# Normalization Ensemble for Handwritten Character Recognition

Cheng-Lin Liu and Katsumi Marukawa
Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
{liucl,marukawa}@crl.hitachi.co.jp

## Abstract

*This paper proposes a multiple classifier approach, called normalization ensemble, for handwritten character recognition by combining multiple normalization methods. By varying the coordinate mapping mode, we have devised 14 normalization functions, and switching on/off slant correction results in 28 instantiated classifiers. We will show that the classifiers with different normalization methods are complementary and the combination of them can significantly improve the recognition accuracy. In experiments of handwritten digit recognition on the NIST Special Database 19, the normalization ensemble was shown to reduce the error rate by factors from 10.6% to 26.9% and achieved the best error rate 0.43%. We also show that the complexity of normalization ensemble can be reduced by selecting seven classifiers from 28 with little loss of accuracy.*

## 1. Introduction

The performance of a character recognizer depends on the pre-processing procedure, feature representation, classifier model, training method and sample data. Due to the imperfection with one or several of these factors, a single classifier cannot achieve the best recognition accuracy. To overcome this limitation, multiple classifier methods have been widely adopted for achieving higher accuracies.

The classification performance of multiple classifier systems (also called ensembles or committees) relies on both the diversity (complementariness or independence) of the participating classifiers and the decision combination method. Complementary classifiers can be generated by diversifying classifier structures, training data, input features, output targets, etc. [1], while the distributed decisions of a given set of classifiers can be combined using various rules [2, 3]. Ensemble methods based on data resampling or filtering and those based on feature selection [4, 5] have received interests. For the special problem of character recognition, classifier diversity can also be obtained by varying the image processing procedures: extracting different features has been commonly adopted, and generating multiple transformed images by perturbation has shown promise [6, 7].

This paper proposes a new ensemble approach, called normalization ensemble, for handwritten character recognition. By this approach, the input character image is transformed to multiple normalized images using different normalization methods. The decisions of multiple classifiers, each on a normalized image, are combined to give the final decision. Unlike the perturbation method that uses the same trained classifier on the transformed images (one of which is expected to best restore the shape deformation), the normalization ensemble uses a trained classifier for each normalization procedure and all the trained classifiers give high accuracies due to the effects of normalization.

The normalization ensemble approach is motivated by the availability of various normalization methods [8, 9, 10, 11] and their differing recognition performance. We will show that the combination of multiple normalization procedures can yield significant improvement compared to the single best one. Our normalization ensemble is based on 14 basic normalization functions varying in linear/nonlinear coordinate mapping, centroid/boundary alignment, aspect ratio mapping, etc. [11, 12], and combining the basic normalization functions with slant correction (deslant) results in 28 instantiated normalization functions. The decisions of 28 classifiers, each on a different normalization function, are combined to give the final classification result.

To validate the effectiveness of normalization ensemble, we have conducted experiments of handwritten digit recognition on the NIST Special Database 19. Our previous works focused on extracting features and training classifiers and have reported superior recognition accuracies [12, 13]. It is, however, difficult to further improve the accuracy using single classifiers. We will show that the normalization ensemble can yield significant improvement.

Combing 28 classifiers in an ensemble imposes heavy computation. A strategy to alleviate this is to select a subset of classifiers that preserve or improve the ensemble accuracy. Guided by the compound diversity measure of [14], we selected seven classifiers from each ensemble and the loss of accuracy is insignificant. In the rest of this paper,

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

Section 2 describes the normalization ensemble; Section 3 presents our experimental results and Section 4 makes conclusion.

## 2. Normalization Ensemble

The architecture of normalization ensemble is shown in Fig. 1. The input character image is transformed to $K$ normalized images using different normalization methods. Each normalized image then undergoes feature extraction and classification. We may extract different features and use different classifier models on the normalized images, but in this study, we use the same feature representation and same classifier model so as to investigate the effects of varying normalization. The classification on each normalized image gives a decision (class label, rank order, or class scores). The decision combiner fuses the $K$ decisions to give the final classification result.

