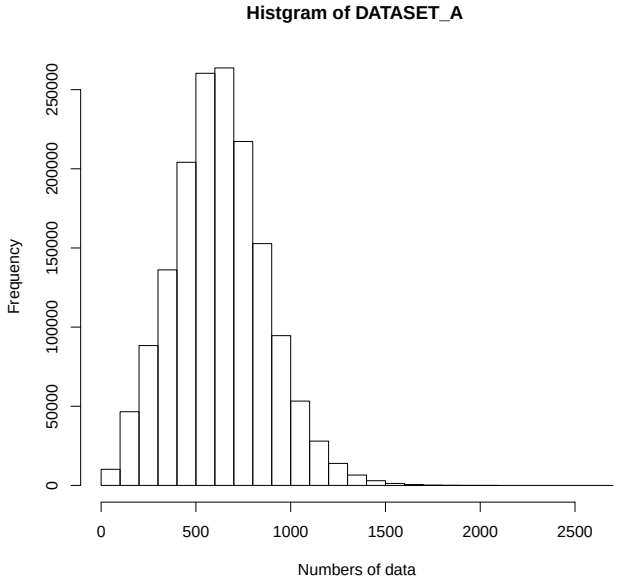

図 4 “DATASET_A” のヒストグラム

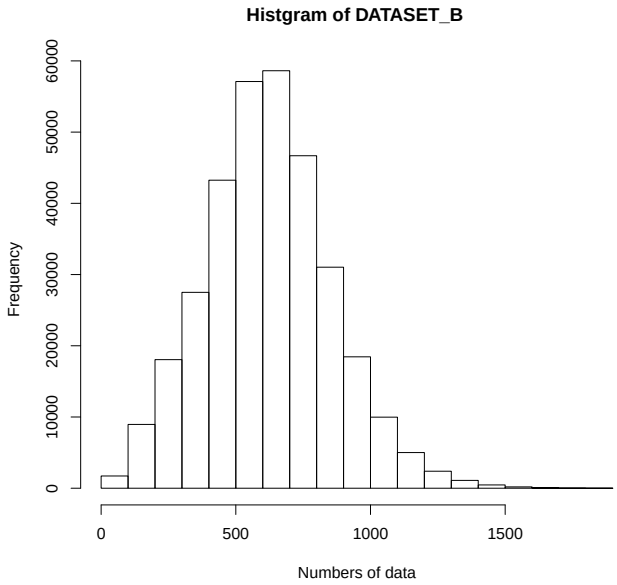

図 5 “DATASET_B” のヒストグラム

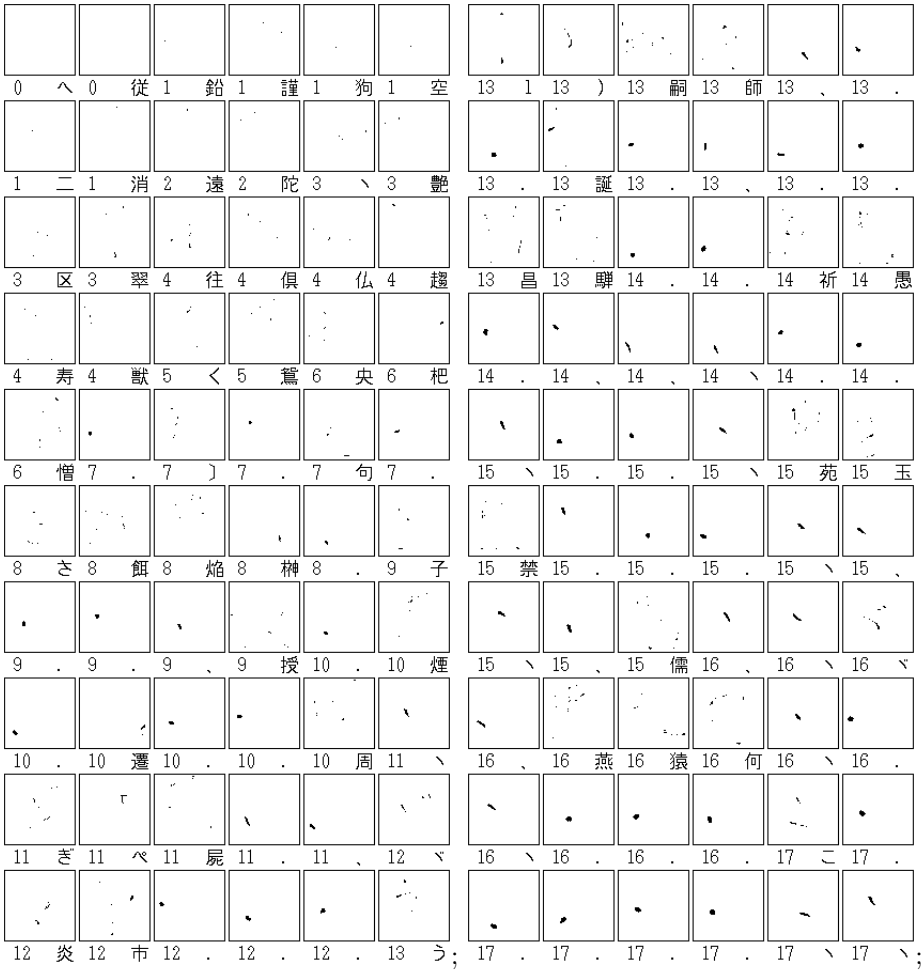

図 6 “DATASET_A”20 画素以下のデータ (1/2)

る，読み取り後の画像データのチェック等で，ある程度のデータの取捨選択は行われていても，データ量が膨大なため，すべてのノイズや読み取りエラーを取り除くことは現実的ではない．

