1085

Bell Laboratories

Cover Sheet for Technical Memorandum

The information contained herein is for the use of employees of Bell Laboratories and is not for publication (see GEI 13.9-3)

Title– Numeric Pattern Recognition Data Base for UNIX Systems

Date– January 5, 1976

TM– 76-1353-1

Other Keywords–

Author(s)	Location and Room	Extension
M. E. Greenberg	HO 4G-636	7923
C. S. Roberts	HO 4F-617	2044

Charging Case– 39394

Filing Case– 39394-100

ABSTRACT

An IEEE pattern recognition data base of approximately 10,000 hand printed digits has been obtained. It has been reformatted so that it may be used with a PDP-11/45 UNIX computer system. This memorandum describes the data base contents and format.

Pages Text 1 Other 7 Total 8

No. Figures – No. Tables – No. Refs. –

Address

E-1532-C (6-73)

SEE REVERSE SIDE FOR DISTRIBUTION LIST

Bell Laboratories

Subject: Numeric Pattern Recognition Data Base for UNIX Systems - Filing Case 39394-100, Charging Case 39394

Date: January 5, 1976

From: M. E. Greenberg
C. S. Roberts
TM 76-1353-1

MEMORANDUM FOR FILE

We have acquired the IEEE Pattern Recognition Data Base 1.2.4 and made it available under UNIX on a PDP-11/45 computer. It consists of almost 10,000 hand-printed digit samples. These samples were collected throughout Japan by the Tokyo Shibura Electric Company, Ltd. in conjunction with the Japanese Ministry of Postal Services for use in designing a postal zip code reader. The samples include patterns from each of the ten numeral classes. They were written in pre-printed boxes so some standardization of size is evident. No other restrictions were placed on writers so there is a great variety of character shapes in each class. Since samples were written with several different writing implements, there is also a wide range of stroke widths and pattern quality in the data base.

The data base comes from IEEE on two 7-track magnetic tape reels. Patterns are stored as binary images with a resolution of 36 x 50. There are 4,160 patterns on one tape reel and 5,262 patterns on the other reel. Appendix A contains the documentation received from the IEEE with the data base. It includes a brief description of the data base, the tape record format, the distribution of characters with the numeral classes, and the octal character code used on the tapes. The data base has been reformatted and transferred to two 9-track, 800 bpi magnetic tape reels for use with a UNIX system. Appendix B gives the format for the reformatted data base. Appendix C contains a listing of a simple C program which prints consecutive patterns, one per page, from the UNIX data base.

M. E. Greenberg /CSR

M. E. Greenberg

C. S. Roberts

C. S. Roberts

MEG
HO-1353-CSR- ms

Att.
Appendices A (5 pages)
B (1 page)
C (1 page)

TOSHIBA RESEARCH AND DEVELOPMENT CENTER
TOKYO SHIBAURA ELECTRIC CO., LTD.

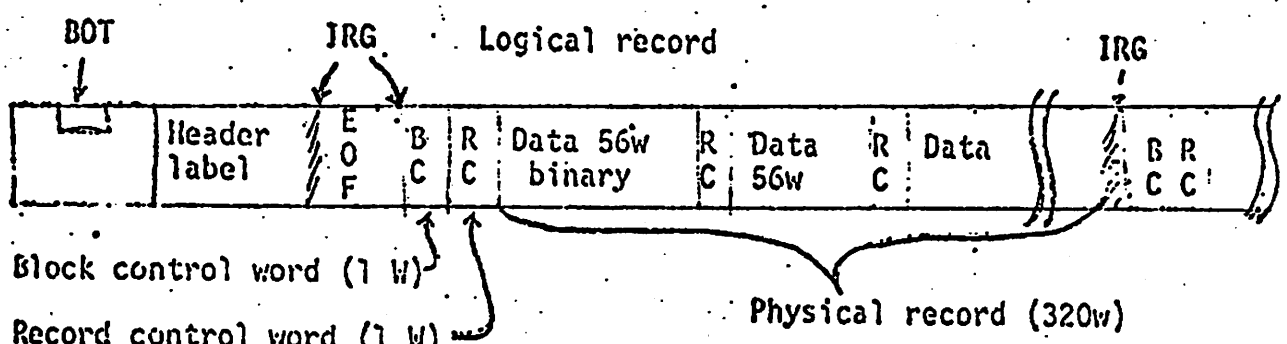
1 KOMUKAI TOSHIBA-CHO, KAWASAKI, 210, JAPAN

