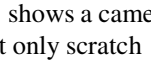
Robust Chinese Character Recognition by Selection of Binary-Based and Grayscale-Based Classifier

Yoshinobu Hotta[1], Jun Sun[2], Yutaka Katsuyama[1], and Satoshi Naoi[1]

[1] FUJITSU LABORATORIES LTD.,
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, 211-8588 Japan
{y.hotta, katsuyama, naoi.satoshi}@jp.fujitsu.com
[2] FUJITSU R&D CENTER Co., Ltd.,
Eagle Plaza B10th floor, Xiaoyun Rd No.26, Chaoyang Dist. Beijing, 10016, P.R. China
sunjun@frdc.fujitsu.com

Abstract. As the spread of digital videos, digital cameras, and camera phones, lots of researches are reported about degraded character recognition. It is found that while the grayscale-based classifier is powerful for degraded character, the performance for clear character is not so good as binary-based classifier. In this paper, a dynamic classifier selection method is proposed to combine the two classifiers based on an estimation of the degradation level and the recognition reliability of the input character images. Experimental results show that the proposed method can achieve better recognition performance than the two individual ones.

1 Introduction

As the spread of digital videos, digital cameras, and camera phones, lot of researches about degraded character recognition are reported [1]-[5]. However, there is a problem that the recognition accuracy of degraded character classifier for characters without degradation is lower than that of conventional classifier that utilize character contour of binary images as a feature vector. Although a general recognition method for characters with degradation or without degradation has been proposed in [1], it is uncertain that the method is effective or not for uncertain fonts because numbers of tested fonts in the paper is limited. In addition, the method is effective only for collapsed characters, not for scratched characters. In general, the contrast of a camera-captured image is lower than that of scanned images and the image is influenced by vibration of a user or so. Therefore collapse or scratch may occur simultaneously when certain binarization methods are applied. Figure 1 shows a camera captured image of "  " and the global binarization result. Not only scratch ("比" and "谷"), but also collapse ("線") are caused.

In this paper, a degraded character classifier is used to deal with grayscale images in recognizing low-resolution characters to avoid the above binarization problem. Meanwhile, characters without degradation are recognized by binary-based classifier in

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 553 – 563, 2006.

which contour direction of character strokes is used as the character feature (hereafter, "binary-based classifier"). To select one of these two classifiers, "degradation level" of input characters is newly defined, and each classifier is used properly based on the estimated degradation level. Some definitions of degradation are proposed [5][6] so far, but the binary image input is assumed and only single font is considered [5] or only alphabetic characters are considered [6].

Fig. 1. Camera captured string and the binarized images

In Section 2, both binary-based and grayscale-based classifiers are described. In Section 3, "degradation level" for grayscale character images is newly defined, and recognition accuracy of both classifiers according to the degradation level is shown. Selection of classifier is described in Section 4. Section 5 shows the experimental result. Finally, the conclusion is set out in Section 6.

Note that background of the characters is assumed to be uniform and input image is grayscale (0-255). Also each character is assumed to be cut out in advance.

2 Character Classifier

2.1 Grayscale-Based Character Classifier

The grayscale-based classifier for degraded character recognition is based on the degradation model [3][4]. First a binary character image is size-normalized to 32×32 pixel. Let black pixel be 1, and white pixel be 0. Next image decimating and zooming are executed to the image, thus 32×32 grayscale image is generated. Many types of degradation image can be generated by changing decimating size of the character. Algorithm of decimating is Supersampling and that of zooming is Cubic interpolation. Pixel values by raster scanned image with 1,024 dimensions are used as the character feature, x. Next, x is normalized to x' by definite canonicalization [7] in order to decrease the blurring influence. Let n be the dimension of the feature.

$$c = (1/\sqrt{n},...,1/\sqrt{n}) \tag{1}$$

$$x' = x - (c \cdot x)c \tag{2}$$

Then x' is regulated into unit length, x''. That is, $(x'', c) = 0$, $\| x'' \|^2 = 1$.

Feature selection by Principal Component Analysis (PCA) is first conducted to reduce computational complexity and memory cost and then similarity between the

input feature and the reference feature is calculated by the subspace method. Procedures are described as follows.

