Recognition with Supplementary Information
—How Many Bits Are Lacking for 100% Recognition?—

Masakazu IWAMURA*, Seiichi UCHIDA†, Shinichiro OMACHI‡, and Koichi KISE*
*{masa,kise}@cs.osakafu-u.ac.jp, Osaka Prefecture University, Sakai, Osaka, Japan
†uchida@is.kyushu-u.ac.jp, Kyushu University, Fukuoka, Japan
‡machi@aso.ecei.tohoku.ac.jp, Tohoku University, Sendai, Japan

Abstract

In this paper, we propose a new model in which the classifier receives not only a pattern itself but also supplementary information that assists recognition. This model enables us to achieve a 100% recognition rate with a 0% rejection rate with certain bits of supplementary information required. For printed characters, experiments show that 4 bits of supplementary information were required in the leave-one-out method and 1 bit was in the resubstitution method. In addition, we generalize the discussion into the relationship among a quantity of supplementary information, a recognition rate and a rejection rate. The theory presented in this paper is applied to the data embedding of a font set for camera-based character recognition [9].

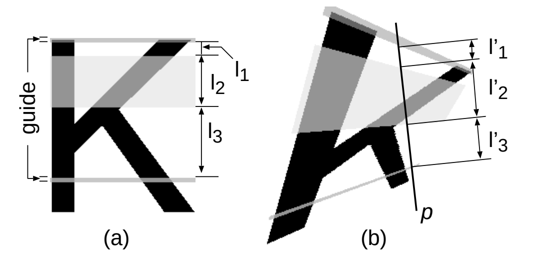


Figure 1. A data embedding method with cross ratio [9]. (a) A character image "K" where a cross ratio pattern is embedded. (b) Projective distortion.

1. Introduction

Let us imagine a task of recovering text data from printed papers without errors. It is not a good idea to employ an OCR undoubtingly since it cannot avoid causing recognition errors. How about using an extra media such as a 2D barcode for recording the whole text data? It is favorable regarding reading errors. However we cannot accept this solution because the original page is spoiled if a large barcode is next to the text (see the last page of this paper and the simulation in Sec. 5.2). It is just a great annoyance to us human beings though it is meaningful for computers.

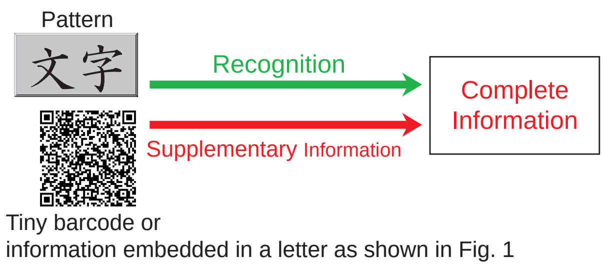
Are there any solutions between these two extremes? It is true that the computer receives a certain amount of information from the recognition results of an OCR. Thus only the remaining information should be required for the retrieving of full text information. In this paper, such pattern recognition is considered.

In the setting of the problem, we have two channels: the rst one is noisy (such as an OCR with recognition errors), and the second one is noiseless (such as a barcode). The problem here is how to design the second channel under the constraint that the amount of information sent through the second channel is minimized. We call the information on the second channel "supplementary information."

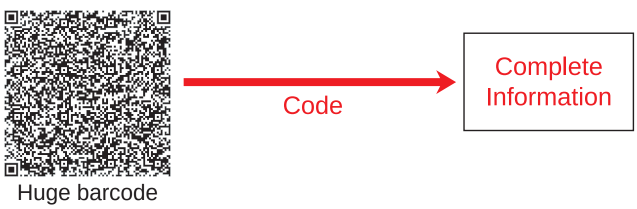
As an application of the supplementary information, we attempt a data embedding method with the cross ratio as in Fig. 1 for the task of document analysis and recognition with camera captured images [9].

