International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 10, No. 7 (2017), pp. 157-166
http//dx.doi.org/10.14257/ijsip.2017.10.7.14

Distance Metric with Kullback–Leibler Divergence for Classification

Dongyun Qian* and Huifeng Jin

Zhejiang Industry & Trade Vocational College, Wenzhou, Zhejiang, China, 325003
qiandongyun@163.com

Abstract

Recently, traditional Euclidean-distance-based algorithms have shown their disabilities because the intrinsic space which samples lie in may not be Euclidean space. An excellent distance metric which can describe similarities between samples correctly can improve the performances of most machine learning tasks greatly. Therefore, learning an excellent distance metric is of vital importance but challenging. Up to now, many distance metric learning methods have been proposed using various techniques. This paper proposed a novel method named Distance Metric with Kullback-Leibler Divergence (DMKD) which fully utilizes Kullback-Leibler Divergence to describe scatters between classes. DMKD introduces the theory of information entropy into the field of distance metric. It maximizes Kullback-Leibler Divergences between classes to improve its discriminative ability. Meanwhile, an iterative optimization strategy is adopted to find the optimization solution of DMKD. The obtained distance metric can separate samples from different classes easily. Various experiments on benchmark datasets have been carried out to verify the excellent performance of this novel method.

Keywords*: Distance Metric Learning, Kullback-Leibler Divergence, Image classification*

1. Introduction

As we all know, the performances of most traditional algorithms based on Euclidean metric may failure facing with samples which don't lie in Euclidean space. Therefore, learning an appropriate distance metric is of vital importance for many machine learning and computer vision tasks [1-3], such as classification [1], face recognition [2] and image retrieval. In order to overcome the limitations of these Euclidean-distance-based algorithms, many researchers have proposed some distance metric learning methods [4-7] from different perspectives. Xing.P [4] proposed a global method which constructs a convex optimization model maximizes the sum of all distances between samples from different classes. Meanwhile, two constraints are added to ensure a feasible solution. Large Margin Nearest Neighbor (LMNN) [5] aims to find an dimensional spaces where the margins between different classes are pulled away as far as possible. Information-Theoretic Metric Learning (ITML) [6] constructs an Bregman optimization which minimizes the differential relative entropy between two multivariate Gaussians. Constraint-Margin Maximization (CMM) [7] embeds original samples into a low-dimensional space and calculates Euclidean distance in this space.

In this paper, we proposed a novel distance metric learning method called Distance Metric with Kullback–Leibler Divergence (DMKD) which adopts Kullback–Leibler (KL) [9] divergence to describe scatters between classes. DMKD maximizes the sum of KL divergences between classes to separate samples from different classes. Meanwhile, a

* *Corresponding Author

ISSN: 2005-4254 IJSIP

constraint is added in DMKD. The constraint ensures that the sum of distances between samples from the same class is less than one certain constant. Then we develop an iterative optimization strategy to find the optimization solution of DMKD.

This paper is organized as follows: Section 2 introduces some basic knowledge of distance metric. In Section 3, the procedure to construct DMKD is illustrated in detail. Meanwhile, the optimization procedure is explained. In Section 4, we conduct several experiments on benchmark datasets to show the excellent performance of this novel distance metric learning method. In Section 5, we summarize this paper and make a conclusion.

2. Related Work

In this section, we introduce some basic knowledge of distance metric briefly. The 4 properties which all distance metric must follow are illustrated. For any two samples x_i and x_j which locates in a d -dimensional space, their Mahalanobis distance can be calculated as follows:

$$d_A\left(x_i, x_j\right)=\sqrt{\left(x_i-x_j\right)^T A\left(x_i-x_j\right)} \tag{1}$$

where d_A is the distance between x_i and x_j using the distance metric matrix A. $A \in R^{d \times d}$ is the metric matrix which is positively semi-definite. Therefore, A can be decomposed into $A=WW^T$, where $W \in R^{l \times d}$ and $l<d$. Therefore, it's equivalent to find an optimal projection matrix W to project samples into the low-dimensional space and calculate the Euclidean distance between samples. Meanwhile, for any distance metric, the four properities should be always satisfied:

1) Triangular inequality: $d_A\left(x_i, x_j\right)+d_A\left(x_j, x_k\right) \geq d_A\left(x_i, x_k\right)$

