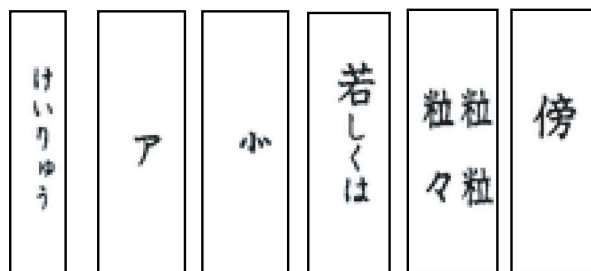


A Study on Decision Rule for Japanese Dictation Test

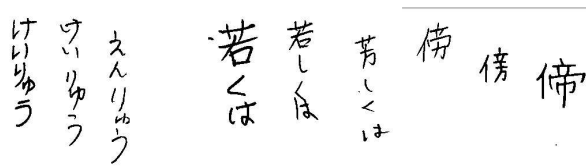
Meng Shi, Wataru Ohyama, Tetsushi Wakabayashi, and Fumitaka Kimura
Mie University, 1515 Kamihama, Tsu, 514-8507, Japan,
{meng, ohyama, waka, kimura}@hi.info.mie-u.ac.jp

Abstract

This paper studies on the decision rule of the automatic grading system for Japanese dictation test which aims to certificate the reading and writing ability of Japanese Kanji and Hiragana. Different from general handwritten character recognition systems which allow to read incorrect (mis-spelled) characters, the grading system is required to discriminate miswritten characters more strictly from correct ones. This paper introduces the core processing stages of the system and focuses on the decision rule of likelihood approach. Two threshold determination methods have been presented and comparatively evaluated. A grading accuracy of error rate less than 0.04% with rejection rate less than 70% is achieved in the performance test.



(a) Example of “right” answers (Six types of questions)



(b) Example of examinee’s answers

1 Introduction

There are more than 1.5 million persons who take the Japanese dictation test in Japan every year [1]. Currently, the grading processes rely on human markers to check the answer given by the examinee. This manual process is labor intensive and has the possibility of incorrect judgement. The proposed automatic grading system is designed to replace the conventional human grading process.

As shown in Fig.1, the examinees frequently give quite confusing answers which are difficult to judge if they are right or wrong. In general character recognition, an input character is assumed to belong one of the pre-defined categories, and some extent of miswriting should be rather allowable. On the other hand the grading system is required to discriminate miswritten characters more strictly from correct ones. The automatic grading system consists of pre-processing, feature extraction and grading. Likelihood approach is employed in grading process rather than general character recognition approach.

Figure 1. Examples of various answers

2 Feature extraction

For a long character string (especially in the case of mixed KANJI-Hiragana string), the probability of rejection increases rapidly*¹ if each component character is segmented and graded one by one. To avoid this problem this system uses the entire string image as an input just like a single character without character segmentation. The most typical string length is one or two, and is at most five or six.

The region which encloses a character string is firstly detected by eliminating the rows and columns which contain less than specified number of black pixels (less than four) from the margin of the prespecified box.

The gradient feature vector [2] of size 392 is extracted from a binary image the size of which is normalized to 64×64. After calculating the eigenvectors of the total covariance matrix of the learning sample, the size of the fea-

¹If the probability of rejecting a single character is 20%, the probability of rejecting a string with 5 characters is $(1 - (1 - 0.2)^5) = 67\%$.

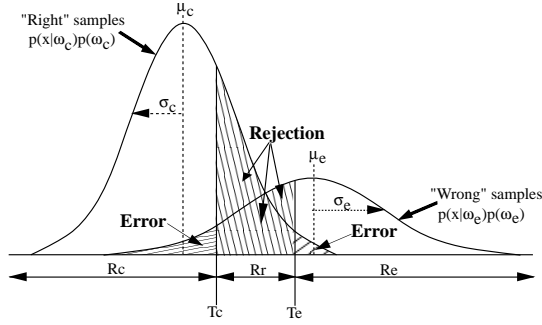


Figure 2. Decision rule based on likelihood approach

ture vectors are reduced to 256 using PCA (Principal component analysis).

3 Grading process

3.1 Distance function

For each question, a human-marked “right” reference set including about 1,000 samples is selected from the actual answer sheet. The mean vector M and covariance matrix Σ are estimated according to the reference set. The distance between an entry X and the reference set is calculated by the Modified Projection Distance Function [3] defined by

$$\begin{aligned} d_X^2 &= MPDF^2(X) \\ &= \|X - M\|^2 - \sum_{i=1}^k \frac{\lambda_i}{\lambda_i + \sigma^2} \{\Phi_i(X - M)\}^2 \end{aligned} \quad (1)$$

where λ_i and Φ_i are the i -th eigenvalue and eigenvector of Σ and k is the number of the dominant eigenvectors ($k < n$) respectively.