We now turn to the real subject: how many bits of the supplementary information are required for a 100% recognition rate? We do not take rejection into account here. Let N be the number of classes. In the proposed model, K kinds of symbols, $K \leq N$, are assigned to the classes as the supplementary information. The K symbols has $\log_2 K$ bits of information. If N kinds of symbols are assigned, a 100% recognition rate with a 0% rejection rate (hereafter we call this "the state of 100% & 0%") is obviously achieved. However, $K = N$ is not always required since it is the same as transmitting a code without recognition as in Fig. 2(c). It is obvious that the better classi er performance becomes, the more K decreases. Thus quantity of information required to achieve "the state of 100% & 0%" is considered to depend

(a) Conventional recognition model.



(b) Proposed model.



(c) Transmitting model of character code.

Figure 2. Recognition and data transmitting models.

The True Class \ Recognition Result	A	B	C	D	E
A	0.6		0.4		
B		0.8		0.1	0.1
C	0.1		0.9		
D		0.1		0.8	0.1
E	0.2	0.1			0.7

Figure 3. An example of a probabilistic confusion matrix. Empty elements represent 0.

on the performance of the classi er .

In this paper, we rst consider the condition that such supplementary information has to satisfy for achieving "the state of 100% & 0%" in Sec. 3. As a table which represents recognition performance of a classi er , we use a confusion matrix (CM) described in Sec. 2. For simplicity, supplementary information is assumed to be retrieved without errors.

As a similar idea of the supplementary information, there is the assist channel coding which is the key idea to improve recognition performance of an OCR with a scanner [5, 3, 2, 4]. The way of building the code is similar to that of assigning the supplementary information in Sec. 3.

Next, we generalize the discussion into the relationship among a quantity of supplementary information, a recognition rate and a rejection rate in Sec. 4. One of the most important points for a real use is the rejection technique. Therefore we also take rejection into account. There is also a case that "how much supplementary information is required to achieve a recognition rate higher than 95%?" Therefore we should consider not only a 100% recognition rate. As the quantity of supplementary information changes from 0 bits to $\log_2 N$ bits, the recognition rate and the rejection rate change. We investigate two kinds of relationships that (1) a quantity of supplementary information and a recognition rate without rejection, and (2) a quantity of supplementary information and a rejection rate with a 100% recognition rate. They are very useful to design a classi er with supplementary information.

In the experiments in Sec. 5, the relationship between quantity of supplementary information and recognition performance are investigated using real CMs, and a task to input a page of text into a computer without errors is simulated.

2. Confusion matrix and its probabilistic expression

2.1. Confusion matrix

A CM is a matrix representing the correspondence between true classes and recognition results. Let $\boldsymbol{C}$ be a CM of $N \times N$ matrix. Usually, the (i, j) element of $\boldsymbol{C}$ represents the number of occurrences where patterns of a class ω_i are recognized as those of a class ω_j.

2.2. Probabilistic confusion matrix

Let $\boldsymbol{W}$ be an $N \times N$ matrix whose (i, j) element represents the probability where a pattern of a class ω_i is recognized as one of a class ω_j, that is $P(\omega_j|\omega_i)$. The (i, j) element of $\boldsymbol{W}$ is calculated as $w_{ij} = \frac{c_{ij}}{C_i}$, where $C_i = \sum_{j=1}^{N} c_{ij}$ as in Fig. 3.

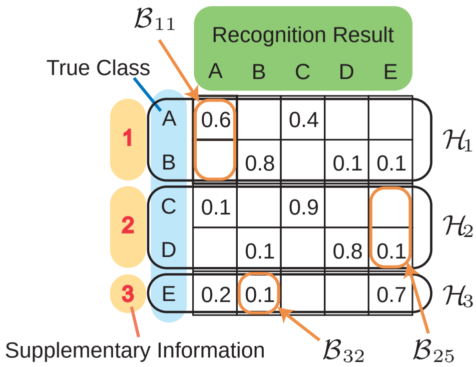


Figure 4. Symbols that achieve a 100% recognition rate with a 0% rejection rate. # of symbols:3, Recog. rate:100%, Reject. rate:0%.