2) Non-negativity: $d_A\left(x_i, x_j\right) \geq 0$

3) Symmetry: $d_A\left(x_i, x_j\right)=d_A\left(x_j, x_i\right)$

4) Distinguishability: $d_A\left(x_i, x_j\right)=0 \Leftrightarrow x_i=x_j$

3. Distance Metric with Kullback–Leibler Divergence

In this section, we proposed a novel distance metric learning method called distance metric with Kullback-Leibler divergence (DMKD). DMKD is a novel distance metric which maximizes KL divergence to separate samples from different classes. There is a constraint adopted to ensure DMKD to obtain a feasible metric. Then the optimization procedure of DMDK is illustrated in detail.

4.1. The Construction Procedure of DMKD

In this section, we introduce the procedure of DMKD in detail. In order to describe scatters between classes, DMKD introduces KL divergence into the field of distance metric. It maximizes KL divergences between classes to improve the discriminative ability. Assume we are given n samples $x_1, x_2, \cdots, x_n$ from c classes which follow Gaussian distributions. Each $x_i \in R^d$ locates in a d -dimensional space. For Gaussian probability density functions $p_i=N\left(x ; u_i, \sum_i\right)$, where $u_i \in R^d$ is the mean vector of the i class measurements, and $\sum_i \in R^{d \times d}$ is the with-class covariance matrix of the i class. Then the KL divergence between the i and j classes is listed as follows:

$$D\left(p_i \| p_j\right) = \int dx N\left(x; u_i, \Sigma_i\right) \ln \frac{N\left(x; u_i, \Sigma_i\right)}{N\left(x; u_j, \Sigma_j\right)} = \frac{1}{2}\left[\ln\left|\Sigma_j\right| - \ln\left|\Sigma_i\right| + tr\left(\Sigma_j^{-1}\Sigma_i\right) + tr\left(\Sigma_j^{-1} D_{ij}\right)\right] \quad (2)$$

where $D\left(p_i \| p_j\right)$ is the KL divergence between two classes. $D_{ij} = \left(u_i - u_j\right)\left(u_i - u_j\right)^T \in R^{d\times d}$ and $\left|\Sigma\right| = \det\left(\Sigma\right)$. However, KL divergence in Eq.(2) is the calculated in the Euclidean space. Because the distance metric matrix is $A = WW^T \in R^{d\times d}$, the KL divergence between these two classes using the new distance metric can be expressed as follows:

$$D_W\left(p_i \| p_j\right) = D\left(p\left(W^T x \mid y = i\right) \| p\left(W^T x \mid y = j\right)\right) = \frac{1}{2}\left[\begin{array}{l} \ln\left|W^T \Sigma_j W\right| - \ln\left|W^T \Sigma_i W\right| + \\ tr\left(\left(W^T \Sigma_j W\right)^{-1}\left(W^T\left(\Sigma_i + D_{ij}\right)W\right)\right) \end{array}\right] \quad (3)$$

Then the construction procedure of metric matrix A is equivalent to find the optimal projection matrix $W \in R^{d\times l}$. DMKD aims to maximize the sum of KL divergences between all classes to separate samples from different classes. And the objective function of DMKD is constructed as follows:

$$\max_W \sum_{1\le i,j\le c} D_W\left(p_i \| p_j\right) \quad (4)$$

Maximizing Eq.(4) makes samples from different classes separate in the new built metric space. Meanwhile, in order to obtain a feasible distance metric, distances between samples from the same class should be restricted. Therefore, DMKD is added a constraint as follows:

$$\begin{aligned} &\max_W \ L_W = \sum_{1\le i,j\le c} D_W\left(p_i \| p_j\right) \\ &s.t. \sum_{1\le a,b\le c} \eta_{ab} \left\|W^T x_a - W^T x_b\right\|^2 < 1 \end{aligned} \quad (5)$$

If x_a and x_b belongs to the same class $\eta_{ab} = 1$, 0 otherwise. DMKD aims to maximizing the sum of KL divergences between classes while compacting samples from the same class. Contrast with other distance metric learning methods, such LMNN, CMM, DMKD utilizes KL divergence to describe scatters between classes and it is more suitable.