3.2 Decision rule

The decision rule based on likelihood approach [4] are defined as following :

$$\begin{aligned} (d_X < T_c) &: \text{Right} \\ (T_c \leq d_X \leq T_e) &: \text{Reject} \\ (d_X > T_e) &: \text{Wrong} \end{aligned} \quad (2)$$

4 Auto-Determination of the thresholds

As shown in Fig. 2, T_c and T_e divide the space of d_X into three regions, “Right”, “Reject” and “Wrong”. The

rejection probability $P(reject)$ and the error probability $P(error)$ are given by :

$$P(reject) = \int_{R_r} p(x)dx \quad (3)$$

$$\begin{aligned} P(error) &= P_c(error) + P_e(error) \\ &= \int_{R_c} p(x|\omega_c)p(\omega_c)dx \\ &\quad + \int_{R_e} p(x|\omega_e)p(\omega_e)dx \end{aligned} \quad (4)$$

The auto-grading system is required to minimize $P(reject)$ when $P(error)$ is less than a specified threshold E_t . For this purpose, two error-probability-override threshold determination are comparatively tested.

4.1 Method A: by approximation in one-dimensional distance space

Approximating the distributions in one-dimensional distance space by two univariate normal densities, the thresholds of T_c and T_e which satisfy Eq. (5) and Eq. (6) can be obtained by the inverse of the cumulative normal density function.

$$\int_{-\infty}^{T_c} P(x|\omega_e)P(\omega_e)dx < \alpha E_t \quad (5)$$

$$\int_{T_e}^{\infty} P(x|\omega_c)P(\omega_c)dx < (1 - \alpha)E_t \quad (6)$$

where α is a weight coefficient for tuning the balance of $P_c(error)$ and $P_e(error)$.

4.2 Method B: by approximation in multi-dimensional feature space

Since the parameters of probability density are only estimated accounting to limited “right” samples, a discrepancy will appear between the reference set and unknown samples (In fact, the distribution of “right” samples’ distance should be approximate by the χ^2 -distribution). While the univariate normal density is an appropriate model for the “wrong” samples’s distance distribution, in the situation of the “right” samples, the distribution of “right” samples can be approximated by multivariate normal density $N(M, \Sigma)$ in the multi-dimensional feature space. Substituting X_β to Eq. (2), the threshold of T_e can be obtained by:

$$\begin{cases} T_e = MPDF(X_\beta) \\ X_\beta = (m_1 + \beta\sqrt{\lambda_1}, m_2 + \beta\sqrt{\lambda_2}, \dots, m_k + \beta\sqrt{\lambda_k}) \end{cases} \quad (7)$$

where β is a tunable parameter (generally is set to enough large value, e.g. $\beta > 2$.) to specify how many Σ is the confusing samples apart from the reference set.

5 Experimental results

The proposed decision rule and threshold determination methods has been evaluated by a database collected from the past tests. The database consists of 24 subsets and includes total of about 10,000 answers for each question. The subsets 1-3 (including 1000 “right” answers) were used as learning samples and the remaining subsets 4-24 (8655 answers) were used as test samples in the experiments.

5.1 Comparison of threshold determination methods

The relationship between “Error Rate” (P_e) and “Rejection Rate” (P_r) are first comparatively evaluated against the “Hiragana string” questions for choosing the threshold determination method. The sample images and their distribution are listed in Table 1.

Table 2 shows that while method A and B give similar T_c , Method A gives less reliable of T_e for most questions, because the distribution of “right” samples approximated by univariate normal density has so small variance in one-dimension that the threshold is sensitive if it is determined by cumulative probability density.

Fig. 3 shows the comparison of the relationship of P_e and P_r between learning samples and test samples. The thresholds determined by method B made decisions for the test samples closer to those for the learning samples.

5.2 Performance test

From all above, the threshold determination method B was chosen for grading. The system performance has been evaluated by test samples for all types of questions.

As shown in Fig. 4, a reasonable accuracy of less than 0.04% error with less than 70% rejection was archived by proposed method.

5.3 Analysis of grading error

Fig.5(a) and (b) show two principal reasons of grading error:

- (a) the failure of the pre-processing (“right” answers are marked as “wrong”);
- (b) the incapability of detail analysis (“wrong” answers are marked as “right”).

