

Camera based Degraded Text Recognition Using Grayscale Feature

Jun Sun

Fujitsu R&D Center Co. Ltd, Beijing, China
sunjun@frdc.fujitsu.com

Yoshinobu Hotta, Yutaka Katsuyama,
Satoshi Naoi

Fujitsu Laboratories Ltd, Kawasaki, Japan
y.hotta, katsuyama, naoi.satoshi@jp.fujitsu.com

Abstract

As the rapid progress of digital imaging technology, camera based character recognition receives more and more attentions. One challenge in camera based OCR is the recognition for degraded text. Conventional OCR engines usually recognize on binary image. However, the performance drops dramatically as the degradation level increases. In this paper, a new recognition method is proposed to recognize degraded character based on dual eigenspace decomposition and synthetic degraded data. Then, the degraded character string is segmented by the combination of binary and grayscale analysis. Experiments on single character and text string recognition prove the effectiveness of our method.

1. Introduction

As the rapid progress of digital imaging technology, camera based character recognition receives more and more attentions. There is a trend that digital camera and video camera will gradually replace the scanner as document image capturing tools. Compared with document images obtained by scanner, the characters in camera-based images suffer more degradation: low resolution, blurring, uneven lighting, et al[1]. Figure 1. shows some samples of camera based character images.

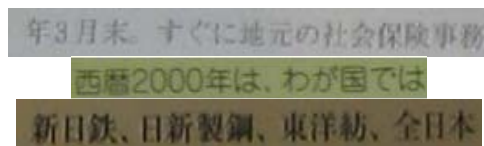


Figure 1. Camera based character images

Many methods are proposed to solve the problem of degraded character recognition, such as degradation recovery[2], advanced binarization[3], and Gabor filter based feature[4].

Most existing methods focus on how to remove the degradation and compare with ideal (binary) patterns in the training set. However, different from scanned document, the degradation in camera based text is much more complicated. Single recovery based method can not solve all the problems.

In this paper, a new recognition method is proposed to recognize degraded character based on grayscale feature and synthetic degraded data. First, a scaling degradation model is used to generate synthetic degraded character images. Representative degraded samples under various font types and blurring levels are obtained by clustering algorithm. These data are used to make synthetic degraded grayscale dictionary. The individual character is recognized by grayscale feature based on dual eigenspace decomposition[5], which is very robust against various types of noises. The combination of dual-eigenspace and degraded dictionary can efficiently improve the recognition capability for individual degraded characters.

Based on the proposed grayscale recognition method, grayscale character string is segmented by the combination of binary and grayscale analysis. Edge based binarization is used as initial coarse segmentation. Then Canny edge detector is performed to separate touching character. The final recognition result is obtained by a modified DP searching method.

Experiments are carried on Japanese characters obtained by digital cameras. Our method shows big advantage over traditional methods.

2. Synthetic degraded characters generation and processing

In order to recognize degraded characters, a large quantity of degraded patterns with different levels of degradation is needed as training set. However, manually collecting these degraded patterns is very time consuming and costly. Thus, a scaling degradation model is used to generate synthetic character images with different levels of degradation.

2.1 Scaling degradation model

The input to the degradation model is clear binary character image. The output of the model is degraded grayscale character image with the same size. The scaling model includes 2 steps:

- 1) Image decimating -- suppose the size of the binary image is $w*h$, given a decimating ratio, r , the original image is decimated by super-sampling interpolation to a grayscale image with size $wl*hl$. Where $wl = r*w$, $hl = r*h$, $r < 1.0$.
- 2) Image zooming -- the decimated grayscale image is zoomed back to a grayscale image with the original size $w*h$ by cubic interpolation.

Since pixel information is lost during the decimating and zooming operation, blurring degradation can be generated by the scaling model. The smaller the decimating ratio, r , is, the more degraded the generated character pattern is. Figure 2. shows some examples of scaling degradation.



Figure 2. Original binary image (left) and synthetic degraded images (right)

2.2. Typical patterns generation by clustering

Blurring degradation can approximate the results of many degradation factors, such like out of focus and low resolution. It can be regarded as one kind of extrinsic degradation. For character recognition, variations caused by different font types should also be considered, which can be regarded as one kind of intrinsic degradation.

In order to handle these two different types of degradation uniformly, clustering algorithm is used to generate "typical patterns":

- 1) For a given character category, binary character images with different font types are collected. This procedure can be achieved by print-and-scanning or automatic groundtruth generation.
- 2) For every binary character image, synthetic grayscale patterns with different degradation level are generated by scaling degradation model.
- 3) C-mean clustering algorithm is used to cluster the patterns of one category into N_c clusters.
- 4) The mean image of every cluster is taken as the typical pattern of that category.

