MVA2015 IAPR International Conference on Machine Vision Applications, May 18-22, 2015, Tokyo, JAPAN

Similar Handwritten Chinese Character Recognition based on Adaptive Discriminative Locality Alignment

Xiwen Qu, Ning Xu, Weiqiang Wang, Ke Lu
School of Computer and Control Engineering,
University of Chinese Academy of Sciences,
Beijing 101408, PR China
Email:queiwen13b@mails.ucas.ac.cn, xuning12@mails.ucas.ac.cn, wqwang@ucas.ac.cn
luk@ucas.ac.cn

Abstract

Discriminative locality alignment (DLA) has been successfully applied in similar handwritten Chinese character recognition (SHCCR). But, the performance of DLA heavily depends on the choice of parameters and the optimal parameters among different groups of similar characters are not consistent. To address this problem, we present an improved method with few parameters, called adaptive discriminative locality alignment (ADLA), whose optimal parameters are the same for different groups of similar characters. Further, the kernel discriminative locality alignment (KADLA) is formulated. The experimental results demonstrate that ADLA has higher performance than DLA in recognition rate, and KADLA has even higher recognition rate. In practice, Since KADLA involves much more time and storage cost, ADLA is a better choice for SHCCR.

1 Introduction

In recent years, handwritten Chinese character recognition (HCCR) has aroused intensive attention due to the development of PC and handheld devices (e.g., smart phone, writing tablets), and impressive achievements in both research and application have been obtained [1]. However, the recognition of unconstrained handwriting still involves some challenges, e.g., varied writing styles, confusion between similar characters and large set of characters [2, 3]. Jin's work [4], as well as classification results of Chinese handwriting recognition competition in ICDAR 2013 (CHRC2013), shows that the decrease of recognition rate in unconstrained HCCR mainly results from the resemblance of similar characters. Some examples of similar characters from the SCUT-COUCH2009 dataset [4] are shown in Figure 1.

Tao *et al.*[5] introduce discriminative locality alignment (DLA) [6, 7] to SHCCR and demonstrate that DLA has better performance than linear discriminant analysis (LDA) [8]. Later, to improve its performance on nonlinear features, kernel discriminative locality alignment (KDLA) is proposed [9]. Although DLA has many advantages, e.g., overcoming the nonlinearity of the distribution of samples, preserving discriminative information over local patches, and avoiding the matrix singularity problem, DLA is very sensitive to the choice of parameters, and the optimal parameters for each group of similar characters are not consistent.

To overcome the problems above, we present an improved DLA, termed adaptive discriminative locality alignment(ADLA). The proposed method only involves one parameter, and it adaptively selects the parameters in DLA. In the training phase, the process of parameters optimization is not needed. Correspondingly, the kernel adaptive discriminative locality alignment (KADLA) is also formulated in this paper.

The rest of the paper is organized as follows. Section 2 presents the ADLA and KADLA methods. Experiments and analysis are presented in Section 3. Section 4 concludes the paper.

2 Proposed methods

For a set of training samples $X = [x_1, \ldots, x_N]$, each sample $x_i \in R^m, i = 1, 2, \ldots, N$, and linear dimensionality reduction is to find a proper projection matrix U which projects x_i to y_i , $y_i \in R^l$, $l < m$, so that $Y = U^T X$, and $Y = [y_1, \ldots, y_N]$.

2.1 Discriminative Locality Alignment

In DLA algorithm, part optimization is first carried out for each sample. Concretely, a local patch is built for a given sample x_i and its neighbors according to class labels, and an objective function is designed to characterize local discriminative information. Then, DLA performs the whole alignment to integrate all part optimizations to form the global coordinate in the projected low-dimensional subspace[6, 7].

2.2 Adaptive Discriminative Locality Alignment

Although DLA has effectively improved the recognition rate in SHCCR compared with LDA [5, 9], some disadvantages still exist. (1) for each given sample

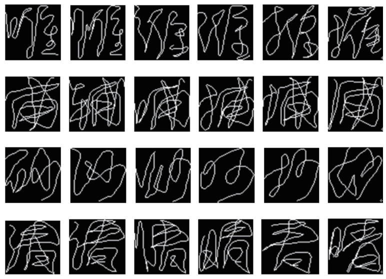


Figure 1. Similar samples with imaginary strokes from the SCUT-COUCH2009.