ここでは，JEITA-HP データベースの中から画素数の少ないものだけを選択して，そこに含まれるノイズと不読文字について調査した．

“DATASET_A” でバイナリ画像の黒となる画素の総和が 20 以下となったものすべてを図 6 および図 7 で示す．各画像の左下には画像の画素数を表示しており，右下には，画像に対応する文字を表示するようにした．ファイル名，カテゴリ番号等は，紙面の都合により割愛している．

“DATASET_A” で 20 画素以下のデータは，全部で 200 個あった．画素数 0 のものは

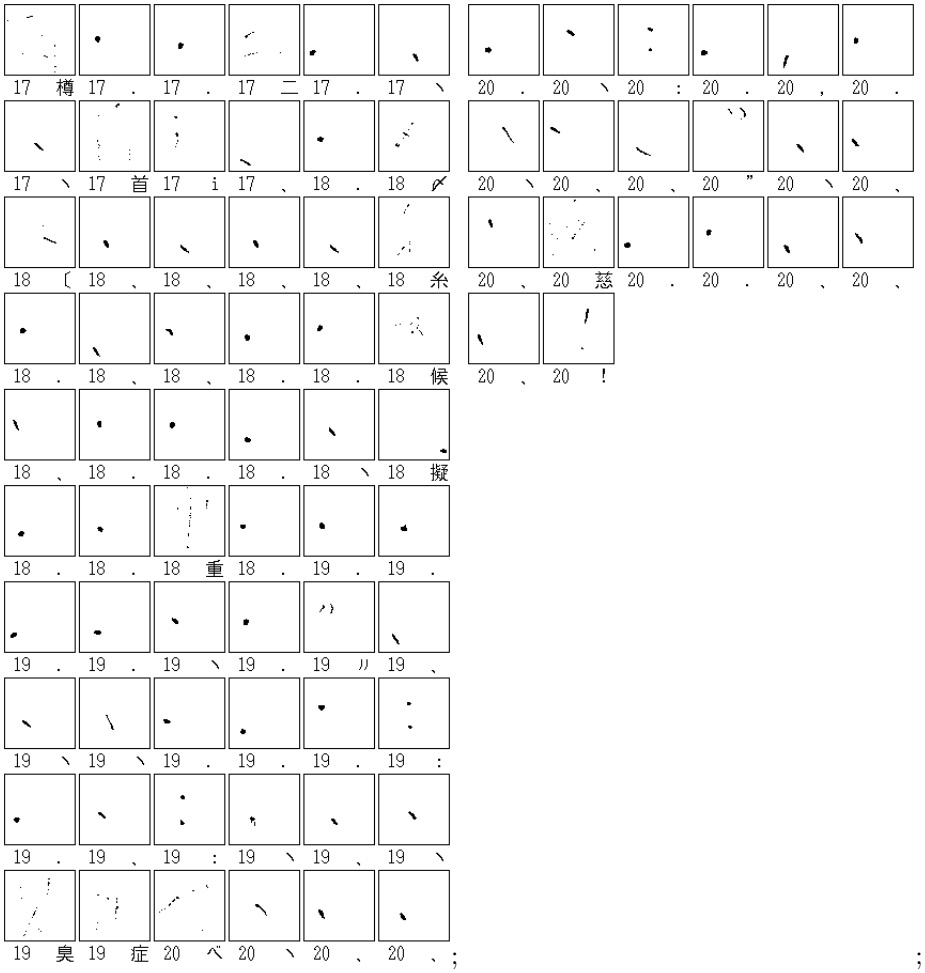

図 7 “DATASET_A” の 20 画素以下のデータ (2/2)