Figure 3. is the result of typical patterns with $N_c = 5$, we can see that these patterns contain different font

shapes as well as different level of blurring degradation.

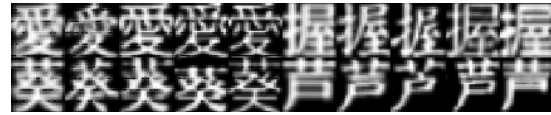


Figure 3. Results of C-mean Clustering

3. Dual eigenspace feature extraction and recognition

For degraded character images, grayscale feature has more advantages over binary feature. In this paper, a dual eigenspace based method is used for feature extraction and recognition.

3.1 Grayscale dictionary generation

The synthetic grayscale patterns generated in previous section are used to generate a two layer structure grayscale dictionary by dual eigenspace decomposition, which contains 3 steps:

1) Construction of unitary eigenspace: suppose the character image with size $w*h$ is represented by a vector $x = [x_1, x_2, \dots, x_{w*h}]^T$ using the raster scanning order. First, a unitary eigenspace is constructed by Principal Component Analysis on the covariance matrix of typical pattern images of all categories:

$$COV = \frac{1}{Pc} \sum_{i=1}^P \sum_{j=1}^{N_c} (m_{ij} - m)(m_{ij} - m)^T, \quad (1)$$

where P is the number of total categories, N_c is the number of clusters in every category, m is the mean vector for all typical images. m_{ij} is the typical image for the j th cluster in the i th category. The first n eigenvectors of matrix COV corresponding to the first n biggest eigenvalues are recorded as: $U = [u_1, u_2, \dots, u_n]^T$, which spans the unitary eigenspace.

2) First layer multi-template dictionary generation: the first layer multi-template dictionary is constructed by casting every typical image generated in 2.2 onto the unitary eigenspace:

$$c_{ij} = U^T (m_{ij} - m) \quad (2)$$

3) Individual eigenspace construction: since the first layer grayscale dictionary is constructed based on the samples of all categories, the discrimination power is not so strong. In order to improve the recognition performance, an individual eigenspace is built for every category using the PCA feature of all samples

belonging to that category. The auto-correlation matrix for the i th category is:

$$W_i = \frac{1}{M_i} \sum_{k=1}^{M_i} (y_i^{(k)} - c_i)(y_i^{(k)} - c_i)^T, i = 1, 2, \dots, P \quad (3)$$

where $y_i^{(k)} = U^T(x_i^{(k)} - m)$ is the PCA feature vector of the k th training sample $x_i^{(k)}$ in the i th category,

$$c_i = \sum_{j=1}^{N_c} c_{ij} / N_c, M_i \text{ is the number of training samples}$$

for the i th category. The first n_i eigenvectors of W_i corresponding to the first n_i eigenvalues are recorded as: $\tilde{U}_i = [u_1^i, u_2^i, \dots, u_{n_i}^i]$, $i = 1, 2, \dots, P$, which spans the individual eigenspace for the i th category. The individual eigenspaces of all categories construct the second layer of grayscale dictionary.

3.2 Dual eigenspace recognition

There are four steps in the recognition phase for dual eigenspace decomposition.

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_{ij} , $i = 1, 2, \dots, P$, $j = 1, 2, \dots, N_c$ 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 every of the d categories from the coarse classification:

$$\eta_i = \tilde{U}_i^T (y - c_i), \quad (4)$$

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

where η_i is the project coefficients 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:

$$\varepsilon_i = \|y - \hat{y}_i\|. \quad (6)$$

The category of optimal reconstruction, that is, the category with minimum reconstruction error, ε , is assigned to the testing sample.

4. Degraded character string segmentation

Character segmentation is another difficulty for degradation text. Compared with scanned document, camera based segmentation mainly suffers from 3 types of degradation (Figure 1.):

- 1) Weak stroke – the contrast between stroke and background is very small.
- 2) Touching characters – as the resolution decreases, the interval between two characters become smaller and smaller, which causes touching easily.
- 3) Fading horizontal stroke – this happens for some popular Japanese and Chinese fonts. The horizontal strokes are much thinner than vertical strokes and easily get invisible under large degradation.

In order to solve these problems, an edge based binarization is first used for initial coarse segmentation. Possible touching characters are segmented by Canny edge detector. A precise registration algorithm is used to find the exact boundary of degraded characters. Finally, the character string is segmented and recognized by a modified DP searching method.