[Learning phase]

1) Construction of unitary eigenspace : the unitary eigenspace is constructed in a way similar to the traditional eigenspace-based method. The mean vector of every category is used to calculate the covariance matrix of unitary eigenspace. Suppose the character image with size *N*N* represent a vector, $x=[x_1, x_2,..., x_{N*N}]^T$, using the raster scanning order. The covariance matrix for the unitary eigenspace is calculated as:

$$COV_{sb} = \frac{1}{P}\sum_{i=1}^{P}(m_i - m)(m_i - m)^T , \tag{3}$$

where P is the number of category. m is the mean vector for all training samples. m_i is the mean vector for the i th category. The first n eigenvectors of matrix COV_{sb} corresponding to the first n biggest eigenvalues are recorded as: $U=[u_1, u_2,..., u_n]^T$, which spans the unitary eigenspace. Usually, the dimension of the unitary eigenspace is far lower than the dimension of original image for noise removal and data compression.

2) 1st feature extraction using the unitary eigenspace: the feature vector for the i th category is obtained by casting its mean vector to the unitary eigenspace:

$$c_i = U^T(m_i - m). \tag{4}$$

Conventional PCA based methods perform recognition in the above unitary eigenspace.

3) Individual eigenspace construction: the performance of conventional PCA method is not always satisfactory, since for many recognition tasks such as character recognition, the feature distribution for every category is not the same. Thus, an individual eigenspace is built for every category using the 1st feature of all the samples belonging to the same category. The auto-correlation matrix for the 1st feature of the i th category is:

$$W_i = \frac{1}{M_i}\sum_{j=1}^{M_i}\left(y_i^{(j)} - c_i\right)\left(y_i^{(j)} - c_i\right)^T , \ i=1,2,...,P \tag{5}$$

where $y_i^{(j)} = U^T(x_i^{(j)} - m)$ is the 1st feature vector of the j th training sample $x_i^{(j)}$ in the i th category, M_i is the number of training samples for the i th category. The first n_1 eigenvectors of W_i corresponding to the first n_1 eigenvalues are recorded as: $\tilde{U}_i = [u_1^i, u_2^i, \ldots, u_{n_1}^i]$, $i=1,2,...,P$, which spans the individual eigenspace for the j th category.

[Recognition phase]

1) 1st feature extraction: for a testing image f , the feature in the unitary eigenspace, y , is extracted using the unitary eigenspace U as $y = U^T(f - m)$.

2) Coarse classification using 1st feature: the coarse classification is performed by comparing the similarity with the 1st feature of the mean vector of every category, c_i, $i=1,2,...,P$ with the 1st feature of testing image. d candidate categories are generated as the result of coarse classification.
3) 1st feature reconstruction: reconstruct the 1st feature of image, f, using the individual eigenspace of the d categories from the coarse classification:

$$\eta_i = \tilde{U}_i^T (y - c_i), \tag{6}$$

$$\hat{y}_i = \tilde{U}_i^T \eta_i + c_i, \tag{7}$$

where η_i is the project coefficient of the 1st feature y on the i th individual eigenspace. $\hat{y}_i$ is the reconstructed feature of y.
4) Final classification using optimal reconstruction: for every of d candidate categories, the reconstruction error of the 1st feature is obtained as:

$$\varepsilon_i = \| y - \hat{y}_i \|. \tag{8}$$

The category of optimal reconstruction, that is, the minimum reconstruction error, ε, is selected as the final recognition result.

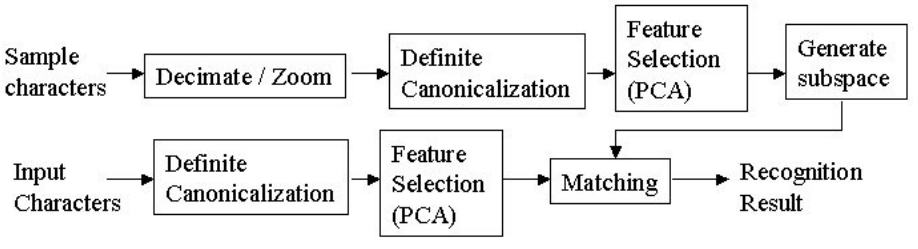


Fig. 2. Whole flow

2.2 Binary-Based Character Classifier

A conventional method is used as a binary-based character classifier. After a binary character image is inputted, size-normalization is conducted first, and contour direction of character stroke is extracted as feature vector. The feature dimension is set to 288 and the city-block distance is used in matching.

