Proposal of Japanese Printed Character Database

Youichi Aimu Kunihito Kato Kazuhiko Yamamoto
Department of Information Science Faculty of Engineering
Gifu University
1-1 Yanagido Gifu 501-1193 JAPAN
e-mail: aimu@yam.info.gifu-u.ac.jp

Abstract

In the character recognition, common database is needed for research and improving method. Currently, however, common Japanese databases such as ETL2 in Japan are not sufficient for data quantity, a variety of font, data quality and so on. So that this paper proposes a database which satisfies these requirements. Proposed database is evaluated throughout character recognition experiment using four direction features. The result of the experiment indicates that most font type is thick Gothic. However, the database is satisfying necessary requirements.

1 Introduction

There are many papers in character recognition [1][2][3], relaxation method[4][5][6], contour dynamic programming[7], mesh pattern method[8], direction pattern matching[9] and etc. The comparing performance of these methods is very useful for development of recognition methods and improvement of performance. The comparing performance requires common database. In Korean, common database is published[10]. If different databases are used for comparing performance of two methods, we cannot compare them exactly. However, many reports use the databases which are made by themselves. The database need many categories and a variety of font, because Japanese have 6879 categories. These categories include simple category such as alphabet and Kana and etc., and complex category such as JIS level-1 Kanji and JIS level-2 Kanji. There are some categories whose differences are only a dot such "富" and "冨", or either dot or small circle such as "ぺ" and "ペ".

However, published Japanese databases such as ETL2[11] are not sufficient for data quantity, a variety of font, data quality.

In this paper, we propose the Japanese printed character database which satisfies necessary requirements. We give the overview of it, and made initial set of the database. The database is evaluated by experiment of character recognition with four direction pattern matching method.

2 Overview of proposal database

The proposed Japanese printed character database is planed in cooperation with Japan Electronic Industry Development Association(JEIDA). Many data sheets which printed some Japanese character with various fonts are collected. There is twelve sheets in each data type. Alphabet, Kana, etc. compose one sheet. JIS level-1 Kanji composes five sheets. JIS level-2 Kanji composes six sheets. The database consists of many images which are scanned these data sheet. In present step, we scan six sheets without six sheets of JIS level-2 Kanji. Because JIS level-2 Kanji are not used in general. Each data sheet is printed a maximum number of 600 characters. One sheet includes 30 lines, and each line includes 20 characters. The line for comment is placed in the lowest line. The comment line includes font ID, character size, sheet condition ID in collecting sheets, sheet index and font name such as shown in Figure 1. Font ID is index for discriminating font type. Character size is either 10pt or 12pt. Sheet condition ID classifies type of data sheet. For example, the data sheet is included fine quality paper, recycling paper, one time copied paper or etc. The category black square "■" as the guide to segment these character placed surroundings of characters on the data sheet such as shown in Figure 2.

Specification of the database is following.