4.1 Edge based binarization

The purpose of edge based binarization is to find weak strokes with low edge strength. Firstly, all candidate edge points in the image are detected. For every candidate edge point, mean edge strength is calculated in its neighborhood area. Only those edge point candidates with strength larger than the mean strength are left as true edge points. Then, local Niblack binarization method is performed on every true edge points. For every image pixel, $COUNT_n$ represents the times it is covered by a Niblack window, $COUNT_s$ represents the times it is binarized as stroke pixel. Final stroke pixels are selected if:

$$COUNT_s > \alpha COUNT_n, \quad (7)$$

where α is a ratio obtained by experiments. Figure 5. shows a comparison of edge based binarization and subpixel Niblack binarization[3]. We can see edge based binarization is much better at keeping the weak stroke pixels.

年3月末。すぐに地元の社会保険事務
年3月末。すぐに地元の社会保険事務

Figure 5. Subpixel Niblack binarization (upper) and edge based binarization (bottom).

4.2 Separation of touching characters

Although edge based binarization can find weak stroke, it sometimes will cause character touching for low resolution text string with tight spacing. We noticed that although the strokes touch with each other, the touching points usually have weak connection in grayscale image. So the touching points are detected by a canny edge detector based algorithm.

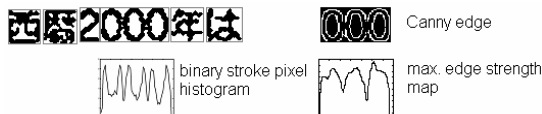


Figure 6. Touching characters separation based on Canny edge detector.

As shown in Figure 6., edge points are first detected using Canny edge detector (top right image). For every column, the maximum edge magnitude of the edge points is calculated (bottom right image). Then a 1-D Niblack algorithm is used to calculate the threshold to find the true “valleys” in the max. edge strength map. The valley points are further validated using the estimated character width within a text line. For comparison, the bottom left image in Figure 6 shows the histogram of binary stroke pixel. We can see that canny based segmentation correctly detects all true separation points while histogram based method has many false-positives.

4.3 Precise character registration

Section 4.1 and 4.2 can only provide initial segmentation result. As we know, PCA based feature extraction method is very sensitive to image shifting error. Also, for low resolution degraded character, the real character boundary position is within subpixel accuracy. Thus, a precise character registration algorithm is used to find the exact boundary of very character and output the regulated grayscale image. There are 3 key points in our registration algorithm:

- 1) The possible region of a character is extended and interpolated to ensure the character boundary can be estimated in subpixel accuracy.
- 2) Only background pixels are estimated and removed since character pixels are usually not dominant in the image as shown in Figure 7.
- 3) Size registration and sigmoid function based grayscale enhancement is performed after the character image is extracted using its precise boundary. This procedure focuses on the

enhancement of weak character strokes. The right part of Figure 7. shows the final result after registration.

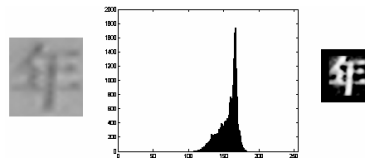


Figure 7. Interpolated character image (left), grayscale histogram (middle), and output of image registration (right).

4.4 Modified DP searching

The segmentation of character string is achieved by DP searching using the weighted recognition distance:

$$D = R_{recog} R_{geo} Dist(0), \quad (8)$$

where $Dist(0)$ is the recognition distance of the first candidate, $R_{recog} = Dist(0) / Dist(1)$ is the recognition confidence parameter, $R_{geo} = size * size / NUM(stroke)$ is the geometry parameter, which is the filling rate of a character. The reason of using geometry parameter is that for dual eigenspace recognition, the recognition distance (reconstruction error) varies according to the character structure. Under the same degradation level, simple structure character like “1” has much less reconstruction error than complex structure character like “愛”. The geometry parameter is used to compensate the inequality of the recognition distance. Finally, the segmentation is implemented by choosing the path with minimum accumulated recognition distance via DP searching.

5. Experiment results

In order to evaluate the effectiveness of our method, experiments are carried out for Japanese character recognition. For all the experiments, the number of the character category in the dictionary is 4299, which includes numeral, hiragana, katakana, kanji, alphabet and symbols. The original training data is 437,520 binary character images belonging to 122 font types.

5.1 Kanji characters Recognition

The first experiment is carried to recognize low resolution Kanji character images. 6 fonts of class-1 Japanese Kanji characters are printed in slides and projected onto the screen. An Olympus u400 digital

camera is used to capture the slide image. Finally, individual character images are extracted as testing set. The number of characters in each font is 2965. The size of the character varies from 10*10 ~ 15*15 pixels.