### 2.1 Normalization methods

We have 14 basic (size) normalization functions, and switching on/off slant correction (deslant) to each of them results in 28 normalization functions in the ensemble. When deslant is switched on, it precedes size normalization to eliminate the slant according to second-order moments [8]. The basic normalization functions include six linear normalization functions (denoted by F0–F5), a nonlinear normalization method (F6), five moment normalization functions (F7–F11), a centroid-boundary alignment method (F12), and a bi-moment method (F13) [11]. We briefly review the normalization methods in the following.

Normalization is performed by mapping the pixels of the input image to a standard normalized image plane. The normalized plane is pre-specified and is usually a square. We control the aspect ratio of the normalized image in so-called aspect ratio adaptive normalization (ARAN) [15], in which the aspect ratio $R_2$ of normalized image is a monotone function of the aspect ratio $R_1$ of input image. If the input image is vertically elongated, then in the normalized plane, the vertical dimension is filled and the horizontal dimension is centered and scaled according to the aspect ratio; otherwise the horizontal dimension is filled and the vertical dimension is centered and scaled.

We roughly group the normalization methods into three categories: boundary alignment (conventional linear and nonlinear normalization), centroid alignment (moment normalization), and curve fitting (both centroid and boundaries aligned). Nonlinear normalization is based on line density equalization [9, 10]. Curve fitting-based normalization methods include a centroid-boundary alignment method (*centr-bound* for abbreviation) and a bi-moment method [11]. In both moment and bi-moment methods, the character boundaries are re-set according to second-order moments. The details of these normalization methods can be found in [11, 12]. By combining the normalization methods with different aspect ratio mapping functions, we have generated 14 basic normalization functions, which are listed as follows.

- F0: linear normalization, fixed aspect ratio,
$$R_2 = 1.$$
- F1: linear normalization, aspect ratio preserved,
$$R_2 = R_1.$$
- F2: linear normalization, square root of aspect ratio,
$$R_2 = \sqrt{R_1}.$$
- F3: linear normalization, cubic root of aspect ratio,
$$R_2 = \sqrt[3]{R_1}.$$
- F4: linear normalization, piecewise linear of aspect ratio,
$$R_2 = \begin{cases} 0.25 + 1.5R_1, & \text{if } R_1 < 0.5 \\ 1, & \text{otherwise} \end{cases}$$
- F5: linear normalization, square root of sine,
$$R_2 = \sqrt{\sin(\frac{\pi}{2}R_1)}.$$
- F6: nonlinear normalization with aspect ratio mapping F5. The line density histograms are computed by the method of [9].
- F7: moment normalization, aspect ratio preserved.
- F8: moment normalization, square root of aspect ratio.
- F9: moment normalization, cubic root of aspect ratio.
- F10: moment normalization, aspect ratio mapping F5.
- F11: moment normalization, aspect ratio $R_2 = 1$.
- F12: centr-bound method, aspect ratio mapping F5.
- F13: bi-moment method, aspect ratio mapping F5.

Examples of 14 normalization functions are shown in Fig. 2, where the leftmost image is the input image and the other images are the normalized ones.

We refer to the basic normalization functions as F0–F13, while the normalization functions preceded by deslant are referred to as D0–D13.