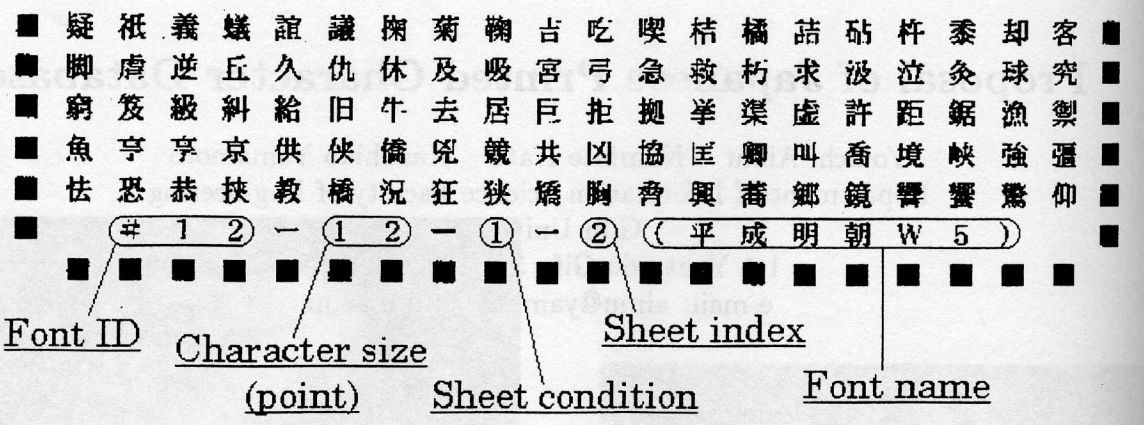


Figure 1: The line for comment in data sheet

- Categories : Categories are all Japanese categories determined by Japan Industrial Standard(JIS).
  - special symbol : 147
  - figure : 10
  - Roman alphabet : 52
  - Hiragana : 83
  - Katakana : 86
  - Greek character : 48
  - Russian character : 66
  - ruled line scrap : 32
  - JIS level-1 Kanji : 2965
  - JIS level-2 Kanji : 3390
- Character Size : Each character is printed with 10pt and 12pt.
- Segmentation Guide : The category black square "■" is used as the guide for segment of each character.
- Data File : Data sheets which are scanned binary 400dpi are saved in no compaction TIFF files.

At present, the number of data which is collected are shown in Table 1. In Table 1, Mincho has thick vertical lines and thin horizontal lines and serif. Kaku Gothic has square stroke tip. Maru Gothic has round stroke tip. The example of data sheet is shown in Figure 2. Each font list of character "漢" is shown in Figure 3.

This database is named JEIDA Fuji format database.

Table 1: The class of data

Font	66types†	
	Mincho	19types(1-19)
	Gothic	30types(20-49)
	Others	17types(50-66)
Size	10pt and 12pt	
Condition	original and copy	
Category	3488categories	

† 14 font types in 66 font types have only original sheet without copy.
Number in a round bracket correspond to number in Figure 3.

3 Experiment for evaluation

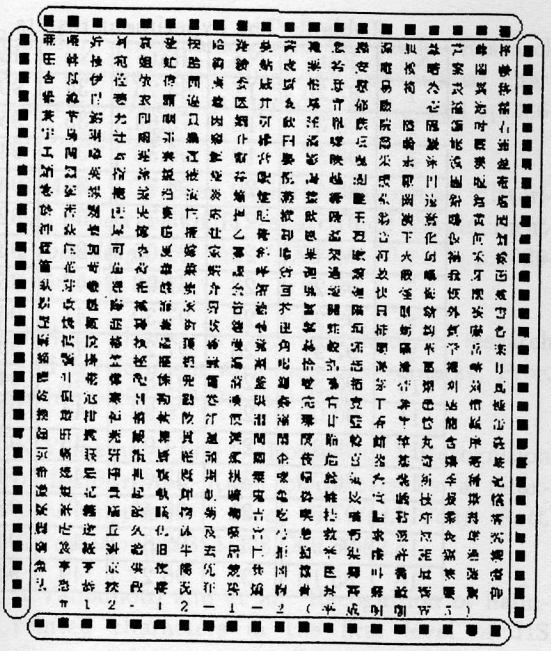
In order to evaluate this database, it is experimented to recognize with four direction feature.

3.1 Experiment system

Each character image is segmented for the data sheet by segmentation guide. A rectangle circumscribing of character is extracted from the input image whose size is 96x96. At this time, a lump whose area is less than fifteen is ignored as noise. Tiny character is placed in center of 60x60 square. If length of long side of rectangle is more than three times the length of short side, the image is placed in center of long side of rectangle square.

Four direction feature patterns[9] whose size are 16x16 are extracted from the rectangle circumscribing, and are blurred using Gaussian filter such as shown Figure 4. These blurring patterns are used in fine classify. These dimension of patterns is decreased 4x4 for rough classifying.

