Online Character Recognition Based on Elastic Matching and Quadratic Discrimination

Hiroto Mitoma, Seiichi Uchida, and Hiroaki Sakoe
Kyushu University, Hakozaki 6–10–1, Higashi-ku, Fukuoka-shi, 812–8581 Japan
{mitoma, uchida, sakoe}@human.is.kyushu-u.ac.jp

Abstract

We try to link elastic matching with a statistical discrimination framework to overcome the overfitting problem which often degrades the performance of elastic matching-based online character recognizers. In the proposed technique, elastic matching is used just as an extractor of a feature vector representing the difference between input and reference patterns. Then quadratic discrimination is performed under the assumption that the feature vector is governed by a Gaussian distribution. The result of a recognition experiment on UNIPEN database (Train-R01/V07, 1a) showed that the proposed technique can attain a high recognition rate (97.95%) and outperforms a recent elastic matching-based recognizer.

1. Introduction

Elastic matching is often employed in online character recognition for establishing sample point correspondence between input and reference patterns. The functions of elastic matching are (i) the adjustment of the difference in pattern length (i.e., the number of sample points) and (ii) the minimization of the difference in feature values (e.g., xy-coordinate feature and directional feature). Dynamic programming (DP) matching is a classic elastic matching technique [1, 2, 3] and still very popular because of its merits. For example, DP matching algorithms can provide globally optimal matching with only $O(IJ)$ computations, where I and J are the length of reference and input patterns, respectively. In conventional DP matching-based recognizers, a matching cost obtained as a by-product of their matching optimization procedure is directly used as a discriminant function.

Although DP matching and other elastic matching-based recognizers generally perform well, they often suffer from misrecognitions due to *overfitting*, which is the phenomenon that the distance between the reference pattern of an incorrect category and an input pattern is underestimated by unnatural matching. For example, "1" and "7" are sometimes misrecognized because elastic matching compensates the difference of the length of their beginning parts.

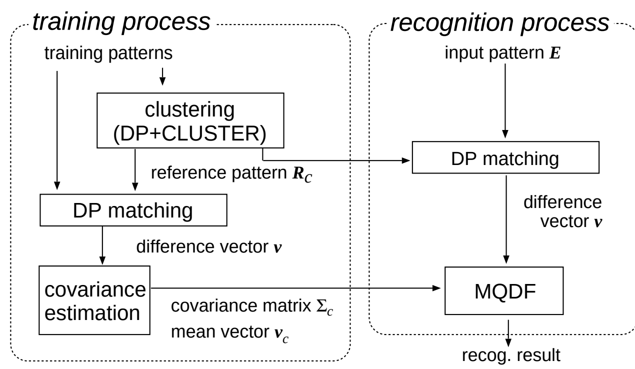

Figure 1. Overview of the proposed technique.

One possible remedy against the overfitting problem is the incorporation of probabilistic/statistical techniques. Statistical DP [4] and hidden Markov model (HMM) [5, 6, 7] are probabilistic extensions of DP and can avoid the overfitting by regulating the probability of feature values. Although they often outperform naive DP techniques, they cannot exclude all overfittings because their Markovian property allows to regulate only a "local" and "individual" probability of the feature value at each sample point. Thus, those techniques cannot regulate a "global" and "mutual" probability of whole sample points.

In this paper, we try to link elastic matching with statistical discrimination framework to overcome the overfitting problem. In the proposed technique, elastic matching is considered as a feature extraction procedure (and its matching cost is disregarded). Specifically, elastic matching is only used to provide a *difference vector* composed of the difference values between every corresponding sample point. Thus, the difference vector captures "global" feature of the input pattern (relative to the reference pattern). Then, Bayes discrimination is performed under the assumption that the difference vector is governed by a Gaussian distribution whose mean and covariance are estimated empirically in advance. For elastic matching results in overfitting, its difference vector will deviate from the distribution. Thus, the a posteriori probability of the difference vector will become a small value, and the misrecognition due to the overfitting will be reduced by the proposed technique. Figure 1 is the diagram of the proposed technique, where MQDF (modified quadratic discriminant function [8]) is a practical version of QDF, which is a Bayesian discriminant function for the patterns having Gaussian distribution.

