手書き漢字データベースのデータ圧縮実験

大倉　充*・今村　太**・塩野　充***

*岡山理科大学大学院理学研究科

システム科学専攻博士課程

**㈱三菱電機東部コンピュータシステム

***岡山理科大学工学部電子工学科

（昭和63年9月30日　受理）

1．まえがき

手書き漢字認識の研究では，開発した認識アルゴリズムを実際にプログラム化し，ソフトウェアシュミレーションによって認識性能を評価することが必要不可欠である。そして，その評価はデータの品質によって大きく左右されるため，各種のアルゴリズムを比較する場合には共通のデータベースが必要となる。そのため，工業技術院電子技術総合研究所（以下，電総研と記す）によって，手書き漢字データベースETL-8（B2）[1]及びETL-9（B2）[2]等が作成され，公開されている。しかし，これらのデータベースは大容量のため，磁気テープに収納されており，オープンリール磁気テープ装置の接続されたシステムでしか活用できないといった不都合な面を持つ。そこで本研究では，これらのデータベースのパソコンやワークステーション上での利用を目的とし，ETL-8（B2）の一部のデータを用いて，パソコンのOSとして最も一般的と考えられるMS-DOS上でのファイル化及びデータ圧縮に関する基礎実験を行い[3][4]，更に，その実験結果の妥当性と最終的なデータベースの大きさの確認のために，ETL-8(B2)の全データを用いて圧縮型データベースの作成を行った。

2．手書き教育漢字データベースETL-8（B2）

2-1　ETL-8（B2）の概要

文字認識アルゴリズムを開発する際，その性能の評価は，認識実験に用いられるデータの品質によって大きく左右されるため，各種のアルゴリズムを比較する場合には共通のデータが必要となる。手書き教育漢字データベースETL-8(B2)（以後，単にETL8と記す）は，そのために，電総研によって作成され公開されている。ETL8は，教育漢字881カテゴリ，ひらがな71カテゴリから形成されており，各カテゴリは，漢字JISコードの順番に並んでいる。また一つのカテゴリは，筆記者の異なった160サンプルを含んでおり，各サンプルは，B2タイプでは，2値化及び位置と重心の正規化が施されているため，例えば，最も基本的な認識手法である重ね合せ的手法などでは，そのまま適用することができる。なお，

ETL8の最終部には，筆記者のために用意されていた見本文字956カテゴリ（956サンプル）も収納されている。

2-2　磁気テープ・フォーマット

図1に，ETL8が収納されている磁気テープの形式を示す。図中，一つのレコードのDATA部に一つのサンプルが，ID部（識別情報部）にそのサンプルのJIS代表読み等の付加情報が格納されている。各サンプルはDATA部の先頭から，図2に示す順番で格納されており，その図形情報の内容を表1に示す。

3．データ圧縮の基本的なアルゴリズム

通常の画像情報のデータ圧縮法は，可逆及び非可逆方式の2種類に大別される。前者では，復元された画像と原画像の間には差異がなく（忠実度が1となる），後者では差異が生じる（忠実度が1未満となる）。これらの圧縮法の選択は，復元された画像がどのように用いられるかによって定まる。本研究では，ETL8が，文字認識アルゴリズムの性能比較に用いられるデータベースであることより，可逆方式を採用した。基本的には，ETL8が2値データであることより，通常，2値ファクシミリ信号の圧縮に用いられる手法を使用しており（ただし，MH符号化[5]のような統計量に基づく手法は，ETL 8以外での使用に難があ