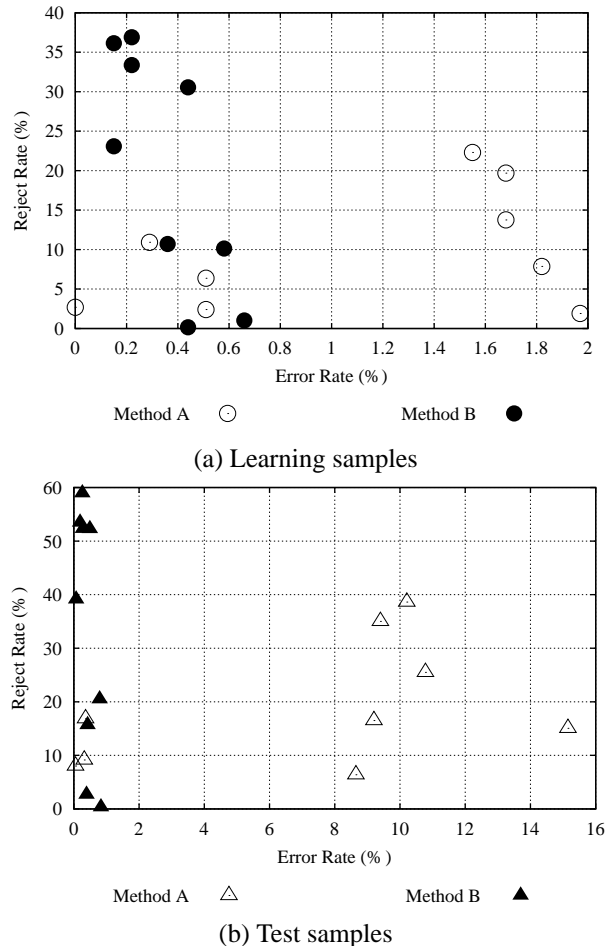


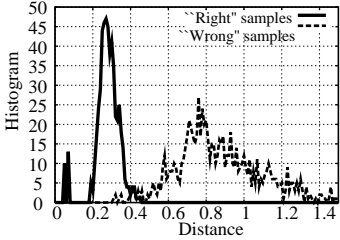
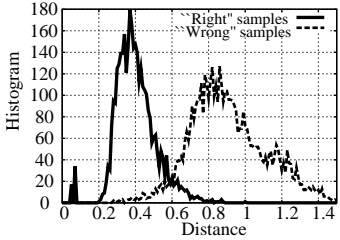
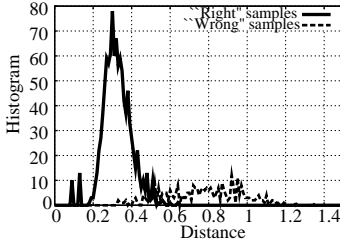
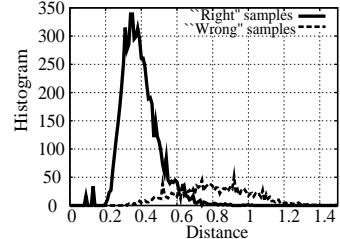
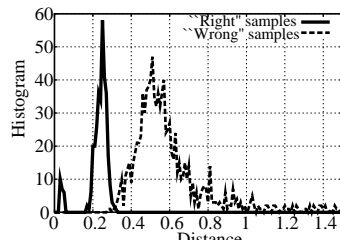
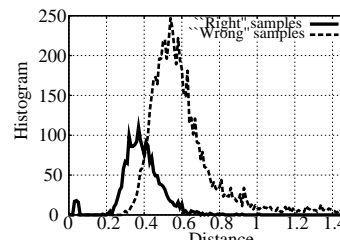
Figure 3. Comparison of P_e - P_r between two threshold determination methods

6 Conclusions

In this paper, a decision rule with two threshold determination methods are proposed for the automatic grading system of Japanese dictation test. An encouraging grading accuracy of less than 0.04% error with less than 70% rejection rate was achieved by the system. Current target performance required by the institution i.e. 0.002% error with 90% rejection was also nearly achieved. The proposed method can be also applied to another decision problem such as “Signature verification”, etc.

Future studies on 1) learning of the “wrong” samples in addition to the “right” samples, i.e. conversion from one class learning to two-classes learning, 2) improvement of pre-processing such as noise elimination based on the connected component analysis, 3) development of user interface, etc., are remaining to develop practical automatic grading system.

Table 1. Samples used for comparing the threshold determination methods

| Sample Image | Distribution of Learning samples | Distribution of Test samples |
|----------------|--|---|
| No.1 しゅっすい |  |  |
| No.2 いんじゆん |  |  |
| No.3 ついでしやう |  |  |

References

- [1] <http://www.kanken.or.jp>
- [2] F. Kimura, T. Wakabayashi, S. Tsuruoka and Y. Miyake, "Improvement of Handwritten Japanese Character Recognition Using Weighted Direction Code Histogram," Pattern Recognition, vol.30, no.8, pp.1329-1337, 1997.
- [3] T. Fukumoto, T. Wakabayashi, F. Kimura and Y. Miyake, "Accuracy Improvement of Handwritten Character Recognition by GLVQ," Proc. of 7th International Workshop on Frontiers in handwriting recognition (IWFHR-VII), pp.271-280, Amsterdam, The Netherlands, 2000.
- [4] K. Fukunaga, "Introduction to statistical pattern recognition", second Edition. Academic press, New York (1990).