Another contribution of the paper is the derivation of *eigen-deformations* of online character patterns. Eigen-deformations are the promising representation of the frequent deformations of online characters and produced as a by-product of QDF. In this paper, we just observe the eigen-deformations and suggest that they will be useful for developing some novel recognizer in future.

2. DP matching and Overfitting

2.1. DP matching

Let $\boldsymbol{R}_c = \boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_i, \ldots, \boldsymbol{r}_I$ denote the reference pattern of category $c \in [1, 2, \ldots, C]$ and $\boldsymbol{E} = \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_j, \ldots, \boldsymbol{e}_J$ denote an input pattern. The vectors $\boldsymbol{r}_i$ and $\boldsymbol{e}_j$ are composed of x-coordinate, y-coordinate, and local direction, and denoted as $\boldsymbol{r}_i = (X_i, Y_i, \Theta_i)^T$ and $\boldsymbol{e}_j = (x_j, y_j, \theta_j)^T$. Although $\boldsymbol{r}_i$ and I should be denoted like $\boldsymbol{r}_{c,i}$ and I_c, simpler notations are used whenever there is no confusion.

The problem of the elastic matching between two patterns $\boldsymbol{R}_c$ and $\boldsymbol{E}$ is defined as the following constrained optimization problem.

[Objective function]

$$\frac{1}{I} \sum_{i=1}^{I} \left\| \boldsymbol{r}_i - \boldsymbol{e}_{j(i)} \right\| \rightarrow \text{minimize} \tag{1}$$

[Control variables]

$$j(1), \ldots, j(i), \ldots, j(I)$$

[Constraint]

$$\begin{cases} j(i) - j(i-1) \in \{0, 1, 2\} \\ j(1) = 1 \\ j(I) = J \end{cases}$$

The notation $\|\boldsymbol{x}\|$ is the Euclidean norm of vector $\boldsymbol{x}$. This optimization problem can be solved effectively by a DP algorithm (whose detail are omitted here).

2.2. Overfitting

In many conventional online character recognition methods, the minimum value of the objective function (1), i.e.,

$$D_{\mathrm{DP}}(\boldsymbol{R}_c, \boldsymbol{E}) = \min_{j(1), \ldots, j(i), \ldots, j(I)} \frac{1}{I} \sum_{i=1}^{I} \left\| \boldsymbol{r}_i - \boldsymbol{e}_{j(i)} \right\| \tag{2}$$

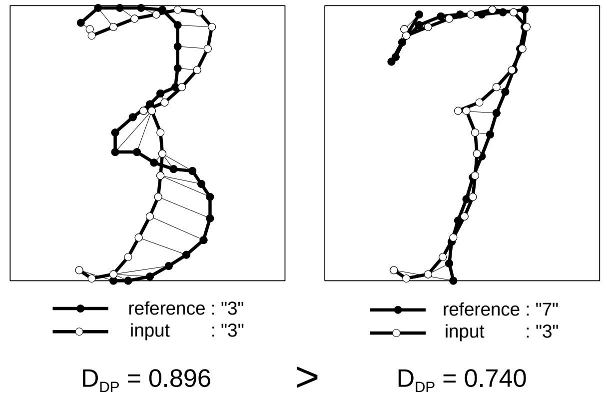

Figure 2. Example of misrecognitions due to overfitting.

has been directly employed as their discriminant functions. This will be because the value D_{DP} is a deformation-tolerant distance between $\boldsymbol{R}_c$ and $\boldsymbol{E}$.

The recognition based on D_{DP}, however, suffers from overfitting. A misrecognition due to the overfitting is shown in Fig. 2, where the input pattern "3" is misrecognized as "7" with D_{DP} underestimated by the unnatural matching. Such overfitting is caused by the regardlessness of category-dependent deformation tendencies. In fact, the deformation represented in the matching between "7" and "3" will be very rare for the category "7".