For comparison, contour directional feature[6], PCA with single template matching are also listed.

Table 1. Recognition rate for 6 font types (%)

Font type	CDF	PCA-1	PCA-2	DE
Mincho 1	68.09	84.86	86.24	90.15
Mincho 2	55.95	86.44	87.96	91.06
Gothic 1	62.31	77.33	79.18	83.94
Gothic 2	71.74	89.71	89.51	91.91
Gothic 3	74.17	90.83	90.99	93.02
MaruGothic 2	72.45	88.40	88.53	91.53
Average	67.45	85.93	87.07	90.27

The abbreviations in Table 1. are defined as:

CDF – contour directional feature based on subpixel Niblack binarized image. The dimension is 288, the classifier is MQDF [7].

PCA-1 – single template PCA feature. The mean feature of every category is taken as feature template.

PCA-2 – Multi-template PCA feature. The number of template is 5. The dimension of the feature in PCA-1 and PCA-2 are all 200.

DE – dual eigenspace based feature. The dimension of the 1st PCA feature is 200. The dimension of 2nd eigenspace is 16.

PCA1, PCA2 and DE are all trained from synthetic character images generated using 3 level of decimating ratio. As shown in Table 1., grayscale features all outperforms binary feature. The multi-template dictionary generated from clustering result is better than single template based PCA feature. Dual eigenspace method gets the best result because of its coarse-to-fine recognition structure.

The effectiveness of the synthetic patterns is reported in previous work of the authors[8].

5.2 Character string recognition

The overall recognition performance of character strings are also tested using 3 sets of text lines from camera images. Table 2. lists the recognition results.

Table 2. Recognition rate for character strings

	Set1	Set2	Set3
Number of strings	293	73	283
Number of characters	4499	1657	5002
Avg. char. size (pixel)	18*18	16*16	12*12
CDF + old segmentation	72.10	69.76	62.79
DE + old segmentation	79.00	78.03	76.01
DE + new segmentation	93.77	86.54	80.13

As shown in Table 2., segmentation algorithm has very big influence to text string recognition. By combining the dual eigenspace recognition with the new segmentation method, the overall recognition rates for all 3 data sets all get great improvements.

6. Conclusion

A degraded text segmentation and recognition method is proposed in this paper. The character string is first coarse segmented using a combination of binary and grayscale analysis. Individual grayscale character images are then extracted by a precise character registration method. After that, a dual eigenspace based method is used for feature extraction and recognition. Experiment results show our method is very effective for degraded Japanese characters.

7. References

- [1] D. Doermann, J. Liang, H. P. Li, "Progress in camera-based document image analysis", In Proceedings of the 7th International conference on Document Analysis and Recognition Volume 1, pp. 606-616, Edinburgh, Scotland, 2003.
- [2] Taylor, M. J., Dance, C. R., Enhancement of document images from cameras. Proc. of SPIE vol. 3305, pp.230-241, 1998
- [3] Kamada, H., Fujimoto, K. High-Speed, High-Accuracy Binarization Method for Recognizing Text in Images of Low Spatial Resolutions. IEEE Fifth international conference on Document Analysis and Recognition, 1999, pp. 139-143
- [4] Wang X. W., Ding X. Q., and Liu C. S. Optimized Gabor filter based feature extraction for character recognition. Proceedings of ICPR, pp.223-226, 2002.
- [5] Zhang, D., Peng, H., Zhou, J., Sankar, K. P. A novel face recognition system using hybrid neural and dual eigenspace methods. IEEE trans. System, Man and Cybernetics – part A 32 (6) pp.787-792, 2002
- [6] Hai, T., Kabuyama, Y., and Yamamoto, E., A method for Handwritten Kanji Character Recognition -- Recognition Method by Multiple Standpoints and Particular Shape Extraction., IEICE Vol.J68-D, No.4,pp.773-780, Apr.1985 (in Japanese)
- [7] F.Kimura, K.Takashina, S.Tsuruoka, and Y.Miyake, " Modified Quadratic Discriminant Functions and the Application to Chinese Character Recognition", IEEE Trans. PAMI. 9, No. 1, 149-153, 1987.
- [8] J. Sun, Y. Hotta, Y. Katsuyama, S. Naoi, "Low resolution character recognition by dual eigenspace and synthetic degraded patterns." 1st ACM workshop on Hardcopy Document Processing. 15-22, 2004.