3.2. The Optimization Procedure of DMKD

In this section, we introduce the optimization procedure of DMKD in detail. DMKD utilizes gradient ascent and iterative projection to achieve the optimal solution. Gradient ascent is utilized by DMKD to maximize the objective function of Eq.(5) while iterative projection is utilized to ensure the constraint of Eq.(5) satisfied. The optimization procedure repeats the two steps until A converges. First, DMKD takes a gradient step $W = W + \alpha \frac{\partial L_W}{\partial W}$ to maximize the objective function of Eq.(5). And $\frac{\partial L_W}{\partial W}$ is calculated as follows:

$$\frac{\partial L_W}{\partial W} = \sum_{1 \le i, j \le c} \partial_W D_W \left(p_i \| p_j \right) \tag{7}$$

where

$$\partial_W D_W \left(p_i \| p_j \right) = \begin{pmatrix} \sum_j W \left(W^T \sum_j W \right)^{-1} - \sum_i W \left(W^T \sum_i W \right)^{-1} + \\ \left(\sum_i + D_{ij} \right) W \left(W^T \sum_j W \right)^{-1} - \\ \sum_j W \left(W^T \sum_j W \right)^{-1} W^T \left(\sum_i + D_{ij} \right) W \left(W^T \sum_j W \right)^{-1} \end{pmatrix} \tag{8}$$

Therefore, we can utilize gradient ascent to update W which can maximize Eq.(5). Then, W should be updated again to meet the constraint of Eq.(5). Finally, the distance metric A can be represented as WW^T.

4. Experiment

In this section, we construct several experiments on various benchmark datasets to show the excellent performance of our proposed method. At first, we summarize the attributes of all datasets and the comparing methods in Section 4.1. And the experiment results are shown in Section 4.2 to show the performance of DMKD.

4.1. Datasets and Comparing Methods

There are 4 datasets utilized in our experiment, including ORL[1], AR[2], Isolet[3] and Caltech 101[4] datasets. All attributes of these datasets are summarized in Table.1 as follows:

Table 1. Attributes of All Datasets

Datasets	Sizes	Classes	Dimensions
ORL	400	40	1024
AR	1680	120	2000
Isolet	7797	26	617
Caltech 101	9146	102	– –

For ORL face dataset, there are 400 faces corresponding to 40 people's faces. And each people has 10 face images. All images are taken at different times, facial expressions, varying the lighting and some facial details. AR face dataset consists of 1680 images corresponding to 120 people's faces. Each people has 14 face images with different expressions. Isolet is a dataset of letters of English alphabet spoken in isolation. It consists of 7797 spoken letters, 2 productions of each letter by 150 speakers. Caltech 101 is an image dataset which contains 9145 images from 102 different objects. And the size of each image is roughly 300 × 200 pixels. And some images from these datasets are shown in Table.1 as follows:

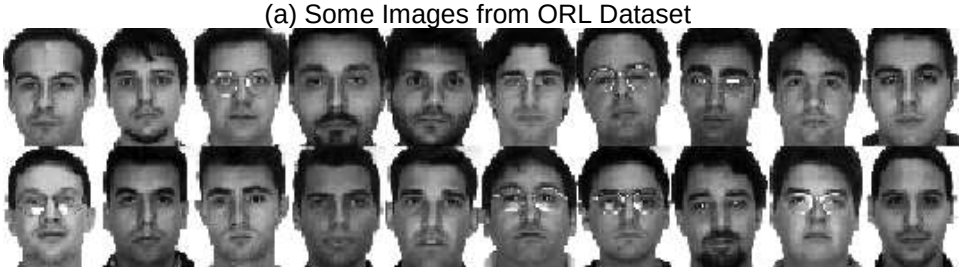
[1] http://www.uk.research.att.com/facedatabase.html