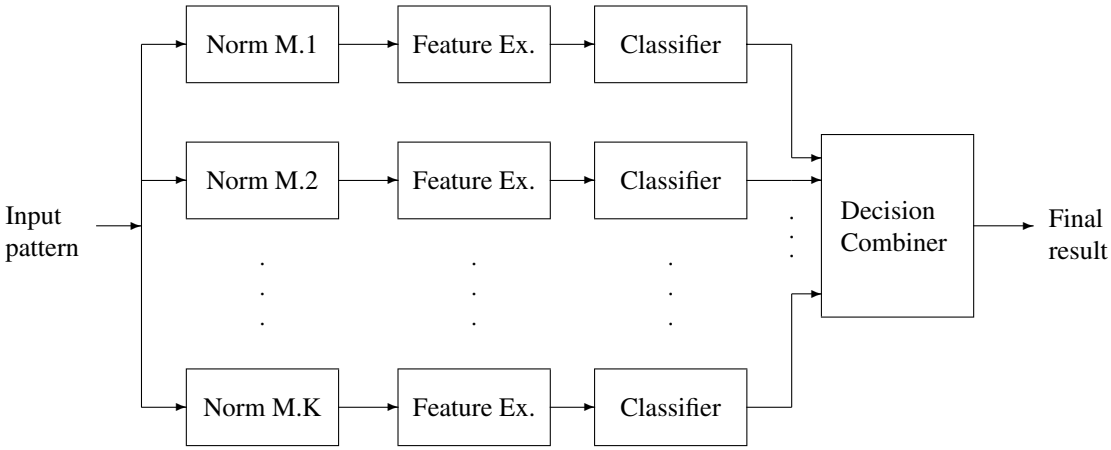


**Figure 1. Architecture of normalization ensemble for character recognition. "Norm M.K" stands for "Normalization Method No.K" and "Feature Ex." is "Feature Extraction".**

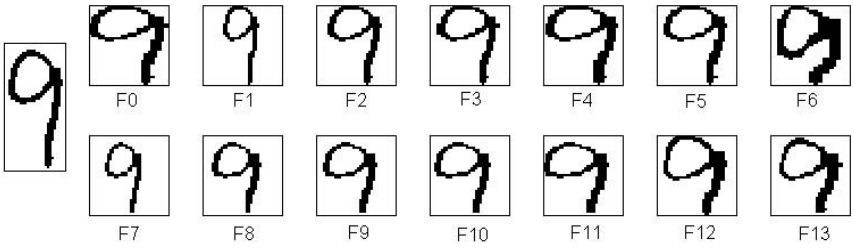


**Figure 2. Examples of character image normalization.**

### 2.2 Classifier combination and selection

We test two combination rules for combining the decisions in normalization ensemble: abstract-level combination by plurality vote (PV) and measurement-level combination by the sum-rule (SR) [2]. For measurement-level combination, we transform the classifier outputs (class scores) into confidence measures that represent the class probabilities. The benefits of confidence transformation have been demonstrated in [16]. The classifier outputs are first rescaled to a moderate range using a scaling function and then transformed into confidence measures using an activation function. For scaling, we take the one-dimensional Gaussian density modeling method of Schürmann [17], and for activation, we approximate multi-class posterior probabilities by combining sigmoid measures using the Dempster-Shafer theory of evidence. The details of confidence transformation can be found in [16].

The complexity of normalization ensemble can be reduced by selecting a subset of classifiers for combination. The subset is selected with the aim of optimizing a selection criterion on a validation dataset. The combination accuracy on the validation set has often been taken as the selection criterion, but some diversity measures are computationally simple and generalize well to unseen data. We use the compound diversity (CD) measure of Giacinto and Roli [14], which is the complement of the double-fault percentage of a pair of classifiers. For evaluating a subset of classifiers, we average the CD measure over all pairs of them. The subset with the maximum average CD is selected. To overcome the exponential explosion of exhaustive search, we use a suboptimal sequential search method, the plus l-take way r (PTA(l,r)) method for selecting a specified number of classifiers.

## 3. Experimental Results

We tested normalization ensembles in handwritten digit recognition on the NIST Special Database 19 (SD19). From this very large database, we compiled medium-size datasets for training, validation, and testing [18]. Our training set contains 66,214 digit samples of 600 writers, the validation set contains 22,271 samples of 200 writers, and the test set contains 45,398 samples of 400 writers. The training set was used for estimating the classifier parameters. The validation set was used for selecting hyper-parameters of clas-

sifiers and for training classifier combination (confidence transformation, weighted combination, subset selection).