3 Degradation Level for Grayscale Image

3.1 Definition of Degradation Level

The degradation level for grayscale character images is defined here. It is difficult to calculate local information such as corner of character stroke when Chinese multifont characters are considered. Therefore, global information is used to evaluate the degradation level. The degradation level is calculated as follows.

1) Inputted grayscale character image is size-normalized to $N \times N$.
2) Definite canonicalization is conducted to the input feature and the feature is regulated into unit length.
3) The density of each pixel value is linearly transformed to 0-255.
4) Count gray pixel value (1-254) among all pixels and define degradation level as follows.

$$\text{Degradation level} = 100 \times (\text{total number of gray pixels}) / (N \times N) \quad (9)$$

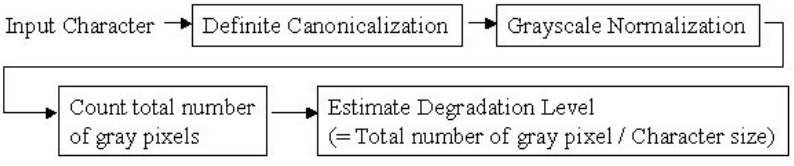


Fig. 3. Calculation flow of degradation level

A character in low quality, "愛 ", and its density histogram are shown in Figure 4. Vertical axis of the histogram means frequency.

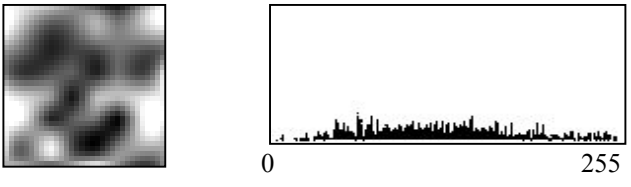


Fig. 4. A degraded character of "愛 " and its density histogram

As the degradation of a character becomes larger, the total number of grayscale pixels also increases. The degradation level here is regarded as a guidepost to evaluate the degradation. When the degradation is small, this value is almost zero regardless of fonts.

3.2 Recognition Accuracy of Each Classifier According to the Degradation Level

Experiments are conducted to investigate the relationship between the degradation level and the recognition accuracy of each classifier. When test data is inputted, recognition result and the degradation level by each classifier is recorded. After many test data are inputted, the recognition accuracy at each degradation level is calculated.

3.2.1 Learning Data

JIS first level kanji, totally 2,965 categories of 19 fonts are used as learning data. Each character image is size-normalized to 32×32 and it is decimated by 7 degrees, 8×8,

12×12 ,..., 32×32, respectively. Next, the decimated image is zoomed back to 32×32. Various fonts such as Mincho, Gothic, Round-gothic, Kaku-gothic etc. are included. The total number of characters used in learning is about 390,000 characters (=2,965×19×7). Figure 5 shows the example of generated degraded characters. The left end shows the image without degradation and the right end shows the image that is decimated to 8×8 and zoomed back to 32×32. On the other hand, the binary-based classifier is trained by clear character images only.

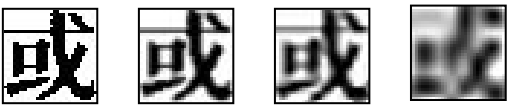


Fig. 5. Generated degraded characters by decimating / zooming

3.2.2 Test Data

JIS first level kanji of 6 fonts, which are different from those of learning data, are used as test data. 7 degrees of degraded images are generated as well. Various fonts such as Gona, Middle-Gothic, etc. are included. The total number of characters is about 120,000 (=2,965×6×7). Figure 6 shows 6 fonts images without degradation.