3. Supplementary Information that Achieves 100% Recognition Rate with 0% Rejection Rate

The proposed model in Fig. 2(b) can achieve a 100% recognition rate with a 0% rejection rate. The condition such supplementary information has to satisfy is derived and formulated with a graph. Hereafter the proposed classi er with supplementary information has the codebook of symbols, and the CM of probabilistic expression, that is the matrix $\boldsymbol{W}$. For simplicity, a priori probability of each class is assumed to be equiprobability. Namely, $P(\omega_i) = \frac{1}{N}$ is asumed.

3.1. Partition of matrix $\boldsymbol{W}$

In the proposed model, symbols are assigned to all the rows in the matrix $\boldsymbol{W}$. A combination of rows to which the k-th symbol is assigned is de ned as

$$\mathcal{H}_k = \left\{ l | l = l_1, \ldots, l_{|\mathcal{H}_k|} \right\}, \tag{1}$$

where $|\mathcal{H}_k|$ is the number of rows to which the k-th symbol is assigned. For example, $\mathcal{H}_1 = \{1, 2\}$, $\mathcal{H}_2 = \{3, 4\}$ and $\mathcal{H}_3 = \{5\}$ in Fig. 4. Note that rows in $\mathcal{H}_k$ are not necessarily continuous.

Next, $\mathcal{H}_k$ is partitioned into each column. A combination of elements in $\mathcal{H}_k$ and in the j-th column is de ne d as

$$\mathcal{B}_{kj} = \left\{ (l, j) \left| l = l_1, \ldots, l_{|\mathcal{H}_k|} \right. \right\}. \tag{2}$$

For example, $\mathcal{B}_{11} = \{(1, 1), (2, 1)\}$, $\mathcal{B}_{25} = \{(3, 5), (4, 5)\}$ and $\mathcal{B}_{32} = \{(5, 2)\}$ in Fig. 4.

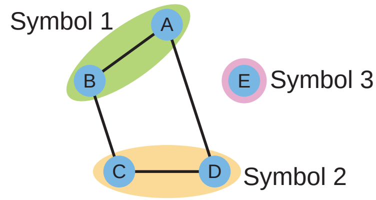


Figure 5. The division of the graph G which corresponds to Fig. 4. A symbol is assigned to each complete graph.

3.2. Condition that supplementary information has to satisfy

With Fig. 3, we consider the condition that the supplementary information for “the state of 100% & 0%” has to satisfy. The gure shows that if a recognition result is the class A, the true class can be either the class A, C or E; the true class cannot be determined by the classi er . If the classi er outputs the class A, it will cause misclassi cation when the true class is either a class C or E. Therefore, supplementary information is required to distinguish the three classes. This means that at least three symbols are required here.

Similarly, if a recognition result is the class B, the true class can be either the class B, D or E. Thus, different three symbols are also required. Consequently, “the state of 100% & 0%” can be achieved when different three symbols are assigned to either “A and B,” “C and D,” and “E” as in Fig. 4 or “A and D,” “B and C,” and “E.” Therefore the condition of the supplementary information which achieves “the state of 100% & 0%” is that “for all k and j, $\mathcal{B}_{kj}$ have less than two nonzero elements.”

3.3 Problem of finding the smallest supplementary information

The problem of nding the smallest quantity of supplementary information is formulated with a graph as in Fig. 5. Let $G = (V, E)$ be a graph where V is a set which consists of N nodes each of which represents a class, and E is a set which consists of edges each of which links two nodes. An edge between two nodes is created if the same symbol can be assigned to the two corresponding classes without misclassi cation. After all such edges are created, the graph G is divided into complete graphs. A symbol is assigned to each complete graph. Thus the problem is to nd a division of the graph G where the number of complete graphs is the smallest. The problem formulated here is a minimization version of PARTITION INTO CLIQUES [1], which is NP-hard.