3. Quadratic Discrimination of Online Characters

3.1. Feature extraction — difference vector

In the proposed technique, a statistical discrimination framework using distribution of a new feature vector representing a global feature of the input pattern is employed to overcome the overfitting. This feature vector is called *difference vector* and derived from the result of DP matching. Accordingly, DP matching is just used as a feature extractor and D_{DP} is disregarded here.

The $3I$-dimensional difference vector of $\boldsymbol{E}$ can be obtained by using $j(1), \ldots, j(I)$, which is the optimal matching between two patterns $\boldsymbol{R}_c$ and $\boldsymbol{E}$, namely

$$\begin{aligned} \boldsymbol{v} \;=\; & \big((X_1 - x_{j(1)}, Y_1 - y_{j(1)}, \Theta_1 - \theta_{j(1)}), \ldots, \\ & (X_i - x_{j(i)}, Y_i - y_{j(i)}, \Theta_i - \theta_{j(i)}), \ldots, \\ & (X_I - x_{j(I)}, Y_I - y_{j(I)}, \Theta_I - \theta_{j(I)})\big)^T . \end{aligned} \tag{3}$$

This difference vector represents the deformation of $\boldsymbol{E}$ from the standard shape $\boldsymbol{R}_c$. In the following discussion, we will

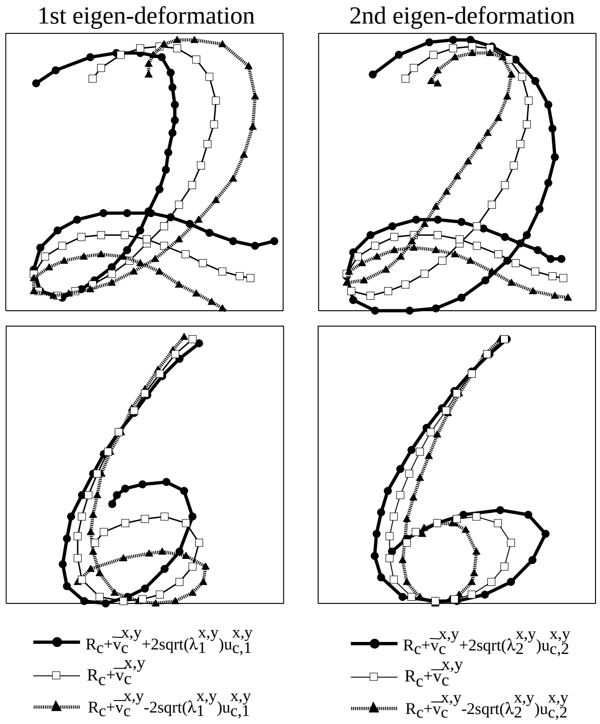

Figure 3. Reference pattern deformed by the first two eigen-deformations $\boldsymbol{u}_{c,1}^{x,y}$, $\boldsymbol{u}_{c,2}^{x,y}$ of "2" and "6".

use subvectors $\boldsymbol{v}^{x,y}$ and $\boldsymbol{v}^d$ defined as

$$\begin{aligned}\boldsymbol{v}^{x,y} &= \big((X_1 - x_{j(1)}, Y_1 - y_{j(1)}), \ldots, \\ &\quad (X_i - x_{j(i)}, Y_i - y_{j(i)}), \ldots, \\ &\quad (X_I - x_{j(I)}, Y_I - y_{j(I)})\big)^T \qquad (4)\\ \boldsymbol{v}^d &= (\Theta_1 - \theta_{j(1)}, \ldots, \\ &\quad \Theta_i - \theta_{j(i)}, \ldots, \Theta_I - \theta_{j(I)})^T. \qquad (5)\end{aligned}$$