Fig. 6. Sample font images used for test

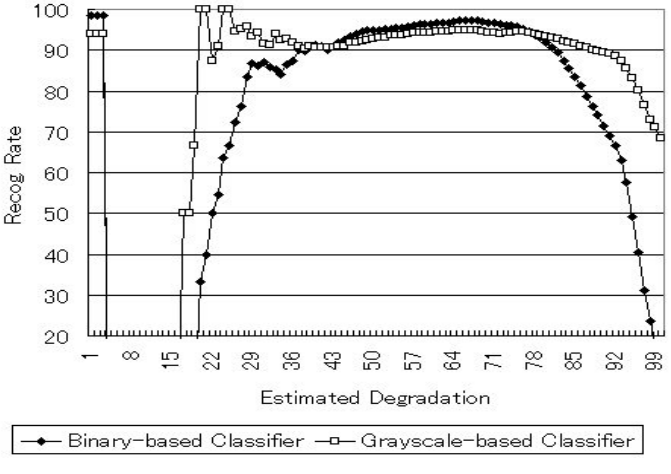


Fig. 7. Recognition accuracy of each classifier according to the degradation level

3.2.3 Experimental Result

Figure 7 shows the recognition accuracy of each classifier according to the degradation level. Horizontal axis means estimated degradation level and vertical axis means its recognition rate. When degradation level is very low (0-3), the recognition rate of binary-based classifier is higher than that of grayscale-based classifier. As for the level from 4 to 16, there is no character with this degradation. When the level is low (17-36), recognition rate of grayscale classifier is higher, but almost all cases are caused by simple shape characters such as " —", and these are special cases. The total number of characters with middle degradation level (37-78) is a lot, and recognition rate of binary-based classifier is higher at these levels. As the degradation level increases, the recognition rate of the binary-based classifier drops dramatically, whereas that of grayscale-based classifier is relatively stable.

Table 1. Overall recognition rate for the entire test dataset

	Recognition rate (%)
Binary-based classifier	85.1
Grayscale-based classifier	91.2

3.3 Reliability Distribution of Each Classifier According to Degradation Level

3.3.1 Experimental Data

Learning data are the same data as that in 3.2.1. The reliability distribution is examined by using part of the same data as that of 3.2.2, the data without degradation and the data with maximum degradation. The total number of characters for one set of data is 17,790 (=2,965×6). Hereafter, level0 shows the data without degradation and level6 shows the data with maximum degradation. Let the recognition distance of 1st candidate be *d1*, and that of 2nd candidate be *d2*, and reciprocals of them are *r1*, *r2*, respectively. We define recognition reliability, *r*, as $r = r1/(r1+r2)$.

3.3.2 Reliability Distribution of Binary-Based Classifier

In Figure 8-11, horizontal axis means standardized reliability of *r* and vertical axis means frequency. Black curve means the frequency of correctly recognized characters and white curve means that of misrecognized characters. In Figure 8, the distribution peaks of two graphs are apart whereas those in Figure 9 almost overlap. From these figures, it can be said that the recognition reliability of the binary-based classifier for degraded characters is unreliable.

3.3.3 Reliability Distribution of Grayscale-Based Classifier

Compared with above figures, the distribution peaks of correctly recognized data and misrecognized data (Figure 11) for level6 data are separated. It can be said that recognition reliability of the grayscale-based classifier for degraded characters is more reliable.