x_i, its local patch is built by k_1 nearest neighbors from the same class and k_2 nearest neighbors from different classes. The recognition accuracy heavily depends on the choice of parameters k_1 and k_2, and the optimal parameter combination of k_1 and k_2 are very different for different similar character set. (2) The weights of samples for different patches are invariable among the samples with the same class as x_i or different classes from x_i. To enhance the DLA algorithm, we present an adaptive discriminative locality alignment (ADLA). For a given training dataset $X = [x_1, \ldots, x_N]$, let n_j denote the number of samples in the jth class, and $l(x_i)$ denote the class label of sample x_i. In part optimization stage, for a given sample x_i, $1 \le i \le N$, we choose r_i samples with class label $l(x_i)$, and denote them by $x_{i^1}, \ldots, x_{i^{r_i}}$. In the proposed method, the number of samples is proportional to the number of selected samples in the same class, i.e., $r_i = \rho n_{l(x_i)}, \rho \in (0, 1)$. The experiments shows that $\rho = 0.95$ is a good choice. For each sample x_j with the same class label as x_i, we use a weight $\omega^{(s)}_{(i,j)}$ to measure its importance in modeling the patch, and it is evaluated by

$$\omega^{(s)}_{(i,j)} = \frac{1}{1 + \exp(-\|x_i - x_j\|)}, \quad l(x_j) = l(x_i), \quad (1)$$

where $\|x_i - x_j\|$ denote the Euclidean distance between two samples x_i and x_j. Finally, we rank the weights of all the samples x_j, $l(x_j) = l(x_i)$, and the r_i samples with the largest weights are chosen to modeling the patch corresponding to sample x_i

According to Euclidean distance, for sample x_i, its $n_{l(x_i)} - 1$ nearest neighbors can be identified. Among the neighbors, the samples with different class labels are chosen to model the patch, and they are denoted by $x_{i_1}, \ldots, x_{i_{m_i}}$, here $(0 \le m_i \le n_i - 1))$. Their corresponding weights are evaluated by

$$\omega^{(d)}_{(i,j)} = \frac{1}{1 + \exp(\|x_i - x_j\|)}, \quad l(x_j) \ne l(x_i). \quad (2)$$

Thus, the local patch for x_i can be formulated as $X_i = [x_i, x_{i^1}, \ldots, x_{i^{r_i}}, x_{i_1}, \ldots, x_{i_{m_i}}]$, and the corresponding patch in low dimensional space is $Y_i = [y_i, y_{i^1}, \ldots, y_{i^{r_i}}, y_{i_1}, \ldots, y_{i_{m_i}}]$. To make distances between y_i and its within-class samples very small and distances between y_i and its between-class samples as large as possible in the projected space, the distance between y_i and its neighbor samples of a same class are as small as possible, and the distance between y_i and its neighbor samples of different classes are as large as possible,a possible objective function can be defined by

$$\min_{y_i} \sum_{j=1}^{r_i} \|y_i - y_{i^j}\|^2 \omega^{(s)}_{(i,j)} - \beta_i \sum_{l=1}^{m_i} \|y_i - y_{i_l}\|^2 \omega^{(d)}_{(i,l)}. \quad (3)$$

It should be noted that the scaling factor β_i varies with different local patch, and in the proposed method it is defined as

$$\beta_i = \frac{(\sum_{k=1}^{r_i} \omega^{(s)}_{(i,k)})/r_i}{(\sum_{t=1}^{m_i} \omega^{(d)}_{(i,t)})/m_i}. \quad (4)$$

Then, Eq.(3) can be rewritten as

$$\min_{y_i} \sum_{j=1}^{r_i+m_i} \|y_{F_i\{1\}} - y_{F_i\{j+1\}}\|^2 \omega^*_{(i,j)} = \min_{Y_i} tr(Y_i L_i {Y_i}^T),$$
$$\omega^*_i = [\overbrace{\omega^{(s)}_{(i,1)}, \ldots, \omega^{(s)}_{(i,r_i)}}^{r_i}, \overbrace{-\beta_i\omega^{(d)}_{(i,1)}, \ldots, -\beta_i\omega^{(d)}_{(i,m_i)}}^{m_i}], \quad (5)$$

where $F_i = \{i, i^1, \ldots, i^{r_i}, i_1, \ldots, i_{m_i}\}$ is the index set of ith patch, $L_i = \begin{bmatrix} \sum_{j=1}^{r_i+m_i} \omega^*_{(i,j)} & -{\omega_i}^{*T} \\ -{\omega_i}^* & diag({\omega_i}^*) \end{bmatrix}$.