The features and classifier models were selected from those that yielded high performance in our previous experiments [12]. We tested three direction features in normalization ensembles: the blurred chaincode feature, the normalization-cooperated feature extraction (NCFE) [19], and the gradient feature on gray-scale normalized images. We only show the results of 8-direction features (200-dimensional) because they yield higher recognition accuracies than 4-orientation features. The NCFE feature measures the distribution of contour directions before normalization (only the stroke positions are mapped by normalization). To enlarge the diversity among the classifiers with NCFE, we also map the stroke contour directions in normalization. The horizontal and vertical directions do not change, while the mapped diagonal directions are decomposed to standard chaincode directions in a similar way to the gradient feature [12, 13]. We call this method modified NCFE (MNCFE).

For classification, we selected the popular neural network multi-layer perceptron (MLP), the polynomial classifier [20] and the discriminative learning quadractic discriminant function (DLQDF) [21]. The PC and the DLQDF were shown to yield high accuracies among the classifiers besides support vector classifiers (SVCs) [13]. The classifier structures were tuned to give high accuracies to the validation set. As a result, the MLP has one layer of 300 hidden units, the PC uses 70 principal components for binomial expansion, and the DLQDF uses 40 eigenvectors for each class.

A normalization ensemble was build for each combination of feature type and classifier model. In addition to combining 28 classifiers, we also tested small ensembles combining 14 basic normalization functions, either with (D0–13) or without deslant (F0–13). In each normalization ensemble, the lowest error rate was given either by moment normalization (F8, F9 or F10) or by curve fitting-based normalization (F12 or F13). Deslant normalization (D0–13) mostly gave higher accuracy than the corresponding normalization function without deslant. The globally lowest error rate, 0.50%, was given by the PC on gradient feature and deslant normalization function D8.

### 3.1 Results of normalization ensembles

The error rates of normalization ensembles on the test set are shown in Table 1. For each ensemble, we show the error rate of the best individual classifier (i.e., the best normalization function), the error rate of Oracle that gives correct classification when at least a participating classifier classifies correctly, and the error rates of combination by plurality vote (PV) and sum-rule (SR). We also show the error reduction rate of each ensemble as compared to the best individual error rate.

The error rate of Oracle is the intersection of classification error of different methods. We can see that this intersection is very small and hence indicates good complementariness between the normalization methods. The combination of classifiers with difference normalization methods yields lower error rate than the best individual classifier. Measurement-level combination (SR) mostly yields lower error rate than abstract-level combination (PV) by better utilizing the output information of classifiers.

The error reduction rate of normalization ensembles ranges from 1.9% to 25.6% for ensembles of 14 classifiers, and from 10.6% to 26.9% for ensembles of 28 classifiers. It turns out that error reduction rate is higher for inferior features (e.g., chaincode feature) and classifier models (e.g., MLP) than for superior ones. The MNCFE, though individually performs comparably with the chaincode feature, yields lower ensemble accuracy due to the insufficient complementariness. The globally best classification performance was yielded by the ensemble of superior feature (gradient feature) and classifier models (PC and LDQDF). The lowest error rate, 0.43%, is a significant improvement compared to the best individual classifier (0.50%).

The improvement of normalization ensembles is significant because the misclassified test samples by individual classifiers are really difficult and even the support vector classifiers (SVCs) do not reduce the error rates significantly. Fig. 3 shows some test samples misclassified by an individual classifier of high accuracy, which have ambiguous shapes and some were even mis-labeled.