3.2. Quadratic Discriminant Function

Under the assumption that the difference vectors have a Gaussian distribution, it is well known that their Bayes discrimination is reduced to a QDF, defined as

$$\begin{aligned}D_{\text{tot}}(\boldsymbol{R}_c, \boldsymbol{E}) &= (\boldsymbol{v} - \overline{\boldsymbol{v}}_c)^T \Sigma_c^{-1} (\boldsymbol{v} - \overline{\boldsymbol{v}}_c) \\ &\quad + \log|\Sigma_c| + 3I \log 2\pi, \qquad (6)\end{aligned}$$

where $\overline{\boldsymbol{v}}_c$ and Σ_c are the mean and the covariance matrix of $\boldsymbol{v}$ of category c, respectively. The last term of the right side of the equation cannot be omitted here because $3I$, the dimensionality of $\boldsymbol{v}$, is different for every category c.

Hereafter, the xy-coordinate feature is assumed to be independent of the directional feature for decreasing the dimension of Σ_c. Under the assumption, (6) can be divided into the discriminant function D_{pos} for the positional feature and D_{dir} for the directional feature, i.e.,

$$D_{\text{tot}}(\boldsymbol{R}_c, \boldsymbol{E}) = D_{\text{pos}}(\boldsymbol{R}_c, \boldsymbol{E}) + D_{\text{dir}}(\boldsymbol{R}_c, \boldsymbol{E}), \quad (7)$$

where

$$\begin{aligned}D_{\text{pos}}(\boldsymbol{R}_c, \boldsymbol{E}) &= (\boldsymbol{v}^{x,y} - \overline{\boldsymbol{v}}_c^{x,y})^T (\Sigma_c^{x,y})^{-1} (\boldsymbol{v}^{x,y} - \overline{\boldsymbol{v}}_c^{x,y}) \\ &\quad + \log|\Sigma_c^{x,y}| + 2I \log 2\pi \\ &= \sum_{m=1}^{2I} \frac{1}{\lambda_{c,m}^{x,y}} ((\boldsymbol{v}^{x,y} - \overline{\boldsymbol{v}}_c^{x,y})^T \boldsymbol{u}_{c,m}^{x,y})^2 \\ &\quad + \log \prod_{m=1}^{2I} \lambda_{c,m}^{x,y} + 2I \log 2\pi. \qquad (8)\end{aligned}$$

The matrix $\Sigma_c^{x,y}$ is the $2I \times 2I$ covariance matrix estimated by the subvector $\boldsymbol{v}^{x,y}$ between the training samples of category c and $\boldsymbol{R}_c$. The vector $\boldsymbol{u}_{c,m}^{x,y}$ and the value $\lambda_{c,m}^{x,y}$ are its eigenvector and eigenvalue, respectively.

It is well known that the estimation errors of higher-order eigenvalues are amplified in (8). Thus, in practice, the modified quadratic discriminant function (MQDF) [8] is employed, where the higher-order eigenvalues $\lambda_{c,m}^{x,y}$ $(m = M^{x,y}+1, \ldots, 2I)$ are replaced by $\lambda_{c,M^{x,y}+1}^{x,y}$, i.e.,

$$\begin{aligned}&D_{\text{pos}}(\boldsymbol{R}_c, \boldsymbol{E}) \\ &\sim \frac{1}{\lambda_{c,M^{x,y}+1}^{x,y}} \|\boldsymbol{v}^{x,y} - \overline{\boldsymbol{v}}_c^{x,y}\|^2 \\ &\quad + \sum_{m=1}^{M^{x,y}} \left(\frac{1}{\lambda_{c,m}^{x,y}} - \frac{1}{\lambda_{c,M^{x,y}}^{x,y}} \right) ((\boldsymbol{v}^{x,y} - \overline{\boldsymbol{v}}_c^{x,y})^T \boldsymbol{u}_{c,m}^{x,y})^2 \\ &\quad + \log \left\{ (\lambda_{c,M^{x,y}+1}^{x,y})^{2I - M^{x,y}} \prod_{m=1}^{M^{x,y}} \lambda_{c,m}^{x,y} \right\} \\ &\quad + 2I \log 2\pi. \qquad (9)\end{aligned}$$

While various methods for determining the parameter $M^{x,y}$ can be considered, the smallest $M^{x,y}$ which satisfy $\sum_{m=1}^{M^{x,y}} \lambda_{c,m}^{x,y} / \sum_{m=1}^{2I} \lambda_{c,m}^{x,y} > \mu^{x,y}$ was used in the recognition experiment of Section 5. The threshold $\mu^{x,y}$ is optimized by recognition experiment using training data.