After part optimization step, the global alignment matrix L can be obtained by summing each L_i according to sample weights in a global coordinate, and N part optimization functions can be unified together as a whole one [7], i.e.,

$$\min_Y tr(YLY^T). \quad (6)$$

Further, let $Y = U^T X$, we have

$$\min_U tr(U^T XLX^T U) \quad s.t. \quad U^T U = I, \quad (7)$$

where I is an $l \times l$ identity matrix. The solution of U consists of l eigenvectors corresponding to the l smallest eigenvalues of XLX^T.

2.3 Kernel Adaptive Discriminative Locality Alignment

In the proposed KADLA, a nonlinear mapping Φ is used to map input data into a high dimensional space, where training samples are denoted by $\Phi(X) = [\Phi(x_1), \ldots, \Phi(x_N)]$. U can be obtained by:

$$[\Phi(X)L^\Phi\Phi(X)^T]U = \lambda U. \quad (8)$$

Different to ADLA, the β_i is given as follows:

$$\beta_i = p \times \frac{(\sum_{k=1}^{r_i} \omega^{(s)}_{(i,k)})/r_i}{(\sum_{t=1}^{m_i} \omega^{(d)}_{(i,t)})/m_i}. \quad (9)$$

where $p \in [0, 1]$ is a regulator, it is empirically set to 0.1 in the following experiments. Assuming that $V_i = [v_{(i,1)}, \ldots, v_{(i,N)}]^T$ and $V = [V_1, \ldots, V_l] \in R^{N\times l}$, because any eigenvector may be described as a linear combination of the observations in feature space, there exist coefficients $v_{(i,j)}$, $i = 1, \ldots, l$, $j = 1, \ldots, N$, such that

$$\begin{aligned} U_i =& \sum_j^N v_{(i,j)}\Phi(x_j) \\ =& \Phi(X)V_i, \end{aligned} \quad (10)$$

then, $U = \Phi(X)V$, Eq.(8) reduces to:

$$\begin{aligned} [\Phi(X)L^\Phi\Phi(X)^T]\Phi(X)V =& \lambda\Phi(X)V \\ \Phi(X)^T\Phi(X)L^\Phi\Phi(X)^T\Phi(X)V =& \lambda\Phi(X)^T\Phi(X)V \\ KL^\Phi KV =& \lambda KV, \end{aligned} \quad (11)$$

where L^{Φ} is global alignment matrix in kernel space, K is the gram matrix, its entry is given by a kernel function $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$. Then, V is composed of l eigenvectors corresponding to the l smallest eigenvalue of $L^{\Phi}K$.

3 Experiments

We evaluate the proposed methods on some related data from the SCUT-COUCH2009 dataset [4]. Concretely, ten groups of similar characters used by [5, 9] are used in our experiments, and they are all listed in Table 1. Each Chinese character class has 188 samples. Some examples of samples are shown in Figure 2.

In the expriments, all character data are first preprocessed, including smoothing, dot density equalization[11], interpolation[10], and then 8-directional features[10] are extracted. We compare the performance of ADLA, KADLA, DLA, KDLA and LDA [8] in terms of recognition rate. To remove redundant information and avoid singularity, for all algorithms, PCA[12] projection is first carried out. For LDA, the $n_i - C$ dimensions is retained after the PCA computation, and 160 dimensions for other methods. In our experiments, the Gaussian kernel $K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ are employed by KDLA and KADLA, and empirically σ to 1.5. All experiments are carried out using Matlab. The KNN classifier is employed to classify data in low dimensional space.

Table 1. The dataset used in our evaluation experiments.

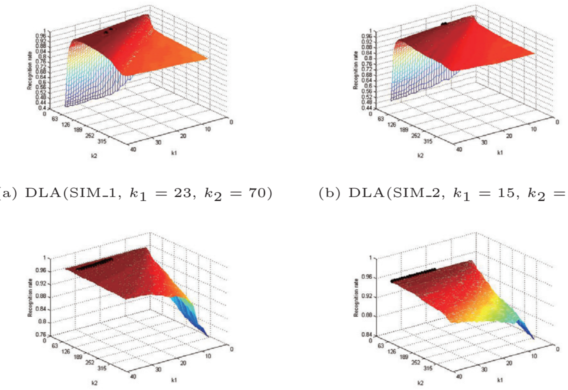
Group ID.	First candidate	Similar characters
SIM_1	氨	氨氮氦氯氢氛氰姜氖氟
SIM_2	柴	柴柒紫染架浆渠梨梁粱
SIM_3	刁	刁习刀刃司了门勺丁可
SIM_4	差	差羞羌姜善羔美养羡盖
SIM_5	凡	凡风凤几见叉冗丸贝八
SIM_6	父	父文久义叉欠叉夕人丈
SIM_7	斤	斤厅斥个介币牙丫布什
SIM_8	盎	盎蛊盅盏盔盗鸯签置盘
SIM_9	笆	笆爸琶筐邑笔芭岂色苞
SIM_10	敞	敞弊敲敬澈版骸歇淑散