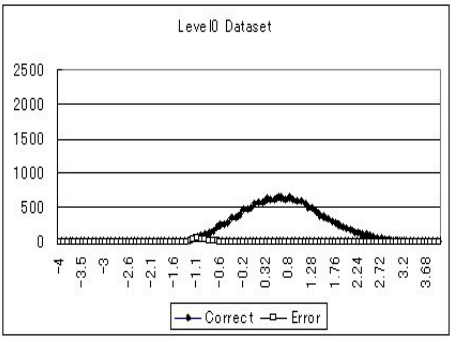


Fig. 8. Reliability distribution of level0 data

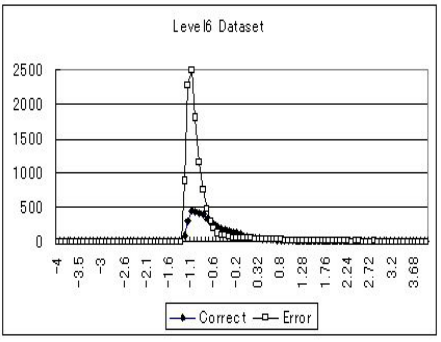


Fig. 9. Reliability distribution of level6 data

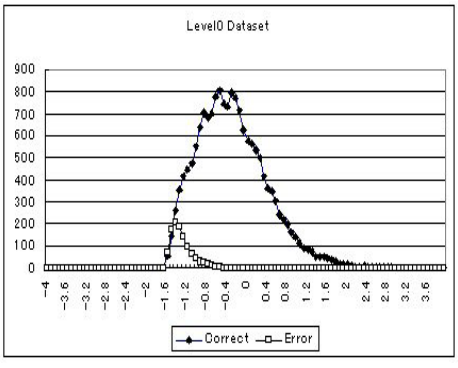


Fig. 10. Reliability distribution of level0 data

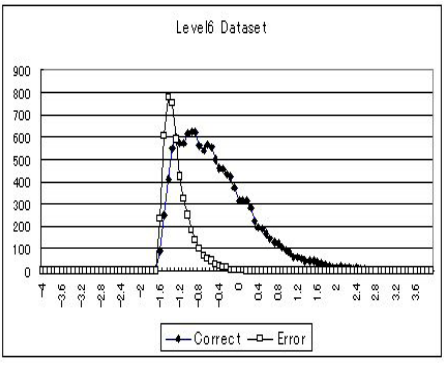


Fig. 11. Reliability distribution of level6 data

4 Selection of Classifier

Lots of researches about classifier combination are reported, in which recognition results or recognition reliabilities of multiple classifiers are used. But from the point of time complexity, it is desirable that fewer classifiers are used. Also recognition reliabilities are unreliable when degradation is large. Therefore we adopt the classifier selection method based on the degradation level of input characters in combining classifiers (Figure 12). First the degradation level of input character is estimated as described in 3.1. If the degradation level is larger than predetermined threshold (*Th1*), then grayscale-based classifier is used. And binary-based classifier is used for the characters with low/middle level degradation. But binary-based classifier sometimes misrecognizes simple-shaped characters shown in Figure 7, therefore grayscale-based classifier is used when the recognition reliability of binary-based classifier is lower than the predetermined threshold (*Th2*). Threshold of degradation level and recognition reliability is determined by the steepest descent method. *Th1* and *Th2* are set to 82 and -0.7, respectively.

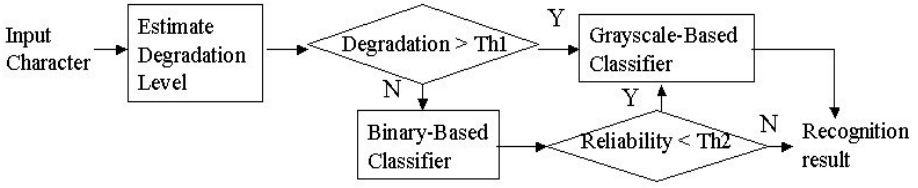


Fig. 12. Selection of binary-based and grayscale-based classifier

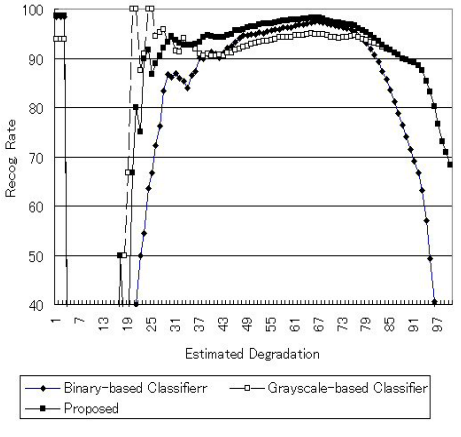
5 Experiment

Learning data for each classifier is the same as that used in 3.2.1. Also the data in 3.2.2 are used as learning data for deciding the threshold of degradation level and recognition reliability. 6 fonts data are used as test data that are different from those in 3.2.1 or 3.2.2 (Figure 15). Seven degrees of degradation images are generated in the same way as 3.2.1 or 3.2.2. The total number of characters is about 120,000.