The function D_{dir} can be derived in the same manner by changing $\boldsymbol{v}^{x,y}$ to $\boldsymbol{v}^d$, $\Sigma_c^{x,y}$ to Σ_c^d, and so on.

4. Eigen-deformations

The eigenvector $\boldsymbol{u}_{c,m}^{x,y}$ is a principal axis of the distribution of the difference vectors $\{\boldsymbol{v}^{x,y}\}$ in feature space. Therefore, a lower order $\boldsymbol{u}_{c,m}^{x,y}$ represents a frequent deformation of category c. Hereafter, $\boldsymbol{u}_{c,m}^{x,y}$ is called (positional) *eigen-deformations* of online character.

Figure 3 shows reference patterns of "2" and "6" deformed by the first two positional eigen-deformations $\boldsymbol{u}_{c,1}^{x,y}$

and $\boldsymbol{u}_{c,2}^{x,y}$. From this figure, deformations frequently observed in actual characters were detected as primary eigen-deformations. The first eigen-deformation of "6" represents the vertical variation of loop part, and the second one represents the horizontal variation of the loop part.

Any online character pattern of the category c will lie on the subspace spanned by these eigen-deformations. This fact will lead a new framework of online character recognition. (See also [9, 10]. Especially in [10], the eigen-deformations of handprinted characters are exploited according to the fact.)

5. Experimental Results

In this section, the recognition result by the quadratic discrimination of Section 3 is provided.

5.1 Dataset

About 16,000 isolated online digit samples from UNIPEN [11, 12] Train-R01/V07 database (1a) were used in our experiment. In this experiment, intractable samples (e.g., mislabeled samples and preprocessing error) was not cleaned off.

The database was divided into three data sets by using the software called utils2compareHWR [13, 14], which is recommended for dividing the UNIPEN database into training/test data. Multi-writer environment that writers of training data are not independent of writers of test data was employed on dividing data. Among three data sets, two data sets were used as training data, and the remaining one data set was used as test data. The following experimental result shows the average of the three trials by the cross-validation method. The training samples were used for the generation of reference patterns in Section 5.2, and the estimation of covariance matrix (i.e., eigen-deformations). The test samples were used as recognition tasks.

All samples were preprocessed as follows. First, each pen-up part was connected by a line segment in order to eliminate stroke-number variations. Thus, all samples subjected to the experiment were one-stroke characters. Then, their size was normalized to 128×128, keeping aspect ratio. Finally, resampling was performed so that the distance between consecutive sample points became constant.

5.2 Reference patterns

Reference patterns $\boldsymbol{R}_c$ were generated automatically using CLUSTER [15, 16, 17], which is a clustering method similar to k-means. By repeating split and merge of clusters, CLUSTER provides suboptimal k clusters $(k = 1, \ldots, K)$.

The number of clusters for each digit, i.e., k, was decided by the following parameter T, which is commonly used for

