3C-4

圧縮型手書き漢字データベース作成のためのデータ圧縮の基礎実験

大倉　充†　今村　太††　塩野　充†
(†岡山理科大学工学部電子工学科　††㈱三菱電機東部コンピュータシステム)

[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，その評価は認識対象とな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，共通のデータベースが必要となる．そのため，電総研によって，手書き漢字データベースＥＴＬ－８[1]やＥＴＬ－９[2]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，磁気テープに格納されており，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，ＥＴＬ８（２値化済みのＢ２タイプを用いた）の一部のデータを用いて，パソコンのＯＳとして一般的と思われる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，ＥＴＬ８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] 基本的なデータ圧縮法

一般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．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，前者を採用した．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用いた（但し，ＭＨ符号化[5]のような統計量に基づく手法は用いていない）．

2－1 ２ビット区切り方式[5]　ランレングス符号化法[5]の一つで，ランレングス（白または黒画素の継続する長さ）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．

2－2 複数ライン一括符号化法[5]　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１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（但し，黒区間の終りは，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．

2－3 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Ｐｋとすると，１ランレングス当りのエントロピーＨは，次式で与えられる[7]．

$$H = -\sum Pk \log_2 Pk \qquad (1)$$

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．一方，平均ランレングスは$\sum k\,Pk$であるから，理想圧縮率Ｃは，

$$C = (H/\sum k\,Pk) \times 100\ (\%) \qquad (2)$$

となる．

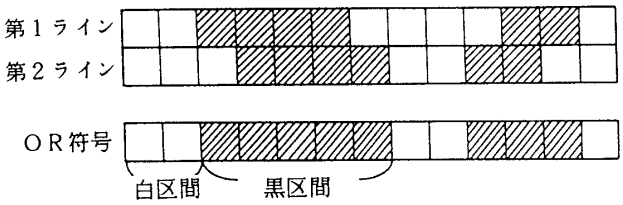

図１．２ラインのＯＲ符号

[3] データ圧縮実験

3－1 ＭＳ－ＤＯＳファイル化　まずＥＴＬ８の図形情報に直接関係しない，ダミーレコードと識別情報部の削除を行う．次に１５カテゴリずつ１枚のフロッピイディスク（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枚のフロッピイディスクに収納される．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．

3－2 実験データ　実験に使用したデータは，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／カテゴリゆえ，計１６０００サンプルとなる．また画面次数は，６３×６４である．圧縮実験に用いた装置はＰＣ－９８０１Ｅ（パソコン），使用言語はＣである．

3－3 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（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）を作成し，その図形に対しても圧縮実験を行った．

3－4 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ＣＲは，次式で定義した．

A Fundamental Experiment on Data Compression for Making Compressed Handprinted KANJI Character Data Base.
Mitsuru OHKURA†, Futoshi IMAMURA††, Mitsuru SHIONO†
†Okayama University of Science.
††Mitsubishi Electric Computer Systems(Tokyo) Co..

表1　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

表2　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3　平均圧縮率②（%）

横走査	縦走査	走査選択
32．6	32．5	31．7

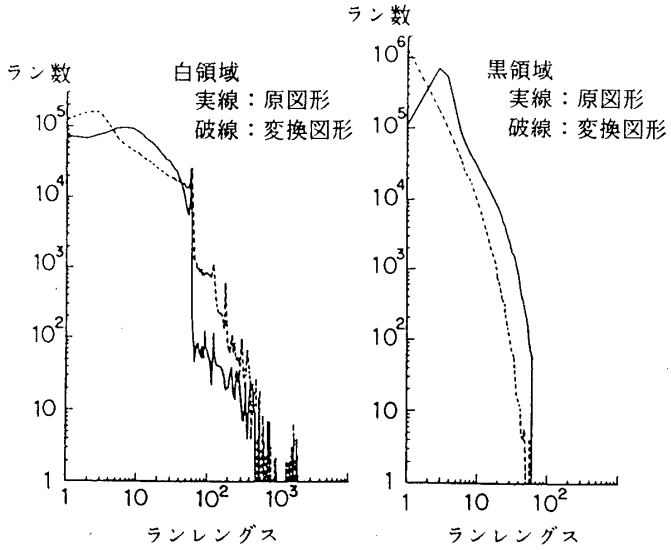

図2．ランレングス分布（横走査）

表4　全サンプルの平均圧縮率（%）

ひらかな 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

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

ここで分母の504は原図形のバイト数を表している．

3－5 圧縮実験結果　表1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水平方向に走査（横走査）した結果である．図2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）を表しており，紙面の都合上，横走査の結果のみ示す（縦走査の場合もほぼ同様の結果となる）．ここで，黒領域の分布で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ラン数が減少していることが分かる．これは，黒領域においてHuffman符号化[5]がなされていることに相当する．また，この結果より求めた一次元符号化の理想圧縮率を表2に示す．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異の生じることが推測される．以上述べたことより，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方法で圧縮実験を行い，得られた結果を表3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，圧縮率の良い方を選択するという方法である．表1，表3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）走査選択で最も良好な圧縮率が得られた．

3－6 圧縮型データベースの作成　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，最も良好な圧縮率の得られた方法で，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4に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分かる．また，この圧縮によって，ETL8の全サンプルは，復元用プログラムを含めて，20枚のフロッピイディスクに収納された．なお圧縮されたデータの復元時間は，PC9801Eを用いて，1カテゴリ（160サンプル）当り約120秒である．

［4］むすび

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，基礎的なデータ圧縮実験を行い，最も良好な圧縮率の得られた手法を用いて，ETL8を基にした圧縮型データベースの作成を行った．しかし実用レベルで考えたとき，そのデータベースはまだ大きすぎるので（特に，ETL9を考えた場合），更に圧縮率を向上させるデータ圧縮法についての検討が必要と考えている．

参考文献

（1）斎藤，山田，山本：”手書文字データベースの解析（V）”，電総研彙報，45，1，2，1981．
（2）斎藤，山田，山本：”手書文字データベースの解析（Ⅷ）”，電総研彙報，49，7，1985．
（3）大倉，今村，塩野：”手書き漢字パターンのデータ圧縮に関する一考察”，昭62中国連大，092109．
（4）大倉，今村，塩野：”手書き漢字データベースのデータ圧縮実験”，昭62関西連大，G8－54．
（5）安居院，中嶋：”画像工学の基礎”，昭晃堂（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（1981）．
（7）瀧　保夫：”通信方式”，コロナ社（1985）．