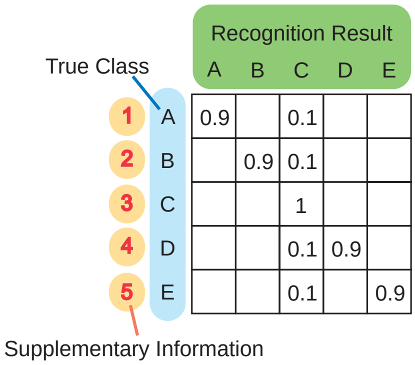


Figure 6. An example of a probabilistic confusion matrix which includes the class into which patterns are often misclassified. The samples of all classes can be recognized as those of the class C.

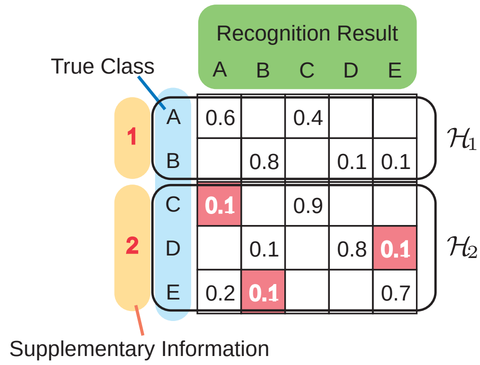


Figure 7. Symbols that accept recognition errors without rejection. The reverse colored elements are the causes of misclassification. # of symbols:2, Misclassi. Rate:0.3/5=6%, Recog. rate:94%, Reject. rate:0%.

Let N_{symmin} be the smallest number of complete graphs. Then, N_{symmin} symbols, which is $\log_2 N_{\text{symmin}}$ bits of information, can achieve "the state of 100% & 0%." N_{symmin} is determined by the largest number of non-zero elements in a column in the CM. To achieve "the state of 100% & 0%," no misclassi cation is permitted even if a classi cation result corresponds to many possible true classes according to the CM.

As the evaluation method of classi ers, the quantity of supplementary information has a different nature from the recognition rate. For example, in the case of the CM in Fig. 3, $N_{\text{symmin}} = 3$ and the recognition rate is 76%. In the case of the CM in Fig. 6, $N_{\text{symmin}} = 5$ and the recognition rate is 92%. This shows that a CM with higher recognition rate do not always require less information to achieve "the state of 100% & 0%." This is con rm ed in the experiment in Sec. 5.1.

4. Relationship Between Quantity of Supplementary Information and Recognition Performance

In Sec. 3, we discussed "the state of 100% & 0%." In this section, we generalize the discussion into the relationship between quantity of supplementary information and recognition performance. Actually, a quantity of supplementary information, a recognition rate and a rejection rate have a trade-off relationship where if two are determined, the rest is automatically determined. Here we focus on (1) relationship between a quantity of supplementary information and a recognition rate without rejection in Sec. 4.2, and (2) relationship between a quantity of supplementary information and a rejection rate with a 100% recognition rate in Sec. 4.3.

4.1. Number of nonzero elements in $\mathcal{B}_{kj}$

To generalize the discussion in Sec. 3, a function q_{kj} which returns the number of nonzero elements in $\mathcal{B}_{kj}$ is de ned. First, a function $z(x)$ is de ned as

$$z(x) = \begin{cases} 0, & \text{for} \quad x = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

Then, the function q_{kj} is de ned as

$$q_{kj} = \sum_{(l,j)\in\mathcal{B}_{kj}} z(w_{lj}). \tag{4}$$

4.2. Relationship between quantity of supplementary information and recognition rate without rejection

As mentioned in Sec. 3, if $\mathcal{B}_{kj}$ for all k and j has less than two nonzero elements, that is $\forall k, j \; q_{kj} \leq 1$, recognition errors do not occur. On the other hand, if $\mathcal{B}_{kj}$ has two or more nonzero elements for some k and j , that is $\exists k, j \; q_{kj} \geq 2$, recognition errors occur. In the case that $q_{kj} \geq 2$, there are q_{kj} possible true classes. The best way to minimize the recognition error is to choose *the most feasible class*. For example, in Fig. 7, if the recognition result is the class A and the symbol is 2, the class E should be chosen. The above consideration leads the formula of the

Algorithm 1 A greedy algorithm.