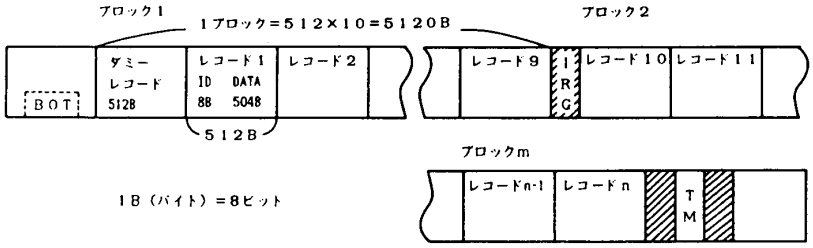


図1　磁気テープ・フォーマット

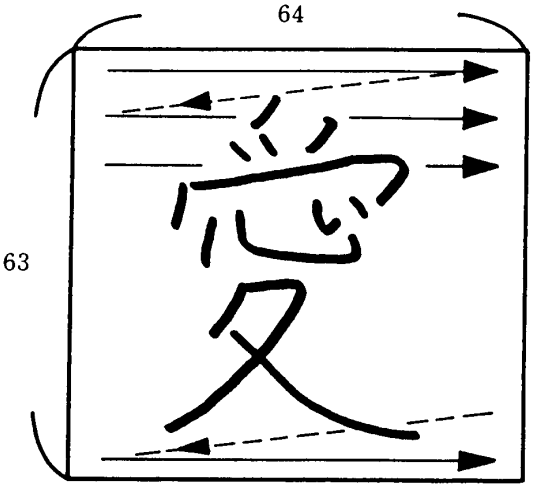


図2　サンプルの格納順

表1　図形情報の内容

標本点数	63×64＝4032　（点）
濃淡レベル	2レベル（1ビット）
総バイト数	504（バイト）

るため用いていない)，以下，それらについて説明を行う。

3-1　2ビット区切り方式[5)]

2値ファクシミリ信号の圧縮に用いられている最も基本的な圧縮手法は，ランレングス符号化法[5)]である。これは，2値画像において，白または黒の画素の継続する長さ(ランレングスという）を符号化する方法である。2ビット区切り方式は，ランレングス符号化法のひとつで，ランレングスを2進数で表示した後に，下位から2ビットずつに区切り，各ブロックごとにその先頭に白領域ランレングスであれば0を，黒領域であれば1をつけて示す方式である。白領域に対してのランレングス符号の割当表を表2に，また，この方式を用いて画素列を符号化した例を図3に示す。

3-2　複数ライン一括符号化法[5)]

ランレングス符号化方式は，走査方向が同じ符号列を処理するという，一次元の相関性に基づくものだが，図面や原稿では，走査線と走査線の間にも相関が存在する。複数ライン一括符号化法とは，この2次元的相関を利用して圧縮効率を上げようというものである。この代表的な方式として，2ラインのOR符号を用いる方式について説明を行う。

まず図4に示すように，2ラインのOR符号で白区間と黒区間を定める。すなわち第1と第2のラインがいずれも白（0，0）の場合は白区間，それ以外は黒区間とする。そして白区間では従来のランレングス符号化を，黒区間では原符号をそのまま用いる。例えば，白区間のランレングス符号化に表3を用い，黒区間は第1と第2のラインの符号を交互に並べると図5に示す結果が得られる。ただし黒区間の終わりは，この結果に示すように00を挿入して区切りを明確にする。本研究では，2ライン及び3ラインの2種類の一括符号化法を用いた。

表2　白領域の符号化

ランレングス	符　号	符号長
1～3	0＊＊	3
4～15	0＊＊0＊＊	6
16～63	0＊＊0＊＊0	9
64～255	0＊＊0＊＊0＊＊0＊＊	12

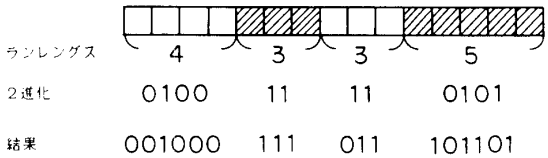


図3　2ビット区切り方式説明図

表3　白区間の符号化

ランレングス	符　号	符号長
1～3	0＊＊	3
4～15	1＊＊0＊＊	6
16～63	1＊＊1＊＊0＊＊	9
64～255	1＊＊1＊＊1＊＊0＊＊	12

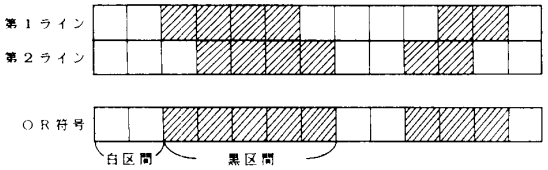


図4　2ラインのOR符号

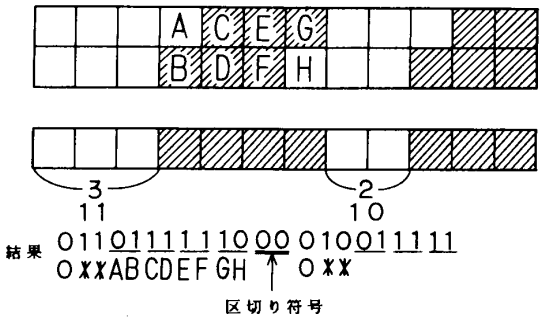


図5 2ライン一括符号化法説明図

3-3 理想圧縮率[6]