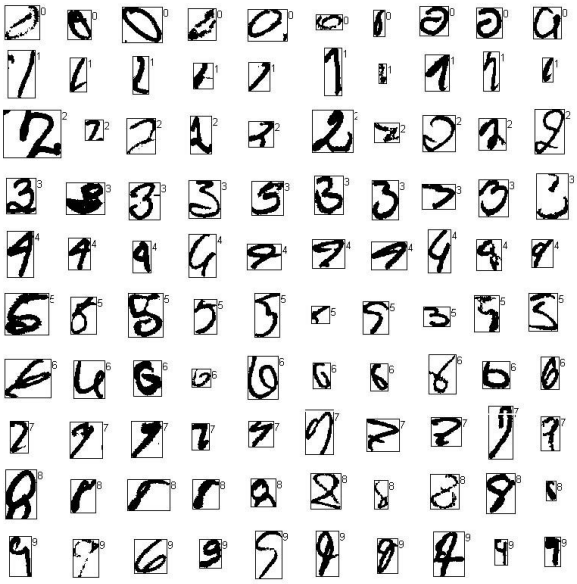

**Figure 3. Digit samples misclassified by an individual classifier of high accuracy.**

Table 2 shows the test error rates of an SVC with Gaussian (radial basis function) kernel (SVC-rbf, comprised of

**Table 1. Classification error rates (%) and error reduction (Red.) rates (%).**

| Classifier | Feature | Normalization | Best | Oracle | PV | SR | Red. |
|---|---|---|---|---|---|---|---|
| MLP | chaincode | F0–13 | 0.82 | 0.12 | 0.61 | 0.61 | 25.6 |
| | | D0–13 | 0.67 | 0.12 | 0.54 | 0.51 | 23.9 |
| | | (F0–13)+(D0–13) | 0.67 | 0.07 | 0.50 | **0.49** | 26.9 |
| | MNCFE | F0–13 | 0.79 | 0.18 | 0.67 | 0.65 | 17.7 |
| | | D0–13 | 0.69 | 0.17 | 0.59 | 0.57 | 17.4 |
| | | (F0–13)+(D0–13) | 0.69 | 0.11 | 0.53 | **0.51** | 26.1 |
| | gradient | F0–13 | 0.73 | 0.17 | 0.59 | 0.57 | 21.9 |
| | | D0–13 | 0.57 | 0.13 | 0.54 | 0.52 | 8.8 |
| | | (F0–13)+(D0–13) | 0.57 | 0.08 | 0.47 | **0.48** | 17.5 |
| PC | chaincode | F0–13 | 0.66 | 0.14 | 0.55 | 0.55 | 16.7 |
| | | D0–13 | 0.56 | 0.12 | 0.49 | 0.47 | 16.1 |
| | | (F0–13)+(D0–13) | 0.56 | 0.08 | 0.48 | **0.47** | 16.1 |
| | MNCFE | F0–13 | 0.58 | 0.19 | 0.56 | 0.58 | 3.4 |
| | | D0–13 | 0.57 | 0.19 | 0.54 | 0.52 | 8.8 |
| | | (F0–13)+(D0–13) | 0.57 | 0.11 | 0.50 | **0.49** | 16.3 |
| | gradient | F0–13 | 0.54 | 0.18 | 0.54 | 0.53 | 1.9 |
| | | D0–13 | 0.50 | 0.15 | 0.47 | 0.46 | 8.0 |
| | | (F0–13)+(D0–13) | 0.50 | 0.10 | 0.44 | **0.44** | 12.0 |
| DLQDF | chaincode | F0–13 | 0.66 | 0.12 | 0.53 | 0.53 | 19.7 |
| | | D0–13 | 0.54 | 0.12 | 0.47 | 0.46 | 14.8 |
| | | (F0–13)+(D0–13) | 0.54 | 0.06 | 0.46 | **0.45** | 16.7 |
| | MNCFE | F0–13 | 0.61 | 0.21 | 0.54 | 0.54 | 11.5 |
| | | D0–13 | 0.59 | 0.22 | 0.58 | 0.54 | 8.5 |
| | | (F0–13)+(D0–13) | 0.59 | 0.13 | 0.50 | **0.49** | 10.6 |
| | gradient | F0–13 | 0.53 | 0.14 | 0.49 | 0.49 | 7.5 |
| | | D0–13 | 0.51 | 0.15 | 0.47 | 0.46 | 9.8 |
| | | (F0–13)+(D0–13) | 0.51 | 0.08 | 0.45 | **0.43** | 15.7 |

ten binary classifiers each separating one class from the others) on normalization function F9. The implementation details of SVC-rbf can be found in [13]. We can see that the lowest error rate of SVC-rbf, 0.49%, is not significantly better than the individual PC (0.50%) or DLQDF (0.51%). The SVCs are very expensive in storage and computation due to the large number of support vectors. The normalization ensemble achieves significantly higher accuracies at comparable or lower complexity.