1: Assign different symbols to all classes. Namely, $\mathcal{H}_k = \{k\}$, for $k = 1, \ldots, N$.

2: Let $L(k)$ be a loss function which is either the misclassi cation rate or the rejection rate. Obviously, $L(N) = 0$.

3: **for** $K = N - 1$ to 1 **do**

4: To calculate the loss $L(K)$, the number of symbols is decreased by one. First, choose a pair of sets of rows that is assigned the same symbol. In other words, choose $\mathcal{H}_s$ and $\mathcal{H}_t$ which satis es $s \neq t$ and $\mathcal{H}_s, \mathcal{H}_t \neq \emptyset$. Then, the same symbol is assigned to $\mathcal{H}_s$ and $\mathcal{H}_t$. Namely,

$$\mathcal{H}_s \leftarrow \mathcal{H}_s \cup \mathcal{H}_t \tag{6}$$

$$\mathcal{H}_t = \emptyset. \tag{7}$$

5: **end for**

misclassi cation rate R_{error} as

$$R_{\text{error}} = \frac{1}{N} \sum_j \sum_k \left\{ \sum_{(l,j) \in \mathcal{B}_{kj}} w_{lj} - \max_{(l,j) \in \mathcal{B}_{kj}} w_{lj} \right\}. \tag{5}$$

In Eq. (5), the rst term in the parentheses is the sum of the elements of the matrix $\boldsymbol{W}$ in $\mathcal{B}_{kj}$, and the second term is the corresponding elements to the class where the classi er outputs.

In this paper, to minimize the misclassi cation rate or the rejection rate, we utilize a greedy algorithm shown in Algorithm 1. In the algorithm, as the number of symbols K decreases from N by one, the misclassi cation rate or the rejection rate is calculated with K symbols. Note that in the case of $K = 1$, it is the same as the usual pattern recognition method without supplementary information. The transition of the symbols and the misclassi cation rate R_{error} when Algorithm 1 is applied to the CM in Fig. 3 is shown in Fig. 8. Figs. 4 and 7 are the cases when the numbers of symbols are 3 and 2, respectively.

4.3. Relationship between quantity of supplementary information and rejection rate with 100% recognition rate

As mentioned in Sec. 3 and Sec. 4.2, if $\mathcal{B}_{kj}$ for all k and j has less than two nonzero elements, that is $\forall k, j\ q_{kj} \leq 1$, recognition errors do not occur. In the case that $\exists k, j\ q_{kj} \geq 2$, in order to achieve a 100% of recognition rate, all the possible true classes have to be rejected because they all can cause recognition errors. For example, in Fig. 9, if the recognition result is the class A and the symbol is 2, both the classes C and E are rejected.

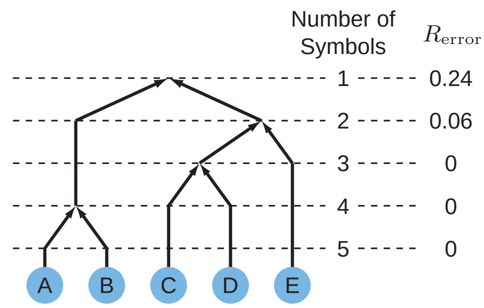


Figure 8. Transition of the symbols and the misclassification rate R_{error} with Algorithm. 1 for the confusion matrix in Fig. 3.

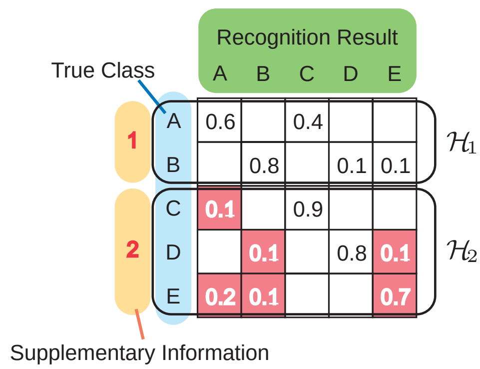


Figure 9. Symbols that achieve a 100% recognition rate by rejection. The reverse colored elements are the causes of rejection. # of symbols:2, Recog. rate:100%, Reject. rate:1.3/5=26%.