一次元符号化においては，各ランレングスの生起確率より理想圧縮率を求めることができる。あるランレングスkの生起確率をP_Kとすると1ランレングス当りのエントロピーH[7]は，次式で与えられる。

$$H = -\Sigma P_K \log_2 P_K \quad (1)$$

これは，一つのランレングスを2進表示するときに必要な平均ビット数の理論限界を表している[6]。一方，平均ランレングスはΣkP_Kであるから，理想圧縮率Cは，

$$C = (H/\Sigma kP_K) \times 100(\%) \quad (2)$$

となる。ただし，ビット数は整数値でなければならないから，これは限界を示すことになる。

4. データ圧縮実験

4-1 MS-DOSファイル化

データ圧縮実験に先立ち，ETL8のMS-DOSファイル化を行った。使用装置は本学情報処理センターの大型計算機FACOM/M380で，使用言語はFORTRAN77である。処理としては，まずETL8の図形情報に直接関係しない，ダミーレコードとID部の削除を行う。次にF6650エミュレータの機能の一つである"ファイル転送"を用いて，15カテゴリずつ1枚のフロッピィディスク（1．25MB）に収納する。その結果，見本文字を含めて，ETL8の全サンプルが65枚のフロッピィディスクに収納される。

4-2 実験データ，使用装置及び使用言語

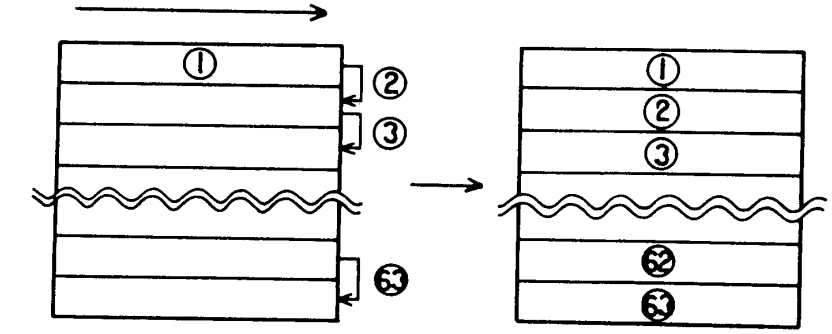
実験に使用したデータは，図6に示す教育漢字の先頭『愛』から『角』までの100カテゴリで，160サンプル/カテゴリより，計16000サンプルとなる。画数は，『一』の1画から『衛』の16画まで各種混在している。実験に用いた装置はPC-9801E(パソコン)，PC-9881K(8インチ・ディスクドライブ)，言語はCである。

4-3 変換図形の作成

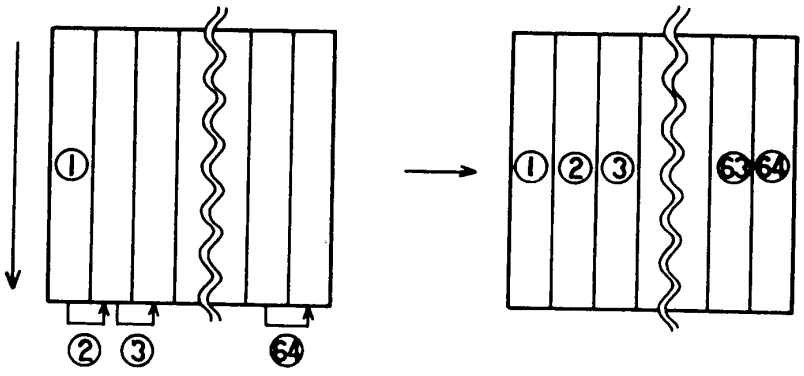
ランレングス符号化法には，ランレングスの値が大きい程圧縮効率が良いという特徴が

愛	悪	圧	安	暗	案	以	位	依	囲
委	意	易	異	移	胃	遺	医	育	一
壱	印	員	因	引	飲	院	右	雨	運
雲	営	栄	永	泳	英	衛	液	益	駅
円	園	延	演	遠	塩	央	往	応	横
王	黄	億	屋	恩	温	音	下	化	仮
何	価	加	可	夏	家	科	果	歌	河
火	花	荷	課	貨	過	我	画	芽	賀
会	解	回	快	改	械	海	界	絵	開
階	貝	外	害	各	拡	格	確	覚	角