[2] http://rvl1.ecn.purdue.edu/~ aleix/aleix face DB.html

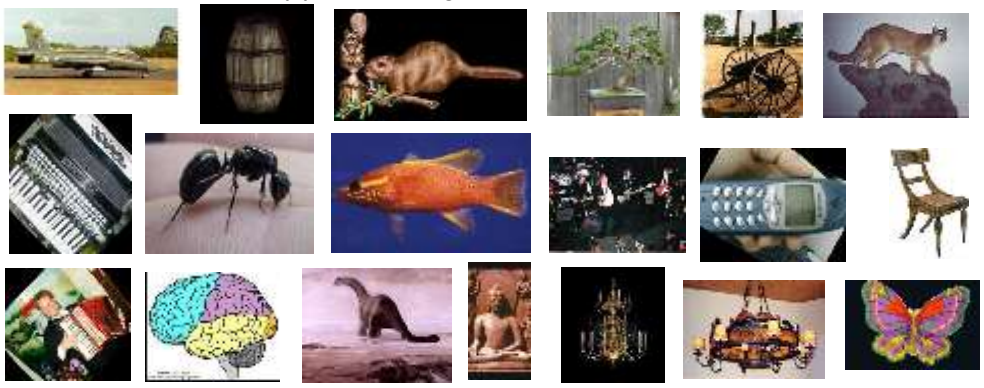
[3] http://archive.ics.uci.edu/ml/datasets/ISOLET

[4] http://www.vision.caltech.edu/Image Datasets/Caltech101/

(a) Some Images from ORL Dataset



(b) Some Images from AR Dataset



(c) Some Images from Caltech 101 Dataset

Figure 1. Some Images from Image Datasets

In order to verify the performance of DMKD, we select 6 famous distance metrics to be comparing methods, including Xing [4], CMM [7], ITML [6], LMNN [5], Euclidean [10], Chebychev [11].

4.2. Classification Experiment

In order to verify the performance of DMKD, this section construct several classification experiments on the datasets above. At first, all distance metrics are trained on training samples using different methods. Then, 1NN classification is utilized to classify the testing samples into different classes. And the classification accuracies are calculated to show the performances of all distance metrics.

For ORL dataset, we randomly select different numbers of samples as training ones. And the left samples are selected as testing ones. Figure2 shows the classification accuracies on ORL dataset .All accuracies are calculated 10 times and we show the mean values as Figure2.

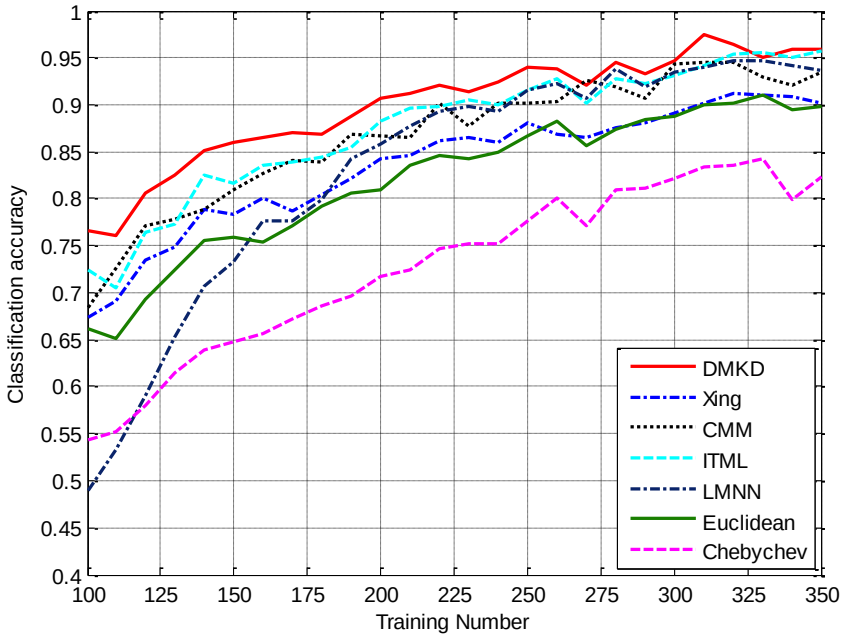


Figure 2. The Classification Accuracies on ORL Dataset

It's obvious that DMKD outperforms the other 6 distance metrics in most situations. Meanwhile, ITML and LMNN can also achieve good performances.

Figure3 shows the classification accuracies on AR dataset. Different numbers of samples are randomly selected as training ones. All distance metrics are trained and 1NN classification is utilized on testing samples. And the classification accuracies are summarized in Figure3 as follows:

