The Impact of Large Training Sets on the Recognition Rate of Off-line Japanese Kanji Character Classifiers

Ondrej Velek and Masaki Nakagawa

Tokyo University of Agr. & Tech. 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan
velek@hands.ei.tuat.ac.jp

Abstract. Though it is commonly agreed that increasing the training set size leads to improved recognition rates, the deficit of publicly available Japanese character pattern databases prevents us from verifying this assumption empirically for large data sets. Whereas the typical number of training samples has usually been between 100-200 patterns per category until now, newly collected databases and increased computing power allows us to experiment with a much higher number of samples per category. In this paper, we experiment with off-line classifiers trained with up to 1550 patterns for 3036 categories respectively. We show that this bigger training set size indeed leads to improved recognition rates compared to the smaller training sets normally used.

1 Introduction

It is a well-known fact that having enough training data is a basic condition for getting good recognition rate. The same as a small child can read only neatly written characters similar to characters in his textbook, the recognizer can correctly recognize only characters which are similar to characters used for training. For western languages it is easier to collect training sets with hundreds or thousands of character samples per category because the number of different categories is low. This is different for far-east languages, such as Japanese and Chinese, where a much higher number of different categories must be taken into account. Because collecting large numbers of Japanese and Chinese character patterns is an extremely time and money-consuming process, until now, only small Kanji character pattern databases have been made publicly available. The ETL-9 database, collected ten years ago in the Electrotechnical Laboratory; the Japanese Technical Committee for Optical Character Recognition, has been the number one database for developing and testing new algorithms in off-line character recognition for many years. During the recent years, nevertheless, new on-line databases were introduced and made publicly available, so that there is sufficient benchmark data available today.

We have used these new databases for training and testing off-line classifiers. In this paper, we will investigate if these training set sizes several times bigger than usually available will significantly increase the recognition rate. Section 2 will first introduce these new databases. Section 3 describes the two off-line classifiers used in our experiments followed by Section 4 presenting our experimental results. Finally, Section 5 contains our general conclusion.

2 Japanese Character Databases

We use four handwritten Japanese character databases in our experiments: the ETL-9 database, and three new databases: JEITA-HP, Kuchibue and Nakayosi.

The new JEITA-HP database collected in Hewlett-Packard Laboratories Japan consists of two datasets: Dataset A (480 writers) and Dataset B (100 writers). Dataset A and B were collected under different conditions. Generally speaking, Dataset B is written more neatly than Dataset A. The entire database consists of 3,214 character classes (2,965 kanji, 82 hiragana, 10 numerals, 157 other characters /English alphabet, katakana and symbols/). The most frequent hiragana and numerals are twice in each file. Each character pattern has resolution 64x64 pixels, which are encoded into 512Bytes. In [1] authors of JEITA-HP shortly introduce new database.

Two large, originally on-line characters databases were collected in recent years in Nakagawa Laboratory of Tokyo University of Agriculture & Technology [2][3][4]. They are now publicly available to the on-line character recognition community. The two databases are called “Kuchibue” and “Nakayosi” respectively. In total, there are over 3 million patterns in both databases, which are written by 120 writers (Kuchibue) and 163 writers (Nakayosi) respectively, with the set of writers being different for both databases. Both on-line databases have the advantage that they account better for the variability in practice because the characters were written fluently; in sentences without any style restrictions and their character shapes are more natural compared to most off-line databases, such as ETL9B. However, on-line databases cannot be directly used for an off-line recognizer. We will generate bitmaps using our newly developed method for generating highly realistic off-line images from on-line patterns.

From the pen trajectory of an on-line pattern, our method generates images of various stroke shapes using several painting modes. Particularly, the calligraphic modes combine the pen trajectory with real stroke-shape images so that the generated images resemble the characters produced with brush pen or any other writing tool. In the calligraphic painting mode based on primitive stroke identification (PSI) [5], the strokes of on-line patterns are classified into different classes and each class of strokes is painted with the corresponding stroke shape template; while in the calligraphic painting mode based on stroke component classification (SCC) [6], each stroke is decomposed into ending, bending and connecting parts, and each part is painted with a stroke shape template. The sample of generated images can be seen in Figure 1.

We generated from each on-line pattern five off-line images, always using different painting modes. The size of the generated images is 96x96 pixels.

The last database used in our experiment is ETL-9. ETL-9 includes 3,036 categories with 200 sample patterns for each category. The image size 64x63 pixels encoded into 504B.

In, Table 1, we compare the characteristics of all the databases. However, in our following experiments we use only Kanji and hiragana characters (3036 categories), which form the intersection of all the databases.

Fig. 1. Patterns generated from on-line database Nakayosi by calligraphic painting mode.

Table 1. Statistics of Kanji character pattern databases.