図6　実験に使用したカテゴリ



（a）原図形　　（b）横走査



（a）原図形　　（c）縦走査

図7　変換図形作成の説明図

あり，複数ライン一括符号化法では，黒区間を符号化していないため，白区間のランレングスの値を大きくすることによって，両手法での圧縮効率の向上が期待できる。そこで本研究では，原図形に対して図7に示すように，各行あるいは各列ごとに排他的論理和をとった変換図形を作成し，（圧縮時における走査方向の違いから，前者を横走査の変換図形，後者を縦走査の変換図形と呼ぶ）その図形に対しても圧縮実験を行った。図8に変換図形

（a）原図形　　（b）横走査　　（c）縦走査

図8　変換図形の一例

表4　平均圧縮率（%）①

圧縮方法	原図形	変換図形
2ビット区切り方式	46.3	37.1
2ライン一括符号化法	42.0	35.1
3ライン一括符号化法	41.2	37.2

の例として，『愛』を示す。

4-4　圧縮率の定義

圧縮後のデータは，8ビット（＝1バイト）の整数倍となっていない場合が多い。そのため，そのようなデータに対しては，データの終わりに1～7個までの0を付け加えて，8ビットの整数倍となるようにしている。また，各サンプルによって圧縮後のデータの長さが異なるため，データの長さを2バイトを使って表し，データをファイルに格納する際，その情報をデータの先頭に付加している。圧縮率の計算には，この2バイトも含めており，圧縮率C_Rは，次式で定義した。

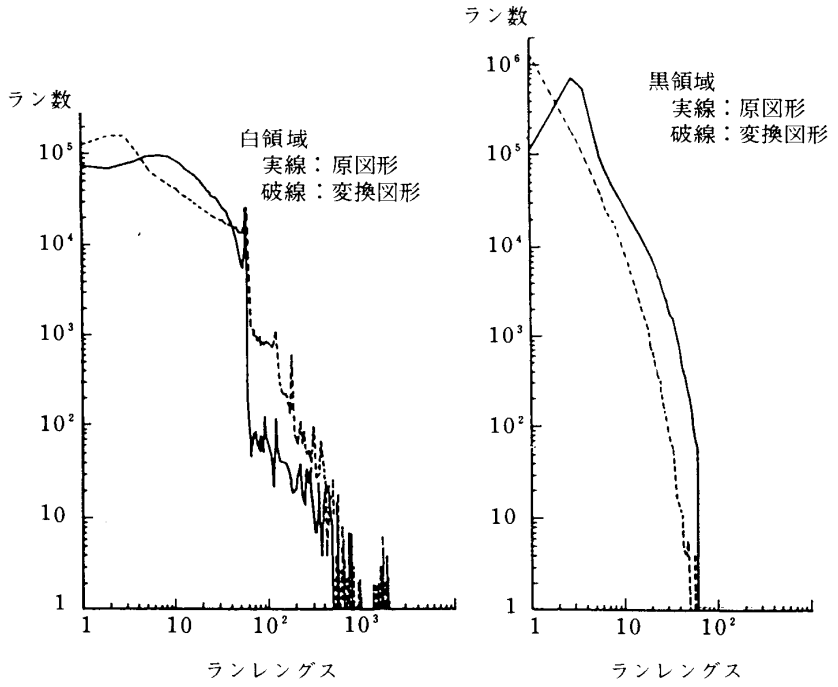
$$C_R = \frac{(\text{圧縮後，実際にファイルに格納されたときのバイト数})}{504} \times 100(\%) \tag{3}$$