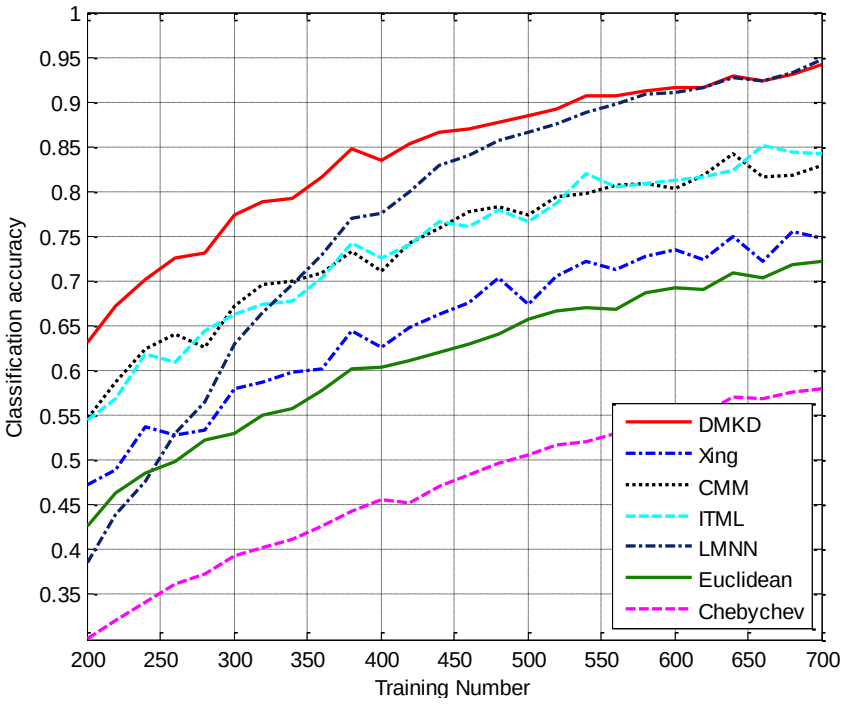
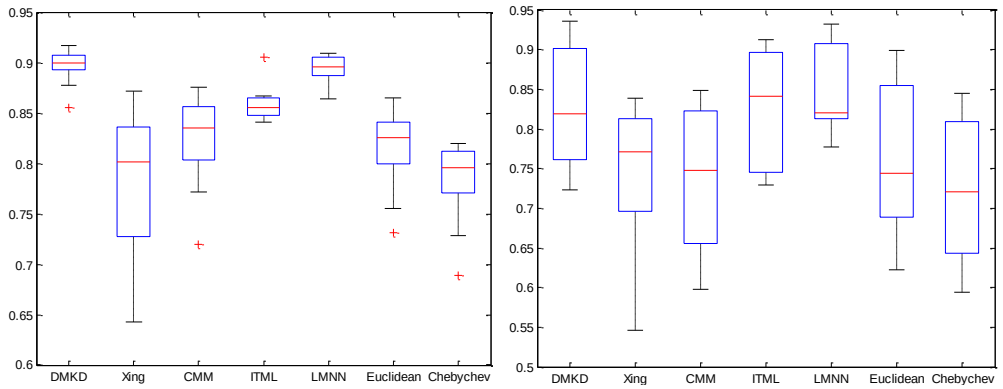


Figure 3. The Classification Accuracies on AR Dataset

We can clearly find that DMKD can achieve better performances than the other 6 distance metrics. With the increase of the number of training samples, the performance of LMNN increases largely.

Because Isolet contains 5 subsets, we choose the first 2 subsets(Isolet1, Isolet2) to conduct our experiment. For these two subsets, we randomly select 400 samples as training ones. The setting of this experiments is just like those above and the classification accuracies are summarized as box plots as Figure4.



(a) Classification Accuracies on Isolet1 (b) Classification Accuracies on Isolet2

Figure 4. Box Plots of Classification Accuracies on Isolet Dataset

For Caltech 101, we randomly select 5000 samples as training ones. Then, LLC[12] features are utilized to represent all samples. All distance metrics are trained and 1NN classifier is utilized to classify all testing samples. We conduct this experiment 10 times and show the mean and min classification errors as Figure5.