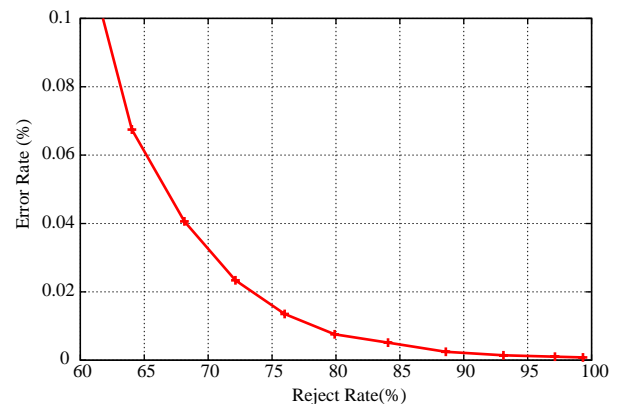
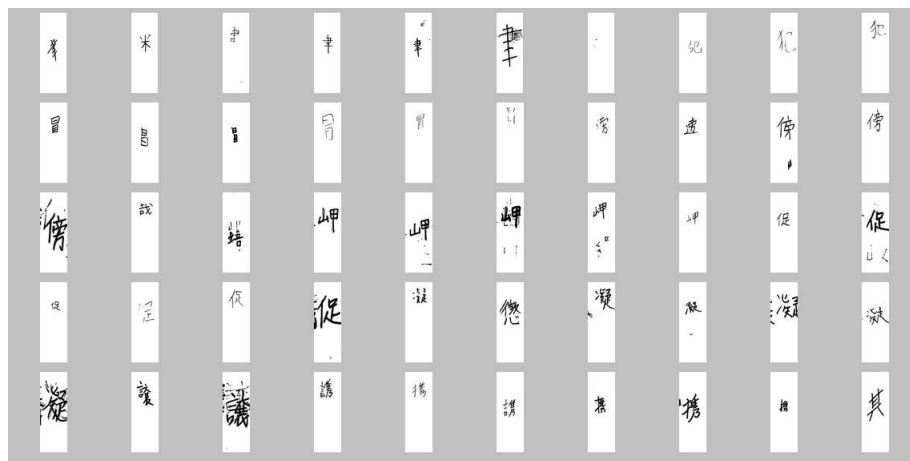


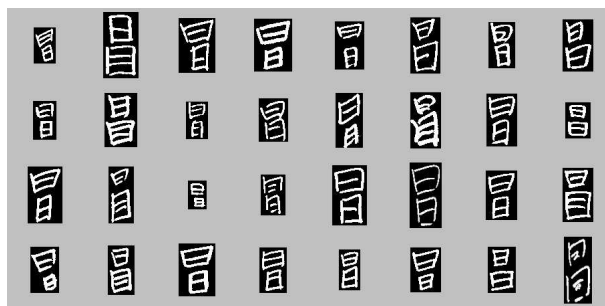
Figure 4. $P_e - P_r$ for test samples

Table 2. Estimated μ , σ thresholds and P_e - P_r

| No. | "Right" | | "Wrong" | | Method A ($\alpha = 1.0$) | | | | Method B ($\alpha = 1.0, \beta = 2.4$) | | | |
|-----|---------|----------|---------|----------|-----------------------------|-------|-----------|-----------|--|-------|-----------|-----------|
| | μ | σ | μ | σ | T_c | T_e | $P_e(\%)$ | $P_r(\%)$ | T_c | T_e | $P_e(\%)$ | $P_r(\%)$ |
| 1 | 0.28 | 0.06 | 0.87 | 0.22 | 0.34 | 0.43 | 0.51 | 6.37 | 0.34 | 0.84 | 0.19 | 53.47 |
| 2 | 0.33 | 0.07 | 0.78 | 0.20 | 0.37 | 0.51 | 1.68 | 19.68 | 0.37 | 0.79 | 0.49 | 52.26 |
| 3 | 0.23 | 0.05 | 0.60 | 0.20 | 0.10 | 0.35 | 0.00 | 24.52 | 0.10 | 0.87 | 0.03 | 89.95 |



(a) Failure of the pre-processing



(b) Incapability of detail analysis

Figure 5. Example of grading errors