“writer-79.img” のデータ番号 3097 の「へ」と，“write-479.img” のデータ番号 1251 の「従」の 2 データがあった．この 2 データは，中身が何もないデータであり，文字として認識することはできないものである．

画素数 1 のものは 6 データ，画素数 2 のものは 2 データあるが，いずれも漢字で，画素数が少なすぎることから，文字として認識することができない．

画素数 3 の データは 4 つある．このうち，ファイル名 “writer-79.img” データ番号とカテゴリ番号が 3039 の「丶」は，比較的，画素数は少ないが，当該文字と認めるには，困難である．

画素数 5 のデータの中に，ファイル名 “writer-79.img” データ番号とカテゴリ番号 2974 の

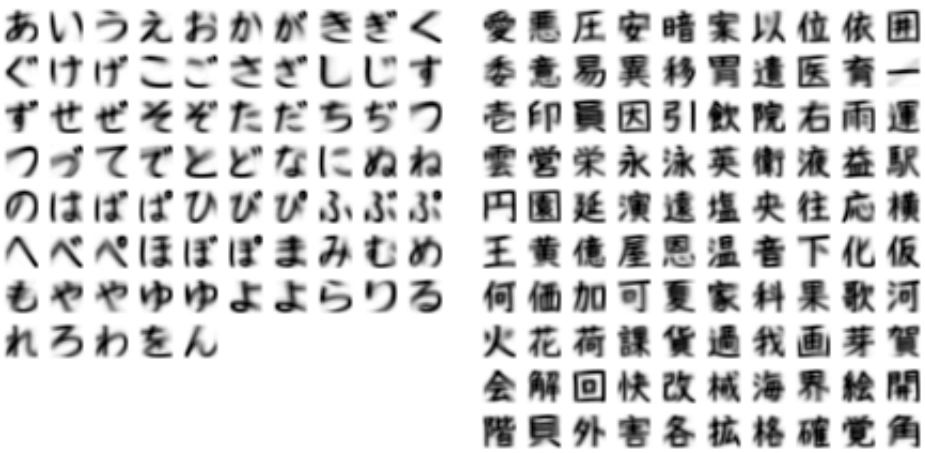

図 8 ETL8 データベース偶数データで作成した平均パターン

データが，ひらがな「く」として収録されているが，これも「く」の文字と認めるのは難しい．

画像と文字コードとが一致していそうなもので，最小の画素数のものは，7 画素で，ファイル名 writer-79.img データ番号 3279，カテゴリ番号 3187 の「.」である．この記号は画素数が少なくて済むため，出現回数も多かった．20 画素以内のデータ全てである 200 件のうち，65 データは「.」記号であった．これに続いて出現回数の多いものは，「、」が 30 件，「ヽ」が 24 件あった．それ以外で，同じ文字が出現したのは「：」が 3 件，「二」と「ゞ」が 2 件で，あとはすべて 1 文字だけであった．

そのほか，このデータ解析で分かったこととして，特定のファイルの出現回数が多かったことが挙げられる．“writer-124.img” は 34 データ，“writer-79.img” は，17 データ “writer-479.img” は 12 データが含まれていた．

ノイズを観察してみると，数点のドットからなるノイズが画像データの中に含まれているものもみられた．

6 単純平均による平均パターンの作成

データベースの文字データを代表するデータとして平均パターンを作成した．もし文字がバラバラであれば，できあがる平均パターンは，平均化されたものとなり，ひどくぼやけた文字となるか，文字と認めることが難しいものとなる．すでに報告している [1] とおり，ETL を使った，手書き数字や手書き英数の平均パターンは，比較的きれいな文字と認められるものになっているが，JEITA-HP でも同程度の平均パターンとなるのかを作成して確かめた．また，比較のため，ETL8 と ETL9 についても同様の平均パターンを作成した．