The above consideration leads the formula of the rejection rate R_{reject} as

$$R_{\text{reject}} = \frac{1}{N} \sum_j \sum_k s_{kj}, \tag{8}$$

where

$$s_{kj} = \begin{cases} 0, & \text{for} \quad q_{kj} \leq 1 \\ \sum_{(l,j) \in \mathcal{B}_{kj}} w_{lj}, & \text{otherwise.} \end{cases} \tag{9}$$

As in Sec. 4.2, the transition of the symbols and the rejection rate R_{reject} when Algorithm 1 is applied to the CM in Fig. 3 is shown in Fig. 10. Figs. 4 and 9 are the cases when the numbers of symbols are 3 and 2, respectively.

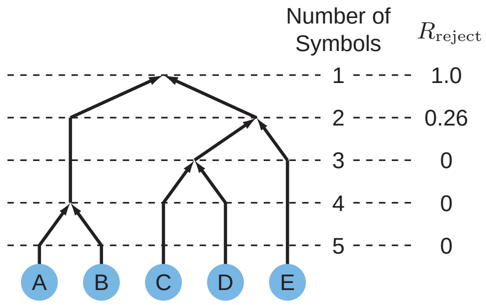


Figure 10. Transition of the symbols and the rejection rate R_{reject} with Algorithm. 1 for the confusion matrix in Fig. 3.

5. Experiments

5.1. Relationship between quantity of supplementary information and recognition performance

In the experiments, eight kinds of CMs which were created by recognition of real data sets were used. For the eight CMs, two kinds of data sets, two kinds of classifiers, and two kinds of recognition experiments, L and R methods, were combined. The L method is the leave-one-out method and the R method is the resubstitution method.

As the two kinds of data sets, handwritten and printed characters were used. Note that embedding supplementary information in handwritten characters is not easy; we used them just for the demonstration of the proposed method. As the handwritten data set, the ETL9B [7], which consists of 3036 Japanese characters and each character has 200 samples, was used. As the printed one, 25 fonts were used. The same 3036 characters as contained in ETL9B were extracted from the data set of printed characters. Each character image in both data sets was normalized nonlinearly [10] to fit in a 64×64 square, and the 196-dimensional directional element feature [8] was extracted. The Euclidean distance and the SQDF [6] were used in classifiers. The conditions of recognition experiments and the corresponding recognition rates are shown in Table 1.

For the eight CMs, relationship between the number of symbols (the quantity of supplementary information) and the recognition rate mentioned in Sec. 4.2, and that between the number of symbols (the quantity of supplementary information) and the rejection rate mentioned in Sec. 4.3 are shown in Figs. 11 and 12. From the figures, we can estimate how much information is required to achieve a certain recognition rate or rejection rate. This information is useful to design a classifier in certain performance. Note that the values in the figures are not the best values and possible recognition performance may be better.

In Table 1, the quantity of supplementary information required to achieve "the state of 100% & 0%" is also shown. The uncertainty of 3036-class problem is $\log_2 3036 \sim 11.57$ bits. For printed characters, the experiments show that 4 bits were required in the leave-one-out (L) method and 1 bit was in the resubstitution (R) method. Though the greedy algorithm does not guaranteed to give the ideal value, quantity of supplementary information required to achieve "the state of 100% & 0%" in the table was confirmed to be the same as the ideal one.

By comparison with the Euclidean distance, the SQDF had an advantage of the recognition rate. However, it did not always have an advantage of the quantity of supplementary information required to "the state of 100% & 0%." This is due to a few classes into which patterns are often misclassified as mentioned in Sec. 3.3. Therefore, a classifier which minimizes the quantity of supplementary information should be developed.