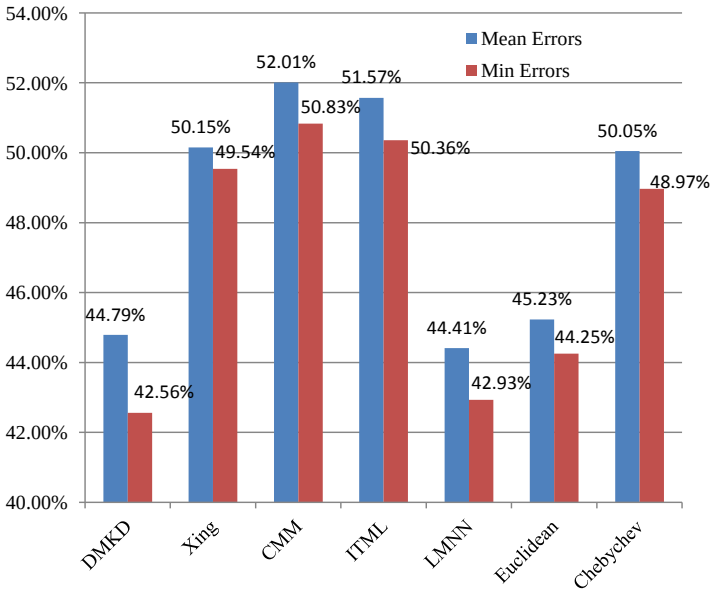


Figure 5. Mean and Min Classification Errors on Caltech 101 Datasets

It's obviously that DMKD is a better distance metric compared with the other 6 distance metrics. Meanwhile, LMNN can also achieve a good performance.

5. Conclusion

In this paper, we proposed a novel distance metric learning method called Distance Metric with Kullback-Leibler Divergence (DMKD). DMKD defines the divergences between two different classes using KL-divergence. Then, a constraint is added to ensure DMKD to obtain a feasible distance metric. Gradient ascent is utilized for DMKD to find the optimal solution. Various experiments on benchmark datasets show that DMKD can achieve good performances in most situations.

Acknowledgments

The authors would like to thank the reviewers for their comments which has improved the quality of the work. This work is supported by the project of the Domestic Visitor Foundation from the Education Commission of Zhejiang Province, China (FX2014177), project of the National Education Information Technology, China (146231964), project of Study on Metric Learning based on Geometric Mean supported from the innovation project of Zhejiang Industry & Trade Vocational College(G160204).

References

[1] C. Schüldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach", Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, England, **(2004)**.
[2] P. J. Phillips, H. Moon and S. A. Rizvi, "The FERET evaluation methodology for face-recognition algorithms", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 10, **(2000)**, pp. 1090-1104.
[3] H. Wang, L. Feng and J. Zhang, "Semantic Discriminative Metric Learning for Image Similarity Measurement", IEEE Transactions on Multimedia, vol. 18, no. 8, **(2016)**, pp. 1579-1589.
[4] Y. Gong, S. Lazebnik and A. Gordo, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, **(2013)**, pp. 2916-2929.
[5] E. P. Xing, A. Y. Ng and M. I. Jordan, "Distance metric learning with application to clustering with side-information", Proceedings of advances in neural information processing systems. British Columbia, Canada, **(2003)**.
[6] K. Q. Weinberger, J. Blitzer and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification", Proceedings of advances in neural information processing systems. British Columbia, Canada, **(2005)**.
[7] J. V. Davis, B. Kulis and P. Jain, "Information-theoretic metric learning. Proceedings of the 24th international conference on Machine learning", Corvallis, USA, **(2007)**.
[8] F. Wang, "Semisupervised metric learning by maximizing constraint margin", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 41, no. 4, **(2011)**, pp. 931-939.
[9] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification", Proceedings of Asian conference of Computer Vision. Queenstown, New Zealand, **(2010)**.
[10] T. M. Cover and J. A. Thomas, "Elements of information theory", John Wiley & Sons, Manhattan, **(2012)**.
[11] P. E. Danielsson, "Euclidean distance mapping", Computer Graphics and image processing, vol. 14, no. 3, **(1980)**, pp. 227-248.
[12] T. Klove, T. T. Lin and S. C. Tsai, "Permutation arrays under the Chebyshev distance", IEEE Transactions on Information Theory, vol. 56, no. 6, **(2010)**, pp. 2611-2617.
[13] J. Wang, J. Yang and K. Yu, "Locality-constrained linear coding for image classification", Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, **(2010)**, June 13-18.

Authors

Dongyun Qian, female, Chinese. She is an associate professor of Zhejiang Industry & Trade Vocational College, Zhejiang province, China. Her research interests are image analysis, pattern recognition, *etc*.

Huifeng Jin, male, Chinese. He is a lecturer of Zhejiang Industry & Trade Vocational College, Zhejiang province, China. His research interests are data mining, image classification, pattern recognition, *etc*.