3.1 Choices of parameters k_1 and k_2

For DLA and KDLA, parameters k_1 and k_2 can be selected in the range of $[1, n_i]$ and $[0, N - n_i]$. For each character class, 35 samples are selected (i.e., $n_i = 35$) for training, and the remaining samples for testing. The parameters β are empirically set to 0.15 and 0.01

Figure 2. Examples of handwritten samples corresponding to similar characters in Table 1.

for DLA and KDLA respectively, and the reduced dimension is fixed at 9. We use two groups of similar characters, SIM_1 and SIM_2, to illustrate the influence of choice of parameters k_1 and k_2 on the recognition for DLA and KDLA. As shown in Figure 3, the black dots are the peaks of recognition rate surface, and apparently they are different for different groups of similar characters. The experimental results show that DLA and KDLA are very sensitive to the choice of parameters. In practice, it is very difficult for DLA and KDLA to find an optimal parameter combination. Comparatively, the proposed methods, ADLA and KADLA, have only one parameter ρ, which is easily adjusted. In the following experiments, two optimal values from SIM_1 and SIM_2 are chosen for all other groups of similar characters.



(a) DLA(SIM_1, $k_1 = 23$, $k_2 = 70$) (b) DLA(SIM_2, $k_1 = 15$, $k_2 = 22$)

(c) KDLA(SIM_1, $k_1 = 32$, $k_2 = 28$) (d) KDLA(SIM_2, $k_1 = 34$, $k_2 = 5$)

Figure 3. Recognition rate vs. parameters k_1 and k_2 for DLA and KDLA.

3.2 Evaluation experiments

For five algorithms mentioned above, we compare their performance on ten groups of similar Chinese characters and summarize the average recognition rates under different reduced dimensions in Table 2. For each algorithm and a given reduced dimension, the result is obtained by averaging the recognition rates for ten groups of similar characters. As shown in Figure 4 and Table 2, when the reduced dimension is more than 9, the recognition rates of different algorithm except LDA keep very stable, and the solution does not exists due to matrix singularity for LDA when $RD = 11$. The experimental results show that the performance of ADLA and KADLA is a little better than that of DLA and KDLA. The recognition performance of KDLA and DLA varies with the different optimal parameters from different groups of training samples, e.g., KDLA and DLA with parameters combination $(k1, k2)$=$(32, 28)$ and $(15, 22)$. The training process for DLA and KDLA involves the search of optimal parameters, which is not easily controlled. Although the results show the kernel methods generally have higher accuracy than their counterpart, they usually involve much higher time and storage cost. For example, if the number of classes is 3755, the dimension of features is 512, the number of training samples for each class is 35, and each element of feature vector needs 4-bytes, the storage cost is $3755 \times 35 \times 512 \times 4 = 269M$ bytes. Additionally, projection matrix and gram matrix also occupy a large volume of storage space. We also carry out the experiments on a Core (TM) i-3470 3.20 GHz computer with 4 GB memory to compare computation cost of different algorithms. The number of candidate similar characters is 10, and the computation time of 1530 test samples for KDLA, KADLA, DLA and ADLA is 2.42s, 2.52s, 0.065s and 0.066s respectively. Obviously, kernel trick brings more time cost. So ADLA is more suitable for practical application than KDLA and KADLA.

Table 2. Average recognition rates (%) of ten groups of similar characters. RD denotes reduced dimension.