Before the experiment, the dictionary for fine



: Guides for Segmentation

Figure 2: The example of the data sheet

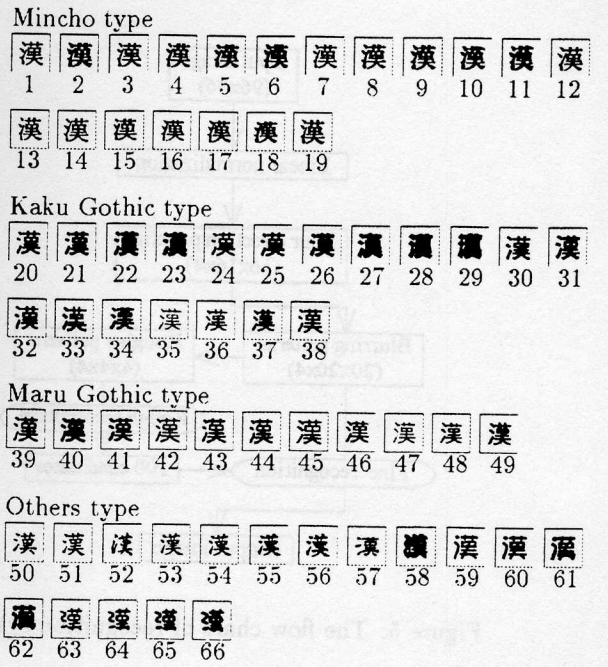


Figure 3: The example of character fonts "漢" (original sheet, 12pt)

classify is made to take the average of the blurring patterns of learning data set. The dictionary for rough classify is made to take the average of 4x4 patterns of learning data set. In experiment, the rough classify selects one hundred candidates. Then, the fine classify selects the category whose similarity is the highest of these candidates.

Figure 5 shows the flowchart of this process.

Each category is learned image of 168 types. These types include two point sizes and two condition types and forty eight font types. Font types are Mincho, Kaku Gothic and Maru Gothic. Because, these fonts are in general use. These images are first to 49th in Figure 3.

We exclude No. 29 in Figure 3 from experiment data, because its stroke is too thick to regard as general. Table 2 shows the data list.

3.2 Result

Table 3 shows the experiment results. The number corresponds to the number in Figure 3. On the whole, font which has thick stroke results bad rate. The results of Kaku Gothic are specially bad rate, because space between two thick strokes is very

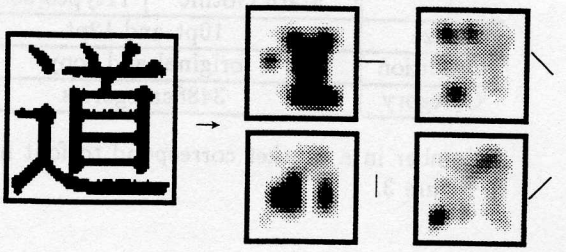


Figure 4: The four direction features of "道"

small or not exist, as 23, 27 28, etc. in Figure 3.

Table 4 shows details about condition and character size. Table 5 shows details about condition and font type. Original sheet of 12pt results best recognition rate in Table 4. Because these images show details of character. On the other hand, in Table 5, copy sheet whose quality ought to go down, results better rate about Gothic. Because space betweeen two strokes is expanded by copy, and the feature of the space of copy sheet is extracted further than the feature of the space of original sheet.

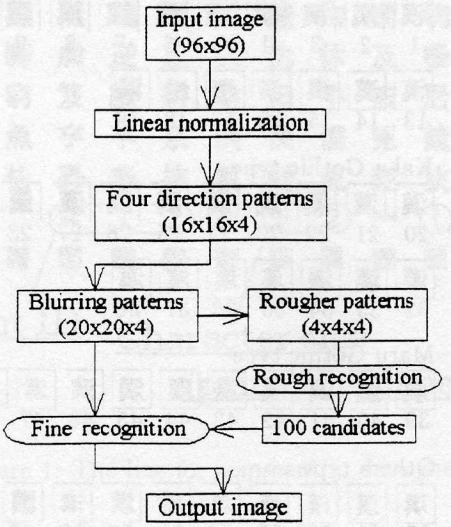


Figure 5: The flow chart of recognition system

Table 2: The data for experiment

Font	48types‡	
	Mincho	19types(1-19)
	Square Gothic	18types(20-37)
	Maru Gothic	11types(38-49)
Size	10pt and 12pt	
Condition	original and copy	
Category	3488categories	

‡ Number in a bracket correspond to font number in Figure 3.

4 Consideration

At present, we gather 236 types of sheet sets, and many data is collecting.