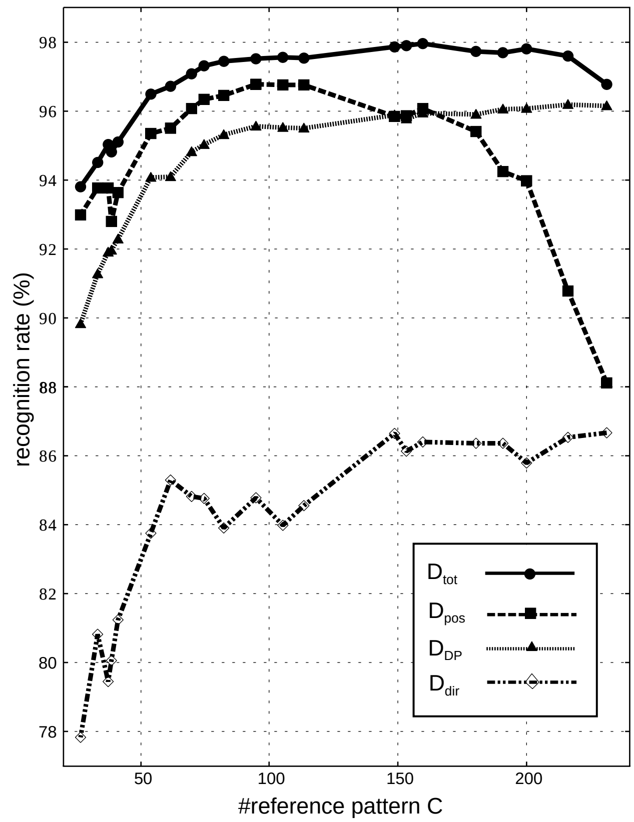

Figure 4. Recognition rate.

all digits. While the number k specifies the number of reference patterns representing the digit, it also affects the estimation of the covariance matrices. This is because, in the experiment, the training samples belonging to the cth cluster were used for the estimation of a covariance matrix Σ_c. Considering those facts, we set k at $\tilde{k}$ which is the smallest k satisfying $Q_k \geq T$, where Q_k is the size of the smallest clusters among k clusters. This implies that the parameter T specifies the minimum number of the training samples for estimating a single covariance matrix. In the following, the number of total clusters, that is $C = \sum_{\text{all digits}} \tilde{k}$, was controlled by changing T.

5.3 Recognition result

Figure 4 shows the recognition rates by D_{tot} as a function of C. The recognition rates by D_{DP}, D_{pos}, and D_{dir} are also shown in the figure for comparison.The highest result, the recognition rate attained by D_{tot} was $97.95\%(C = 160)$. In [4] it is reported that statistical dynamic time warping (SDTW), which is a recent and sophisticated version of DP matching, attained 97.10% on the same data set, when $C = 150$. In the same condition, our technique attained $97.85\%(C = 150)$ and thus outperforms the SDTW.

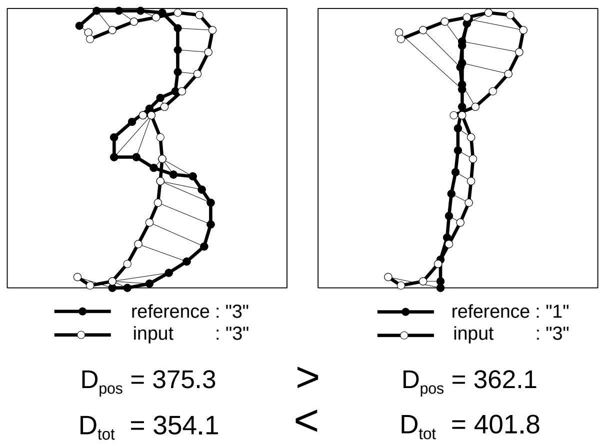

Figure 5. Sample correctly recognized by the proposed technique.

From Fig. 4, the recognition rates of $D_{\rm tot}$ were always higher than those of $D_{\rm DP}$. This can be considered that the misrecognitions due to overfitting were suppressed by QDF.

The recognition rates attained by $D_{\rm tot}$ were always higher than that by $D_{\rm pos}$ and $D_{\rm dir}$. Figure 5 shows one of improves. The input pattern "3" of Fig. 5 was misrecognized as "1" by $D_{\rm pos}$, because the difference of positional feature was small due to the unusual smoothness of the curvature around the middle part of the input pattern "3". The difference of directional feature, however, becomes larger, and finally $D_{\rm tot}$ could provide correct result. Thus, $D_{\rm pos}$ and $D_{\rm dir}$ are collaborative and should be used together in QDF.