RD	Methods						
	LDA	DLA (23,70)	DLA (15,22)	ADLA	KDLA (32,28)	KDLA (34,5)	KAD –LA
1	37.2	37.3	39.2	37.4	41.5	30.1	39.3
2	61.2	64.6	62.4	62.2	64.1	47.0	64.6
3	73.1	77.2	77.2	75.0	76.5	63.8	75.7
4	79.6	84.0	84.0	83.4	83.8	77.4	83.4
5	83.7	88.1	88.7	88.6	87.6	82.6	87.4
6	85.5	90.5	91.3	91.6	90.5	86.5	90.4
7	87.1	92.2	93.0	93.3	92.2	90.2	92.1
8	87.9	93.0	93.9	94.0	93.6	91.8	93.7
9	88.1	94.1	94.5	94.8	95.58	94.96	95.68
10	88.0	94.1	94.5	94.8	95.58	94.96	95.68
11	NaN	94.1	94.5	94.8	95.58	94.96	95.68

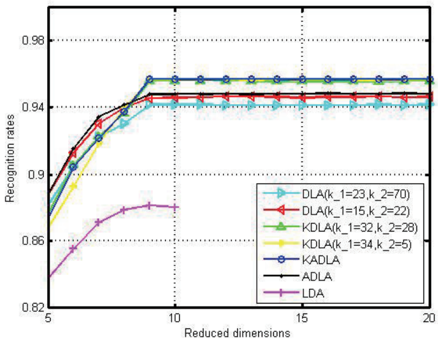


Figure 4. Average recognition rates (%) vs. reduced dimension.

4 Conclusions

In this paper, we present an improved DLA method with few parameters, called adaptive discriminative locality alignment (ADLA). Compared with DLA, the proposed method has better recognition rate. It inherits all the advantages of DLA and makes the training process become easy due to no computation of parameter optimization. The preliminary experimental results demonstrate the effectiveness of ADLA and KADLA for SHCCR in terms of recognition rate. For SHCCR, ADLA is very competitive, since it has higher computation efficiency and less memory demand than its kernel version.

Acknowledgements

We would like to thank L.-W. Jin et. al. for helping to providing the experimental data. This work is supported by the National Science Foundation of China (NSFC) under Grant no.61232013, no.61271434, no.61175115.

References

[1] C.-L. Liu, S. Jaeger, M. Nakagawa: “Online recognition of Chinese characters: the state-of-the-art,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.26, no.2, pp.198–213, 2004.

[2] C.-L. Liu, K. Nakashima, H. Sako, H. Fujisawa: “Handwritten digit recognition: investigation of normalization and feature extraction techniques,” Pattern Recognition vol.37, no.2, pp.265–279, 2004.

[3] F. Yin, Q.-F. Wang, X.-Y. Zhang, C.-L. Liu: “ICDAR 2013 Chinese handwriting recognition competition,” 2013 International Conference on Document Analysis and Recognition, Washington, DC, American, pp.1464–1470, 2013.

[4] L.-W. Jin, Y. Gao, G. Liu, Y.-Y. Li,K. Ding: “SCUT-COUCH2009 – a comprehensive online unconstrained Chinese handwriting database and benchmark evaluation,” International Journal on Document Analysis and Recognition, vol.14, no.1 pp.53–64, 2010.

[5] D.-P. Tao, L.-Y Liang, L.-w. Jin, Y. Gao: “Similar Handwritten Chinese Character Recognition using Discriminative Locality Alignment Manifold Learning,” 2011 International Conference on Document Analysis and Recognition, Beijing, China, pp.1012-1061, 2011.

[6] T.-H. Zhang, D.-C. Tao, X.-L. Li, and J. Yang: “Patch alignment for dimensionality reduction,” IEEE Transactions on Knowledge and Data Engineering, vol.21, no.9, pp.1299–1313, 2009.

[7] T.-H. Zhang, D.-C. Tao, J. Yang: “Discriminative locality alignment,” 10th European Conference on Computer Vision, Marseille, France, pp.725–738, 2008.

[8] R. A. Fisher: “The Use of Multiple Measurements in Taxonomic Problems,” Annals of Eugenics, Vol.7, pp.179–188, 1936.

[9] D.-P. Tao, L.-Y. Liang, L.-W. Jin, Y. Gao: “Similar handwritten Chinese character recognition by kernel discriminative locality alignment,” Pattern Recognition Letters, vol.35, no.1, pp.186–194, 2014.

[10] Z.-L. Bai, Q. Huo: “A Study On the Use of 8-Directional Features For Online Handwritten Chinese Character Recognition,” 2011 International Conference on Document Analysis and Recognition, Beijing, China, pp.232-236, 2005 .

[11] Z.-L. Bai, Q. Huo: “A Study of Nonlinear Shape Normalization for Online Handwritten Chinese Character Recognition: Dot Density vs. Line Density Equalization,” The 18th International Conference on Pattern Recognition, Hong Kong, vol.2, pp.921-924, 2006.

[12] H. Hotelling: “Analysis of a complex of statistical variables into principal components,” Journal of Educational Psychology, vol.24, no.6, pp.417-441, 1933.