ここで分母の504は原図形のバイト数を表している。

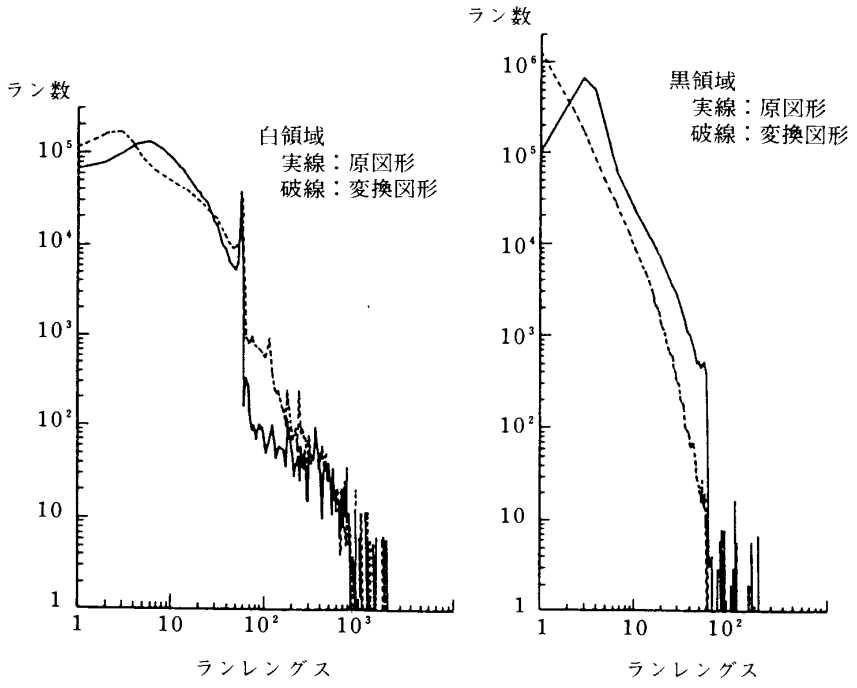
4-5　圧縮実験結果

表4に得られた，漢字100カテゴリ全体の平均圧縮率を示す。これらは，図形の左上から右下へ横方向に走査した（横走査）結果である。また，ここでの"変換図形"の欄は，横走査の変図換形に対しての結果である。なお，2ライン一括符号化法では，原図形の終わりに64個の0を付け加えて，64行としている。この結果より，全ての方法において変換図形に対しての圧縮率の方が良く，また，この3種類の圧縮法では，2ライン一括符号化法の変換図形に対しての圧縮率が最も良いことがわかる。

図9に実験に用いた100カテゴリについての，ランレングスの分布を示す。横軸はランレングスを，縦軸はそのランレングスの出現した回数（ラン数）を表している。同図(a)は，図形の左上から右下へ横方向に走査したもの（横走査）で，同図(b)は，図形の左上から右下へ縦方向に走査したもの（縦走査）である。両走査とも変換図形では，白領域のランレ



(a) 横走査



(b) 縦走査

図9　ランレングス分布

ングスの値は大きくなり，黒領域のものは逆に小さくなるという傾向があり，このことが表4の平均圧縮率に反映されていることがわかる。黒領域の分布で特筆すべきことは，変換図形において分布のピークが，ランレングス1のところに移動し，ランレングスの増加と共になだらかに分布が減少していることである。これは，黒領域においてHuffman符号

表5 一次元符号化の理想圧縮率(%)

圧縮方法	原図形	変換図形
横走査	33.8	27.1
縦走査	32.1	27.0

表6 平均圧縮率(%)②

横走査	縦走査	走査選択
32.6	32.5	31.7

化[5]がなされていることに相当する。従って，2ビット区切り方式において黒領域には原符号を割り当てる符号化によって，圧縮率の向上が期待できる。表5にランレングス分布より求めた，一次元符号化の理想圧縮率を示す。これより変換図形においては，横走査及び縦走査で圧縮率に差がないことがわかるが，文字形状によっては，両走査の結果に大きな差の生じることが推測される。以上のことより，2ビット区切り方式に対して，変換図形の黒領域には原符号を割り当てるという改良を加えた方法で実験を行い，その結果を表6に示す。表中，走査選択というのは，サンプルごとに両走査を行い，圧縮率の良いものを選択するという方法である。表4，表6より，本実験においては，変換図形に対しての，2ビット区切り方式（黒領域原符号）走査選択（2ビット区切り走査選択法と呼ぶ）で最も良好な圧縮率が得られた。