5.2. Comparison to transmitting character code

We have explained that the proposed model in Fig. 2(b) combines advantages of the conventional approach of pattern recognition in Fig. 2(a) and transmitting identification codes in Fig. 2(c). The proposed model is valuable when patterns are available but recognition errors occur. For example, let us assume a task to input a page of text into a computer without errors. If you have a media that can contain all the information, the problem is solved. However, when you use a printable media such as a 2D barcode, you will realize another problem occurs: it requires large area. On the other hand, in the proposed model a smaller 2D barcode can reduce the area. This is confirmed by a simulation using the QR code (ISO/IEC 18004). We assumed a page contains 1000 Japanese characters in the simulation.

There are 40 variations of the QR code which differ in printed area and containable data size. Since 12 bits are required for a code of 3036 categories,

$$12(\text{bits}) \times 1000(\text{characters})/8 = 1500(\text{bytes})$$

of supplementary information are required per page. On the other hand, since 1 bit is required for the proposed method,

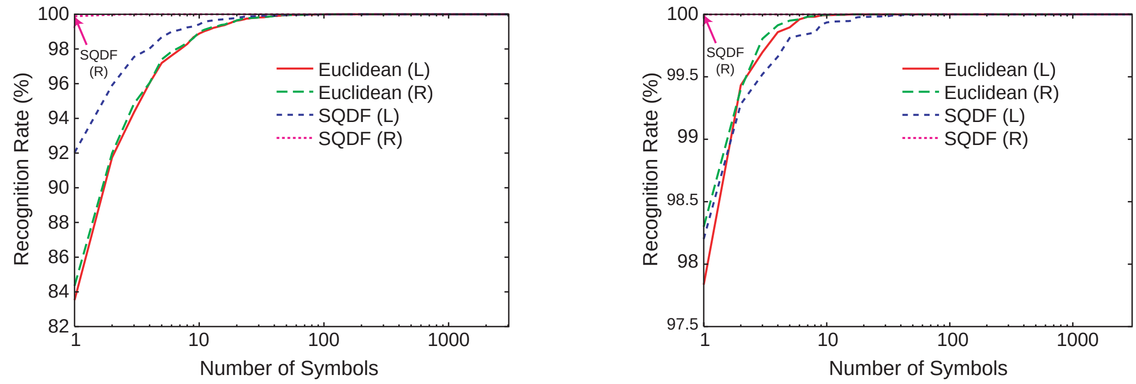
$$1(\text{bit}) \times 1000(\text{characters})/8 = 125(\text{bytes})$$

are required per page.

In the binary mode with the error correction level M, the version 8 and 32 are the smallest ones whose capacities are more than 125 and 1500 bytes, respectively. The QR codes of version 8 and 32 are shown in Fig. 13. In comparison to version 32, the QR code of version 8 is about one third in

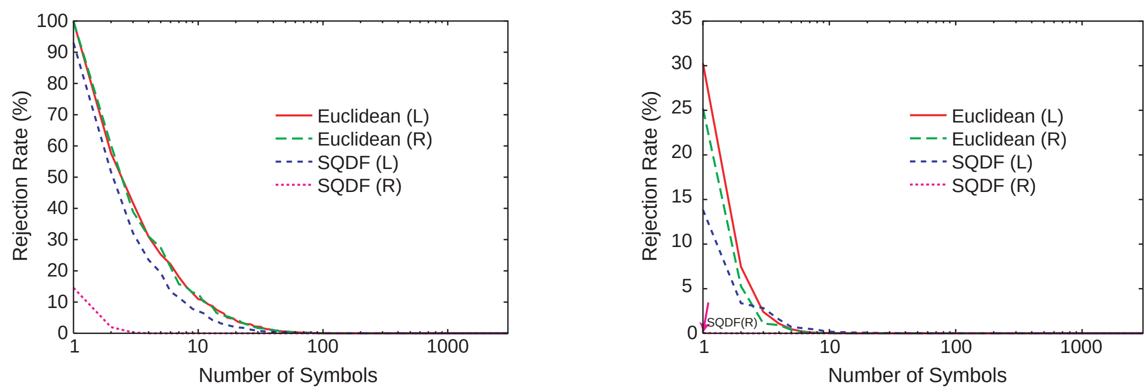
Table 1. Conditions of recognition experiments, recognition rates and quantity of information required to achieve a 100% recognition rate with a 100% recognition rate (“the state of 100% & 0%”).