Table 2 shows the comparison of recognition accuracy. Figure 13-14 shows the recognition accuracy according to degradation level. As for learning data, recognition rate of the proposed method is higher than other classifiers at almost all the degradation levels (Figure 13). As for test data, the recognition rate of the proposed method is

Table 2. Comparison of recognition accuracy

	Recog. rate (learning data)	Recog. rate (test data)
Binary-based	85.1%	83.5%
Grayscale-based	91.2%	92.1%
Proposed	93.6%	92.9%



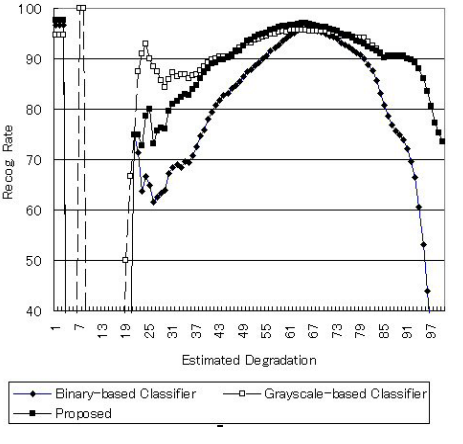


Fig. 13. Recognition accuracy of learning data

Fig. 14. Recognition accuracy of test data

Fig. 15. Sample font images used for test

Table 3. Comparison of recognition accuracy for test data at each font

	Font1	Font2	Font3	Font4	Font5	Font6
Binary-based	84.0%	88.9%	90.2%	88.6%	85.6%	63.6%
Grayscale-based	90.0%	95.1%	95.5%	94.3%	92.3%	85.5%
Proposed	92.6%	96.2%	96.2%	95.3%	93.1%	83.4%

higher only when the degradation level is at the middle levels (around 67). Even so, the total number of characters at middle level is too much, thus recognition accuracy of proposed method for entire data is highest. Table 3 shows the comparison of recognition accuracy for test data at each font. The proposed method is effective for many fonts except a very thick font (font6). Even when the degradation is low, the recognition reliability of binary-based classifier seems to be unreliable for a very thick font.

6 Conclusion

In this paper, the degradation level for grayscale character image is newly defined and the binary-based classifier that uses character contour feature and the grayscale-based classifier that uses pixel value of images as a feature are investigated according to the degradation level. It is found that recognition result or reliability of binary-based classifier is not reliable when input character is degraded, thus selection of binary-based classifier and grayscale-based classifier is considered based on the degradation level and the recognition reliability. Experimental results using lots of character images with variety of degradation level show that the proposed method achieves the recognition rate of about 92.9% whereas the binary-based classifier and the grayscale-based classifier achieve 83.5% and 92.1% respectively.

Since artificially generated character images are used in order to investigate the performance of each classifier according to the degradation level, our future work is to test the proposed method with real data.

References

[1] S.Omachi, F.Sun and H.Aso, "A Noise-Adaptive Discriminant Function and Its Application to Blurred Machine-Printed Kanji Recognition," IEEE Trans. PAMI, vol.22, no.3, pp.314-319, March 2000.

[2] O.Shiku, A.Nakamura, S.Miyahara and T.Ohyama, "Blurred Character Recognition by Complementing Features of Blurred Regions," IEICE D-II, Vol.J87-D-II, No.3, pp.808-817, 2004 (in Japanese).

[3] H.Ishida, S.Yanadume, T.Takahashi, I.Ide, Y.Mekada and H.Murase," Recognition of Low-Resolution Characters by a Generative Learning Method," Proc. of the 1st Int'l Workshop on Camera-Based Document Analysis and Recognition (CBDAR2005), pp.45-51, 2005.
[4] J.Sun, Y.Hotta, Y.Katsuyama and S.Naoi, "Low Resolution Character Recognition by Dual Eigenspace and Synthetic Degraded Patterns," Proc. of the 1st ACM Workshop on Hardcopy Document Processing, pp.15-22, 2004.
[5] M.Sawaki, H.Murase and N.Hagita,"Character Recognition in Bookshelf Images by Automatic Template Selection," Proc. of ICPR '98, Vol. 2, pp.1117-1120.
[6] H.S.Yam and E.H.B.Smith, "Estimating Degradation Model Parameters from Character Images," Proc. of Seventh ICDAR, pp.710-714, 2003.
[7] T.Iijima, "Theory of Pattern Recognition," Series of basic information technology 6, *Morikita Publishing Company Ltd*, 1989.