The comparison of normalization ensemble and perturbation method is of interest. Since the classification accuracy depends on the feature representation and classifier model as well as the dataset, it is hard to directly compare their performances. On two NIST test sets, Ha et al. reported error reduction rates 16.4% (from accuracy 99.45% to 99.54%) and 9.4% (from accuracy 96.80% to 97.10%) [7]. The effect of our normalization ensembles is at least comparable to this. As to the computational complexity, both normalization ensemble and perturbation method perform normalization and classification multiple times, yet the normalization ensemble is more expensive in storage because it uses multiple classifiers. This complexity can be reduced by classifier selection.

**Table 2. Error rates of support vector classifier with Gaussian kernel on normalization function F9.**

| Feature | deslant | #SV | Error (%) |
|---|---|---|---|
| chaincode | no | 8,170 | 0.59 |
| | yes | 6,722 | 0.53 |
| MNCFE | no | 8,077 | 0.55 |
| | yes | 6,784 | 0.56 |
| gradient | no | 7,141 | 0.50 |
| | yes | 5,822 | 0.49 |

#SV: number of distinct support vectors

### 3.2 Results of classifier subset selection

By selecting seven classifiers from 28, the classification error rates on the test set are shown in Table 3, where we also show the selected normalization functions. Comparing the error rates of selected subsets with those of ensembles combining all 28 classifiers (Table 1), the combination per-

formance deteriorates considerably only for one ensemble (MLP on gradient feature). The selected subsets are apparently biased to deslant normalization functions but show reasonable diversity, e.g., the subset of DLQDF on chaincode feature contains two linear normalization functions (F3,D0), a nonlinear normalization (F6), three moment normalization functions (F7,D8,D10), and a bi-moment normalization (D13).

**Table 3. Error rates (%) of combining selected normalization functions. The error rate is highlighted when it is not higher than that of combining 28 classifiers.**

| Classifier | Feature | Selected | PV | SR |
|---|---|---|---|---|
| MLP | chaincode | F7,12;D4,5,7,10,13 | 0.53 | **0.49** |
| | MNCFE | F7,8;D2,4,6,10,12 | 0.56 | 0.55 |
| | gradient | F6,8;D2,5,6,10,11 | 0.52 | 0.52 |
| PC | chaincode | F4,10;D5,6,7,10,13 | **0.48** | **0.46** |
| | MNCFE | F0,8;D4,6,7,10,12 | **0.50** | **0.49** |
| | gradient | F1,6;D0,6,7,9,10 | 0.45 | 0.45 |
| DLQDF | chaincode | F3,6,7;D0,8,10,13 | 0.47 | 0.46 |
| | MNCFE | F0,6,7,D0,6,11,12 | 0.51 | 0.50 |
| | gradient | F2,9;D0,2,6,8,10 | **0.44** | **0.43** |

## 4. Conclusion

We proposed a normalization ensemble approach for handwritten character recognition and have demonstrated its effectiveness. The complexity of normalization ensembles can be reduced by classifier subset selection with little loss of accuracy. The performance of normalization ensembles can be enhanced with more complementary normalization methods. On the other hand, the comparison of normalization ensemble and perturbation method as well as the possible hybridization of them should be furthered in the future.

## References


[1] T.G. Ditterich, Ensemble methods in machine learning, *Multiple Classifier Systems*, J. Kittler and F. Roli (Eds.), LNCS Vol.1857, Springer, 2000, pp.1-15.