A PC (Intel(R) Xeon(TM) CPU 3.06GHz) required 23.6 msec at $C = 160$. For recognizing a single character, that is, the proposed technique is practical from the view point of computational complexity.

6. Conclusion

A new online character recognition technique was proposed, where an elastic matching technique is combined with a statistical discrimination framework. In the proposed technique, elastic matching is used just as an extractor of the feature vector representing the difference between input and reference patterns. Then quadratic discrimination is performed under the assumption that the feature vector is governed by a Gaussian distribution. The result of a recognition experiment on UNIPEN database (Train-R01/V07, 1a) showed that the proposed technique could attain a high recognition rate (97.95%) and outperformed a recent elastic matching-based recognizer of [4]. During the derivation of the proposed technique, eigen-deformations were also introduced as the promising representation of the deformations of online characters. It was suggested that eigen-deformations can be utilized for developing another new recognizer by future work.

Acknowledgement: This work was supported in part by SECOM Science and Technology Foundation and Ministry of Internal Affairs and Communications in Japan under Strategic Information and Communications R&D Promotion Programme (SCOPE).

References

[1] K. Ikeda, T. Yamamura, Y. Mitamura, S. Fujiwara, Y. Tominaga, and T. Kiyono, "On-line recognition of handwritten characters utilizing positional and stroke vector sequence," Pattern Recognition, vol. 13, no. 3, pp. 191-206, 1981.

[2] K. Yoshida and H. Sakoe, "Online handwritten character recognition for a personal computer system", IEEE Trans. Consumer Electronics, vol. CE-28, no. 3, pp. 202-209, 1982.

[3] D. J. Burr, "Designing a handwriting reader", IEEE Trans. PAMI, vol. PAMI-5, no. 5, pp. 554-559, 1983.

[4] C. Bahlmann and H. Burkhardt, "The writer independent online handwriting recognition system flog on hand and cluster generative statistical dynamic time warping," IEEE Trans. PAMI, vol. 26, no. 3, pp. 299-310, 2004.

[5] R. Nag, K. H. Wong, and F. Fallside, "Script recognition using hidden Markov models," Proc. ICASSP, pp. 2071–2074, vol. 3 of 4, 1986.

[6] J. Hu, M.-K. Brown, and W. Turin, "HMM based online handwriting recognition," IEEE Trans. PAMI, vol. 18, no. 10, pp. 1039-1045, 1996.

[7] M. Nakai, N. Akira, H. Shimodaira, and S. Sagayama, "Substroke approach to HMM-based on-line Kanji handwriting recognition," Proc. ICDAR, pp. 491–495, 2001.

[8] F. Kimura, K. Takashina, S. Tsuruoka, "Modified quadratic discriminant functions and the application to Chinese character recognition," IEEE Trans. PAMI, vol. 9, no. 1, pp. 149-153, 1987.

[9] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models - their training and application," Comput. Vis. Image Und., vol. 61, no. 1, pp. 38–59, Jan. 1995.

[10] S. Uchida and H. Sakoe, "Handwritten character recognition using elastic matching based on a class-dependent deformation model," Proc. ICDAR, vol. 1 of 2, pp. 163–167, 2003.

[11] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, "UNIPEN project of on-line data exchange and recognizer benchmarks," Proc. ICPR, pp. 29-33, 1994.

[12] `http://hwr.nici.kun.nl/unipen/`

[13] E. H. Ratzlaff, "Methods, report and survey for the comparison of diverse isolated character recognition results on the UNIPEN database," Proc. ICDAR, pp. 623-628, 2003.

[14] `http://www.alphaworks.ibm.com/tech/comparehwr`

[15] A. K. Jain and R. C. Dubes, Algorithms for clustering data, Prentice Hall, 1988.

[16] R. Dubes and A. K. Jain , "Clustering techniques: the user dilemma" Pattern Recognit., vol. 8, pp.247-268, 1976.

[17] S. D. Connell and A. K. Jain, "Template-based online character recognition," Pattern Recognit., vol. 34, no. 1, pp. 1-13, 2001.