4-6 考 察

本実験を通して得られた，圧縮率の向上を妨げる原因についての考察を行う。まず，ETL 8の各サンプルの文字ストロークが，16×16や24×24ドットのROM漢字パターン等のストロークのように線幅1の直線でなく，不均一な線幅を持つこと，これに関係して，ストロークの輪郭線上に凹凸があり，短いランレングスでの切り替わりが多いこと，変換図形を作成しても左右斜め方向の黒領域成分が残るため，白領域のランレングスが分断されてしまうこと等が挙げられる。また，ETL 8には，これらの他に，通常の2値図形の圧縮に用いられる種々の方法(例えば，フリーマンのチェーン符号化[5])の適用に不適当な要因，即ち，分岐点の存在，数ドットの雑音，線幅のため必然的に生じる内部のホールを形成する輪郭線の存在，複数個の単独に存在するストローク等が含まれている。

以上の事柄を考慮すると，本実験で得られた圧縮率より良好な圧縮率を得るには，原画像に対して，既存の圧縮手法だけを適用したのでは困難と思われる。また，より良好な圧縮率を得るための上述の問題点を全て解決した圧縮手法の開発も非常に困難と考えられる。従って，原図形になんらかの前処理を施して，上述の問題点を少なくする手法を探求する方（新たな変換図形の作成法の開発）がより実際的と考えられる。

5. ETL-8（B2）圧縮型データベース

上述の実験結果の妥当性及び最終的なデータベースの大きさの確認のために，2ビット区切り走査選択法を用いて，ETL8の全サンプルのデータ圧縮を行い，圧縮型データベースを作成した。

表7　全サンプルの平均圧縮率（%）

ひらがな 75カテゴリ	教育漢字 881カテゴリ	見本文字 956サンプル	全サンプル 153916サンプル
23．4	32．0	27．3	31．3

表8　圧縮型データベースの内容

Vol. ナンバー	ファイル名	カテゴリ数
1	ENCOD1～64．DAT	64
2	ENCOD65～115．DAT	51
3	ENCOD116～163．DAT	48
4	ENCOD164～207．DAT	44
5	ENCOD208～252．DAT	45
6	ENCOD253～299．DAT	47
7	ENCOD300～348．DAT	49
8	ENCOD349～395．DAT	47
9	ENCOD396～455．DAT	50
10	ENCOD446～493．DAT	48
11	ENCOD494～540．DAT	47
12	ENCOD541～588．DAT	48
13	ENCOD589～634．DAT	46
14	ENCOD635～682．DAT	48
15	ENCOD683～728．DAT	46
16	ENCOD729～775．DAT	47
17	ENCOD776～823．DAT	48
18	ENCOD824～870．DAT	47
19	ENCOD871～916．DAT	46
20	ENCOD917～956．DAT	40
20	ENCOD957．DAT	
20	DECODE	

5－1　全サンプルの平均圧縮率

表7に，2ビット区切り走査選択法を用いて得られた全サンプルの平均圧縮率を示す。この結果より，圧縮実験に用いた漢字100カテゴリに対して得られた結果と大差ないことがわかる。また，この圧縮によって，ETL8の全サンプルは，圧縮されたデータの再生プログラムを含めて，20枚のフロッピィディスクに収納された（表8参照）。

5－2　圧縮後のデータの格納形式

サンプルごとに走査方向が異なるため，圧縮後のデータは，図10に示すような形でファイルに書き込まれる。圧縮後のデータが8ビットの整数倍になっていない場合には，符号化したデータの終わりに“0”を1～7個のいずれかを付加（図中，付加ビット）して，全体を8ビットの整数倍とし，そのトータルのバイト数を2バイトで表現した後，データの先頭に付加する（図中，付加情報）。また，走査方向については，先頭2バイトの先頭ビッ