Data set	Distance in Classi er	L / R method	Recognition rate (%)	Quantity of supple. info. required to “the state of 100% & 0%” (Bits) (# of symbols required to “the state of 100% & 0%” in parentheses)
Handwritten Characters	Euclidean	L	83.53	7.69 (206)
		R	84.35	7.69 (206)
	SQDF	L	92.03	8.84 (459)
		R	99.89	2.00 (4)
Printed Characters	Euclidean	L	97.84	4.00 (16)
		R	98.30	3.91 (15)
	SQDF	L	98.20	6.15 (71)
		R	99.99	1.00 (2)



(a) Handwritten characters. (b) Printed characters.

Figure 11. Relationship between the number of symbols and the recognition rate without rejection.



(a) Handwritten characters. (b) Printed characters.

Figure 12. Relationship between the number of symbols and the rejection rate with a 100% recognition rate.

(a) Version 32. This equals the character code.

(b) Version 8. This equals the supplementary information of the proposed method.

Figure 13. QR code required to represent a page of Japanese text.

width and height, and about one ninth in area. It is demonstrated that the proposed model can reduce the quantity of information required. Consequently it is confirmed that the proposed model is efficient in such situations in comparison to a barcode containing the whole text information as in Fig. 2(c).

6. Conclusion

In this paper, we proposed a new model that the classifier receives not only a pattern itself but also supplementary information that assists recognition. In the model, we can achieve a 100% recognition rate with a 0% rejection rate ("the state of 100% & 0%"). For printed characters, experiments showed that 4 bits were required in the leave-one-out method and 1 bit was in the resubstitution method.

For the eight real CMs, (1) relationship between a quantity of supplementary information and a recognition rate, and (2) that between a quantity of supplementary information and a rejection rate were observed. These information is useful to design a classifier using supplementary information in certain performance.

Furthermore, we demonstrated that the quantity of supplementary information required to achieve "the state of 100% & 0%" has a different nature from the recognition rate; the quantity of supplementary information is a new evaluation criterion of classifiers. In this sense, we have to design a classifier which minimizes the supplementary information.

Acknowledgement Thanks are due to an anonymous reviewer for valuable comments on the assist channel coding.

References

[1] M. R. Garey and D. S. Johnson. *Computers and Intractability*. W. H. Freeman and Company, New York, 1979.

[2] D. H. Greene and E. H. Kuo. Assist channel coding with vertical block error correction. United States Patent 6,768,560.

[3] D. H. Greene, L. T. Niles, and E. H. Kuo. Assist channel coding with character classifications. United States Patent 6,862,113.

[4] D. H. Greene and A. C. Popat. Assist channel coding with convolution coding. United States Patent 6,628,837.

[5] E. H. Kuo. Assist channel coding for improving optical character recognition. Master's thesis, MIT, 2000.

[6] S. Omachi, F. Sun, and H. Aso. A new approximation method of the quadratic discriminant function. *Lecture Notes in Computer Science*, 1876:601–610, Sept. 2000.

[7] T. Saito, H. Yamada, and K. Yamamoto. On the data base ETL9 of handprinted characters in JIS Chinese characters and its analysis. *Trans. IEICE*, J68-D(4):757–764, 1985. Written in Japanese.

[8] N. Sun, Y. Uchiyama, H. Ichimura, H. Aso, and M. Kimura. Intelligent recognition of characters using associative matching technique. In *Proc. Pacific Rim Int'l Conf. Artificial Intelligence (PRICAI'90)*, pages 546–551, Nov. 1990.

[9] S. Uchida, M. Iwamura, S. Omachi, and K. Kise. Data embedding for camera-based character recognition. In *Proc. First International Workshop on Camera-Based Document Analysis and Recognition*, Aug. 2005.

[10] H. Yamada, K. Yamamoto, and T. Saito. A nonlinear normalization method for handprinted kanji character recognition — line density equalization —. *Pattern Recognition*, 23:1023–1029, 1990.