Database	Format	#Categories	#Writers	#Pattern		
				Per category	Per writer	Total
Kuchibue	vector	3356	120	120-47760	11962	1435440
Nakayosi	vector	4438	163	163-45314	10403	1695689
Etl9B	bitmap	3036	4000	200	N/A	607200
JEITA-HP	bitmap	3214	580	580-1160	3306	1917480

3 Classifiers Used in Our Experiments

We use two off-line recognition schemes in our experiments. The first recognizer represents each character as a 256-dimensional feature vector. It scales every input pattern to a 64x64 grid by non-linear normalization and smoothes it by a connectivity-preserving procedure. Then, it decomposes the normalized image into 4 contour sub-patterns, one for each main orientation. Finally, it extracts a 64-dimensional feature vector for each contour pattern from their convolution with a blurring mask (Gaussian).

For our first classifier [MQDF] a pre-classification step precedes the actual final recognition. Pre-classification selects the 100 candidate categories with the shortest Euclidian distances between their mean vectors and the test pattern. The final classification uses a modified quadratic discriminant function (MQDF2) developed by Kimura from traditional QDF. Our second classifier [LVQ] uses only 196-dimensional feature vectors; i.e., 49 features for each orientation. It is based on the MCE training method (minimum classification error) proposed by Juang and Katagiri, which is a well-known LVQ algorithm.

4 Training Off-line Recognizers with Large-Size Training Sets

From four databases described in Table 1, we will create two large training sets and several sets for testing.

The first training set **Train_A** includes 750 samples per each of 3,036 categories and is formed only from originally off-line samples. The first 200 samples are from ETL-9 and the rest, 550 is from JEITA-HP (464 from dataset A and 86 from dataset B). Testing sets used with this training file are these: **Test_1**, 16 samples from dataset A; **Test_2**, 14 samples from dataset B of JEITA-HP. **Test_3** is from independent samples mostly captured from post addresses.

The second training set **Train_B** includes 800 samples per category and is created from Nakayosi database by our methods for generating hi-realistic images. Appropriate testing set **Test_4** was generated by the same methods, but from the second on-line database Kuchibue.

Two classifiers LVQ and MQDF from section 3 were trained with Train_A and Train_B. MQDF classifier was trained also with merged $\{\text{Train_A} \cup \text{Train_B}\}$ set, so that the maximum number of training sample achieved 1,550 patterns for each of 3,036 categories.

Table 2. LVQ classifier. Train_A set: 50-750 training samples.

[LVQ]	50	100	200	400	750
Test_1	85.01	87.16	88.71	91.36	92.98
Test_2	90.58	92.35	93.49	95.42	96.19
Test_3	85.4	87.1	88.71	91.21	91.99

Table 4. MQDF classifier Train_A $\cup$ Train_B set: 50-1550 training samples.

mqdf	50	100	200	400	750	1550
Test_1	88.34	90.29	91.46	93.02	93.87	94.41
Test_2	93.31	95.07	95.84	97.07	97.45	97.53
Test_3	89.74	91.49	92.67	94.04	94.31	94.52

Table 3. LVQ classifier Train_B set: 50-800 training samples.

[LVQ]	50	100	200	400	800
Test_4	81.86	84.65	86.63	88.05	89.31

Table 5. MQDF classifier Train_B $\cup$ Train_A set: 50-1550 training samples.

mqdf	50	100	200	400	800	1550
Test_4	84.84	88.32	90.4	91.01	92.28	93.88

Graph 1. Error rate for LVQ classifier **Graph 2.** Error rate for MQDF classifier

5 Conclusion

In this paper, we have demonstrated the effect of large training sets on Japanese Kanji character classifiers. The collection of training sets of up to 1,550 samples for each of 3,036 categories was only possible thanks to several newly collected Kanji character databases and our new method for generating highly realistic images from on-line patterns, which allows us to employ also on-line databases in off-line recognition. We have proved that increasing the training set beyond the size typically used still significantly increases recognition rate for Japanese character recognition.

References

- [1] T. Kawatani, H. Shimizu; Handwritten Kanji Recognition with the LDA Method, Proc. 14th ICPR, Brisbane, 1998, Vol.II, pp.1031-1035
- [2] M. Nakagawa, et al., On-line character pattern database sampled in a sequence of sentences without any writing instructions, Proc. 4th ICDAR, 1997, pp.376-380.

- [3] S. Jaeger, M. Nakagawa, Two on-line Japanese character databases in Unipen format, Proc. 6th ICDAR, Seattle, 2001, pp.566-570.
- [4] K. Matsumoto, T. Fukushima, M. Nakagawa, Collection and analysis of on-line handwritten Japanese character patterns, Proc. 6th ICDAR, Seattle, 2001, pp.496-500.
- [5] O. Velek, Ch.Liu, M. Nakagawa, Generating Realistic Kanji Character Images from On-line Patterns, *Proc. 6th ICDAR*, pp.556-560, 2001
- [6] O. Velek, Ch.Liu, S.Jaeger, M.Nakagawa, An Improved Approach to Generating Realistic Kanji Character Images and its Effect to Improve Off-line Recognition Performance, accepted for ICPR 2002.