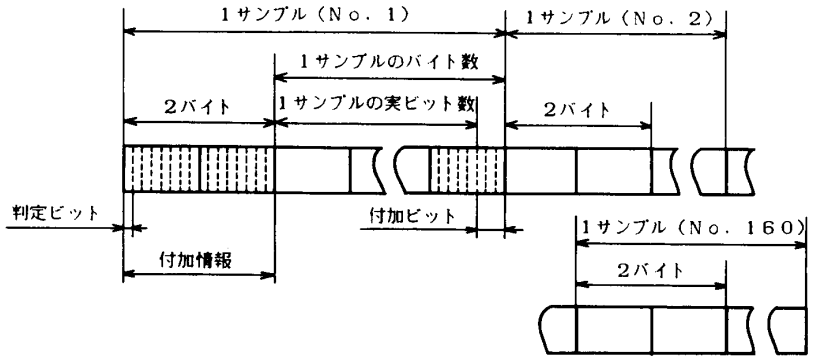


図10 圧縮後のデータの格納形式

ト（図中，判定ビット）の"0"，"1"で判定を行う。

5-3 圧縮型データベースの内容

ETL 8の956カテゴリ＋見本文字956サンプルが，20枚のフロッピィディスクに収納されている。ファイル名はENCOD 1．DAT～ENCOD957．DATで，ファイル名中の数字がカテゴリのJISコードの順番に対応している(但し，ENCOD957．DATは見本文字を示す)。従って，ENCOD 1．DATには，ひらがなの『あ』が，ENCOD956．DATには，漢字の『話』が収納されている。また，各ファイルには，1カテゴリ160サンプルが，オリジナルのETL 8に納められている順番どうりに収納されている。但し，オリジナルとは異って，ID部及びダミーレコードは削除されてデータ部のみとなっている。表8に，20枚の各フロッピィディスクの内容を示す。Vol．20のディスクには，見本文字のファイル及びデータの再生プログラム（DECODE）が納められている。なお，再生プログラムDECODEについてのマニュアルを作成しているため，初心者でもこの圧縮型データベースの使用は可能である。データの再生に要する時間は，PC9801Eを用いて，1カテゴリ当り約120秒である。

6．むすび

電総研の手書き漢字データベースの，パソコンやワークステーション上での利用を目的として，ETL8のデータの一部を用いて，MS-DOSファイル化及びデータ圧縮に関する基礎実験を行った。その結果，本実験では，2ビット区切り走査選択法において最も良好な圧縮率が得られ，この手法を用いて，ETL8の全サンプルを圧縮し，フロッピィディスク20枚から成る圧縮型データベースを作成した。しかし，単純計算で，この結果をETL9に当てはめると，フロッピィディスク80枚というかなり膨大な量の圧縮型データベースとなる。従って，ETL9についての圧縮型データベースの作成のためには，更に圧縮手法に関する研究が必要である。しかし，新たな圧縮手法の開発よりは，むしろ，新たな変換図形の作成手法の開発の方が，より実際的と考えられる。

参考文献

1）斎藤泰一，山田博三，山本和彦，森　俊二："手書き文字データベースの解析(V)—教育漢字データベースのパターン・マッチング法による評価—"，電子技術総合研究所彙報，Vol. 45，Nos. 1，2，pp.49—77（1981—02）.

2）斎藤泰一，山田博三，山本和彦："手書き文字データベースの解析（Ⅷ）—方向パターン・マッチング法によるJIS第1水準手書き漢字データベースETL 9の評価—"Vol. 49，No. 7，pp. 487—525(1985—07).

3）大倉　充，今村　太，塩野　充："手書き漢字パターンのデータ圧縮に関する一考察"，昭和62電気四学会中国支部連合大会，092109，pp. 203（1987—10）.

4）大倉　充，今村　太，塩野　充："手書き漢字データベースのデータ圧縮実験"，昭和62電気関係学会関西支部連合大会，G 8—54，pp. G272（1987—11）.

5）安居院猛，中嶋正之："画像工学の基礎"，昭晃堂，東京（1986）.

6）吹抜敬彦："画像のディジタル信号処理"，日刊工業新聞社，東京（1981）.

7）瀧　保夫："通信方式"，コロナ社，東京（1985）.

Data Compression Experiments of Handwritten KANJI Data Base

Mitsuru OHKURA* · Futoshi IMAMURA** and Mitsuru SHIONO***

*Graduate School, Okayama University of Science,
Ridaicho 1-1, Okayama, 700 Japan
**Mitsubishi Electric Computer
Systems (Tokyo) Corporation
***Faculty of Engineering,
Okayama University of Science

(Received September 30, 1988)

At the study of recognition of handwritten KANJI characters, it is also necessary to evaluate the recognition efficiency of the developed method. The common data base is necessary when we compare the various algorithms. So, the handwritted KANJI data base ETL-8 (B2) is made and open to public by Electrotechnical Laboratory in Japan. This data base is stored in 3 open reel magnetic tapes because of its large volume.

In this paper, we converted this data base into the data base on MS-DOS and showed the results of data compression experiments with it made for the purpose of using it. on a personal computer or a work-station.