作成手法は，図 1〜図 3 の作成方法と同じ方法で正規化したものをたし合わせて平均値を

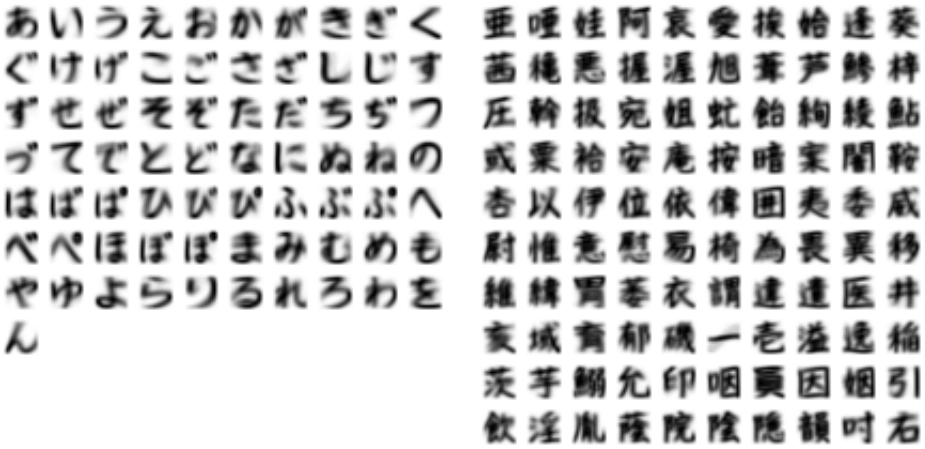

図 9 ETL9 データベース偶数データで作成した平均パターン

図 10 JEITA-HP(DATASET_B) データベース偶数データで作成した平均パターン

とり，それを 256 階調に揃えたものである．データは偶数の組と奇数の組に分けて作成した．ETL8 は 80 文字の平均パターン，ETL9 100 文字の平均パターンである．JEITA-HP の DATASET_A は 236 文字，DATASET_B は 50 文字から作成したものである．

平均パターンで見る限り，JEITA-HP のデータも ETL と同程度の字形の平均パターンができている．ETL，JEITA-HP ともに，ひらがなの濁音，半濁音記号などの見分けは難しいものの，文字としての大体の形は，平均パターンでも残っていることが分かる．

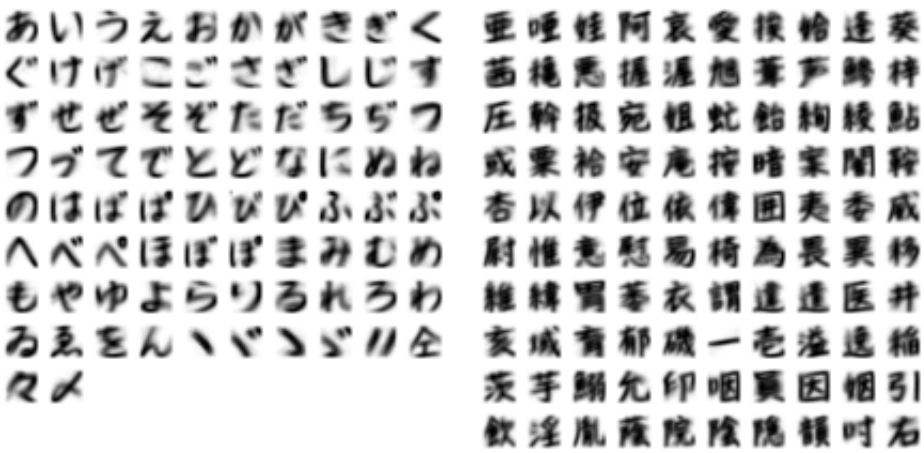

図 11 JEITA-HP(DATASET_B) データベース偶数データで作成した平均パターン

7 さいごに

すでに実用化され，さらに今後，益々の利用が見込まれる文字認識のためのデータベースが日本語に限ってみると，公開されているものは，ほとんど無いのが現状である．しかも，JEITA-HP は新しいデータベースではないが，あまり利用されていないようであった．

日本語文字認識システムの発展には，まだまだ多くの文字データが必要となると思われるが，まずは，このデータベースを整理して利用することにより，もっと広く利用されることを望んでいる．

参考文献

[1] 安田，中島，北村，二階堂，” 文字認識と svm 法”，pp.33-43,vol.20, 明星大学情報学部紀要，2012

[2] 栗田，”http://home.hiroshima-u.ac.jp/tkurita/lecture/svm/index.html”

[3] 甘利，麻生，津田，村田，” パターン認識と学習の統計学”，岩波書店，2003

[4] ” 認識形入力に関する調査研究報告書”, 一般社団法人 電子情報技術産業協会，認識形入力方式標準化専門委員会，2012