[2] J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, On combining classifiers, *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3): 226-239, 1998.

[3] R.P.W. Duin, The combining classifiers: to train or not to train, *Proc. 16th ICPR*, Quebec, Canada, 2002, Vol.2, pp.765-770.

[4] S. Günter, H. Bunke, Creation of classifier ensemble for handwritten word recognition using feature selection algorithms, *Prof. 8th IWFHR*, Ontario, Canada, 2002, pp.183-188.

[5] R. Bryll, R. Gutierrez-Osuna, F. Quek, Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets, *Pattern Recognition*, 36(6): 1291-1302, 2003.

[6] M. Yasuda, K. Yamamoto, H. Yamada, Effect of the perturbed correlation method for optical character recognition, *Pattern Recognition*, 30(8): 1315-1320, 1997.

[7] T. Ha, H. Bunke, Off-line handwritten numeral recognition by perturbation method, *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5): 535-539, 1997.

[8] R.G. Casey, Moment normalization of handprinted character, *IBM J. Res. Dev.*, 14: 548-557, 1970.

[9] J. Tsukumo, H. Tanaka, Classification of handprinted Chinese characters using non-linear normalization and correlation methods, *Proc. 9th ICPR*, Roma, Italy, 1988, pp.168-171.

[10] H. Yamada, K. Yamamoto, T. Saito, A nonlinear normalization method for hanprinted Kanji character recognition—line density equalization, *Pattern Recognition*, 23(9): 1023-1029, 1990.

[11] C.-L. Liu, H. Sako, H. Fujisawa, Handwritten Chinese character recognition: alternatives to nonlinear normalization, *Proc. 7th ICDAR*, Edinburgh, Scotland, 2003, pp.524-528.

[12] C.-L. Liu, K. Nakashima, H. Sako, H. Fujisawa, Handwritten digit recognition: investigation of normalization and feature extraction techniques, *Pattern Recognition*, 37(2): 265-279, 2004.

[13] C.-L. Liu, K. Nakashima, H. Sako, H. Fujisawa, Handwritten digit recognition: benchmarking of state-of-the-art techniques, *Pattern Recognition*, 36(10): 2271-2285, 2003.

[14] G. Giacinto, F. Roli, An approach to the automatic design of multiple classifier systems, *Pattern Recognition Letters*, 22(1): 25-33, 2001.

[15] C.-L. Liu, M. Koga, H. Sako, H. Fujisawa, Aspect ratio adaptive normalization for handwritten character recognition, *Advances in Multimodal Interfaces—ICMI 2000*, T. Tan, Y. Shi, and W. Gao (Eds.), LNCS Vol.1948, Springer, 2000, pp.418-425.

[16] C.-L. Liu, H. Hao, H. Sako, Confidence transformation for combining classifiers, *Pattern Analysis and Applications*, 7(1): 2-17, 2004.

[17] J. Schürmann, *Pattern Classification—A United View of Statistical and Neural Approaches*, Wiley-Interscience, 1996.

[18] C.-L. Liu, H. Sako, H. Fujisawa, Performance evaluation of pattern classifiers for handwritten character recognition, *Int. J. Document Analysis and Recognition*, 4(3): 191-204, 2002.

[19] M. Hamanaka, K. Yamada, J. Tsukumo, Normalization-cooperated feature extraction method for handprinted Kanji character recognition, *Proc. 3rd IWFHR*, Buffalo, NY, 1993, pp.343-348.

[20] U. Kreßel, J. Schürmann, Pattern classification techniques based on function approximation, *Handbook of Character Recognition and Document Image Analysis*, H. Bunke and P.S.P. Wang (Eds.), World Scientific, 1997, pp.49-78.

[21] C.-L. Liu, H. Sako, H. Fujisawa, Discriminative learning quadratic discriminant function for handwriting recognition, *IEEE Trans. Neural Networks*, 15(2): 430-444, 2004.