MAGNETIC TAPE RECORDING FORMAT

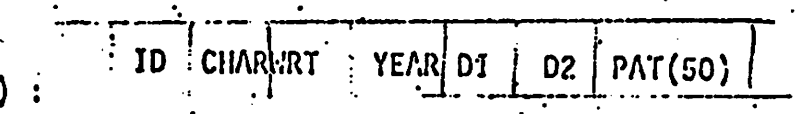
Number of tracks	:	7
Recording mode	:	NRZI
Density	:	556 bits/inch
A word	:	36 bits
A character	:	6 bits

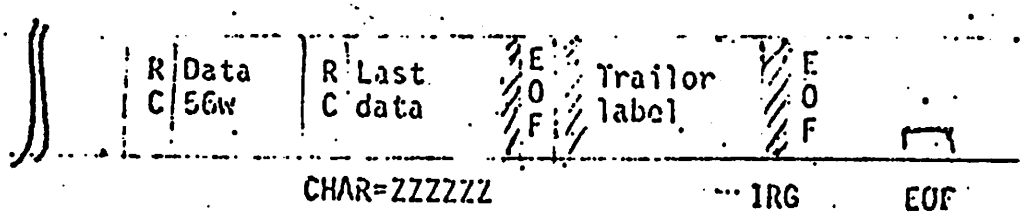
A single pattern consists of 56 words;

ID (1 w)	:	Identification number (Integer)
CHAR (1 w)	:	Character name code (BCD)
WRT (1 w)	:	Writing instrument (BCD)
YEAR (1 w)	:	Year (BCD)
D1 (1 w)	:	Dummy (BCD)
D2 (1 w)	:	Dummy (BCD)
PAT (50 w)	:	36x50 bit binary patterns



Data (56w) :





TOSHIBA RESEARCH AND DEVELOPMENT CENTER
TOKYO SHIBAURA ELECTRIC CO., LTD.

1 KOMUKAI TOSHIBA-CHO, KAWASAKI, 210, JAPAN

The octal code expresion of characters in the magnetic tapes is as follows, where an odd parity check is adopted.

char.	octal code	char.	octal code	char.	octal code	char.	octal code
0	00	b	20	J	41	+	60
1	01	A	21	K	42	/	61
2	02	B	22	L	43	S	62
3	03	C	23	M	44	T	63
4	04	D	24	N	45	U	64
5	05	E	25	O	46	V	65
6	06	F	26	P	47	W	66
7	07	G	27	Q	50	X	67
8	10	H	30	R	51	Y	70
9	11	I	31	-	52	Z	71
		.	33	$	53	,	73
		(	35	*	54	=	75
				)	55		
				'	57		

Appendix B

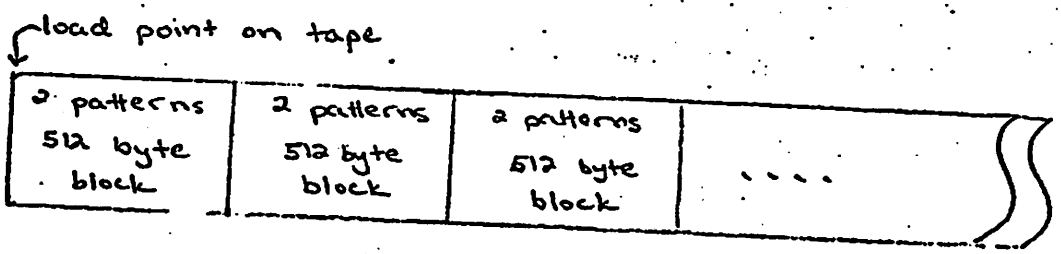
Magnetic Tape Recording Format
UNIX Tapes

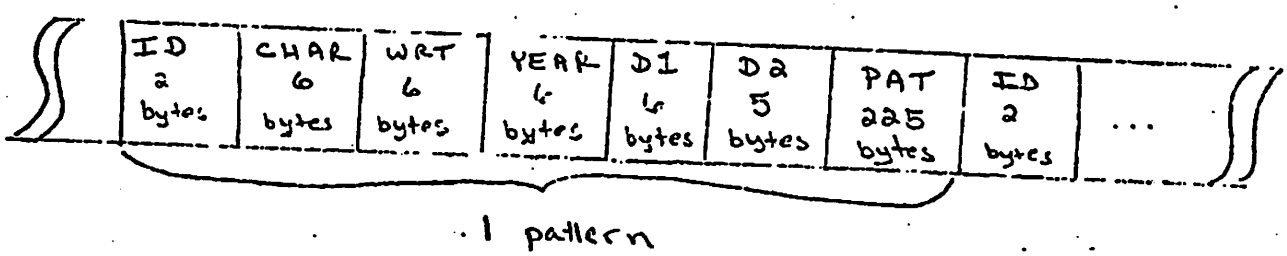
Number of tracks	:	9
Recording mode	:	NRZI
Density	:	800 bits/inch
A word	:	16 bits
A character	:	8 bits (1 byte)