About font type, Gothic has large data, because Gothic includes two types. Most font has also thick stroke. About collecting condition, sheet condition type is original sheet or copy sheet. However, the number of the classes are increased, as recycling paper and several times copy paper.

5 Conclusion

We need common database of Japanese printed character and discussed common format to promote to gather huge fonts with JEIDA committee. We proposed JEIDA Fuji format and gather several fonts. In this paper, its overview is shown. The quality of the database is evaluated with character recognition. The result indicates that most font has thick stroke.

We need to collect more data sheet than this scanned sheets for common database. Anyway, we hope that the database is widely useful for research and improving method.

Table 3: The experiment results

No.	Rate	No.	Rate	No.	Rate
1	97.20	17	83.26	34	90.76
2	96.34	18	79.64	35	95.49
3	98.57	19	96.56	36	94.69
4	98.45	20	96.57	37	94.87
5	97.76	21	93.71	38	92.32
6	94.45	22	82.05	39	93.17
7	98.22	23	61.93	40	85.91
8	98.23	24	97.30	41	96.06
9	97.92	25	97.22	42	93.75
10	96.81	26	91.89	43	94.14
11	90.00	27	76.96	44	93.60
12	96.81	28	74.94	45	93.61
13	96.75	30	93.64	46	94.37
14	97.59	31	90.49	47	95.49
15	95.99	32	92.85	48	94.22
16	85.10	33	96.22	49	94.14

Table 4: The result of copy condition and character size

	original	copy	Total
10pt	91.14 %	89.93 %	90.62 %
12pt	94.90 %	94.25 %	94.62 %
Total	93.02 %	92.09 %	92.62 %

Table 5: The result of copy condition and font type

	original	copy	Total
Mincho	95.42 %	91.97 %	94.02 %
Kaku Gothic	90.23 %	90.69 %	90.43 %
Maru Gothic	93.45 %	94.43 %	93.89 %

Acknowledgment

The authors wish to thank the members of JEIDA committee for their useful discussion and supports.

References

[1] S.Mori, K.Yamamoto and M.Yasuda. Research on machine recognition of handprinted characters. *Trans. PAMI*,vol.PAMI-6,no.4,pp.386-405,July 1984.

[2] S.Mori, C.Y.Suen and K.Yamamoto. Historical review of OCR research and development. *Proceedings of the IEEE*,vol.80,no.7,pp.1029-1058,July 1992.

[3] S.Mori. Issues in document analysis research and development with historical remarks. *Proc. DAS*,pp.297-319,October 1996.

[4] K.Yamamoto and A.Rosenfeld. Recognition of handprinted Kanji characters by a relaxation method. *Proc. 5th Int. Conf. Patt. Recognition*,pp.395-398,Oct. 1982.

[5] K.Yamamoto and et al. Recognition of handprinted characters in the first level of JIS Chinese characters. *Proc.8th ICPR*,pp570-572,October 1986.

[6] T.Nagasaki, T.Yamamoto and M.Nakagawa. The behavior of dynamic relaxation in an elastic stroke model for character recognition. *Proc. 4th ICDAR*,vol.1,pp16-22,August 1997.

[7] H.Yamada. Contour DP matching method and its application to handprinted Chinese character recognition. *Proc.7th ICPR*,pp.389-392,1984.

[8] M.Ohkura, Y.Shimada, M.Shiono and R.Hashimoto. On discrimination of handwritten similar KANJI characters by subspace method using several features. *Proc. 2nd ICDAR*,pp589-592,October 1993.

[9] M.Yasuda, K.Yamamoto and H.Yamada. Effect of the perturbed correlation method for optical character recognition. *Pattern Recognition*,vol.30,no.8,pp1315-1320, 1997.

[10] D.H.Kim, Y.S.Hwang, S.T.Park, E.J.Kim, S.H.Paek and S.Y.Bang. Handwritten Korean character image database PE92. *Proc. 2nd ICDAR*,pp470-473,October 1993.

[11] K.Yamamoto, H.Yamada and T.Saito. Current state of recognition method for Japanese characters and database for research of handprinted character recognition. *From Pixels to Features III*,pp.105-116,May 1992.