A single pattern consists of 256 bytes:

ID (2 bytes)	: Identification number (Integer)
CHAR (6 bytes)	: Character name code (ASCII, 6 characters)
WRT (6 bytes)	: Writing instrument (ASCII, 6 characters)
YEAR (6 bytes)	: Year (ASCII, 6 characters)
D1 (6 bytes)	: Dummy
D2 (5 bytes)	: Dummy
PAT (225 bytes)	: 36x50 bit binary pattern, interpretated as a bit string.

NOTE: The last character of D2 from the original tape has been deleted.





Dec 17 16:34 ieeels.c Page 1

```
main()
{
 struct ieeerec {
       int idno;
       char bytes[254];
 } arec;
 int nn,k;

again:
 nn = read(0,&arec,256);  /* read 256 byte pattern with header */
 if (nn < 256) return;
 printf("%d %29s\n",arec.idno,arec.bytes);  /* print header information */
 for (k=0; k<10; k++) putchar('\n');
 xxpat(&arec.bytes[29]);  /* print pattern */
 for (k=0; k<10; k++) putchar('\n');
 goto again;  /* print next pattern */
}

xxpat(bits)
 char bits[225];
{
 char mask,xtype,line[38],*iblnks;
 int i,lcnt,j,xcnt,m;

 iblnks = "      ";
 line[36] = 0;
 i = lcnt = 0;
 j = xcnt = 0;
 mask = 1;
 while (i<225) {
       if (bits[i] & mask)  { xtype = 'X'; xcnt++; }
       else xtype = ' ';
       line[j++] = xtype;
       if (j >= 36) {  /*  end of one row in pattern */
          if (xcnt) {  /* nonblank line */
             m = 35;
             while (line[m] == ' ') --m;
             line[m+1] = 0;
             printf("%s%s\n",iblnks,line);  /* print row of pattern */
          }
          else putchar('\n');  /* row is blank */
          xcnt = j = 0;
       }
       mask = mask << 1;
       if (mask == 0) { i++; mask = 1; }
 }
}
```

Data Base Number	Data Type	Description	Recording Medium	Source	Cost
1.2.4	Handprinted Numeric Characters	The data base consists of 10,000 hand written numeric characters collected from live as well as experimental mail throughout Japan. The data base is contained on two magnetic tapes with each tape having 5,000 characters. Images are binary with a resolution of 36 x 50.	Magnetic Tape, 7 Track, 2 Reels, 556 BPI	Toshiba Research and Development Center Tokyo Shibaura Electric Co., Ltd. Mr. Hiroshi Genchi 1 Komukai Toshiba-Cho, Kawasaki, 210, JAPAN	

TOKYO SHIBAURA ELECTRIC CO., LTD.
1 KOMUKAI TOSHIBA-CHO, KAWASAKI, 210, JAPAN

Character distribution in magnetic tape #1001

character	number	percent
0	500	10.0
1	500	10.0
2	500	10.0
3	500	10.0
4	500	10.0
5	500	10.0
6	500	10.0
7	500	10.0
8	500	10.0
9	500	10.0
total	5000	100.0

Character disiribution in magnetic tape #1002

character	number	percent
0	605	12.1
1	549	11.0
2	486	9.7
3	501	10.0
4	576	11.5
5	535	10.6
6	569	11.4
7	460	9.2
8	374	7.5
9	348	7.0
total	5000	100.0

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

COMPUTER GROUP

Data Base Description

DATA BASE NUMBER 1.2.4

Data Type

Handprinted Numeric Characters

Source

Toshiba Research and Development Center
Tokyo Shibaura Electric Co., Ltd.
Mr. Hiroshi Genchi
1 Komukai
Toshiba-Cho, Kawasaki, 210, JAPAN

Date

October 1970

Recording Medium

Magnetic Tape, 7 Track, 2 Reels, 556 BPI, 6 Bits per Character

Description

The data base consists of 10,000 hand written numeric characters collected from live as well as experimental mail throughout Japan. The data base is contained on two magnetic tapes with each tape having 5,000 characters. Images are binary with a resolution of 36 x 50. A single pattern consists of 56 words.